Security Protocols and Resource Allocation for Fifth Generation Networks

Arsenia Chorti

ETIS UMR 8051, CY Université, ENSEA, CNRS

Security Protocols and Resource Allocation for Fifth Generation Networks

Arsenia (Ersi) Chorti
Maître de Conférences à l'ENSEA
Habilitation à Diriger des Recherches de CY Cergy Paris Université
Section CNU 27, Informatique

Defended on the 12th October 2020 in front of the jury composed of

Jean-Marie GORCE Professor, INSA-Lyon, President of the jury

Gerhard FETTWEIS, Professor, TU Dresden, Vodafone Chair, Reviewer
Marceau COUPECHOUX, Professor, Télécom ParisTech, LTCI, Reviewer
Ghaya REKAYA, Professor, Télécom ParisTech, LTCI, Reviewer

Camilla HOLLANTI, Professor, Aalto University, Dep. of Mathematics and System Analysis, Examiner
Inbar FIJALKOW, Professor, ENSEA HDR Guarantor
Iryna ANDRIYANOVA, Professor, CY Cergy Paris University, HDR Referee

I hereby declare that this thesis and the work reported herein was composed by and originated entirely from me, the following PhD students I co-supervise: M. Mitev, S. Skaperas, G. S. Nunez and M. Bello, and their theses's directors, unless otherwise stated. Information derived from the published and unpublished work of others has been acknowledged in the text and references are given in the list of sources.

Arsenia (Ersi) Chorti (2020)

Abstract

In my role as a research advisor, I strove to give to my students the opportunity to work on promising and far fetching – whenever possible – research topics in the general framework of fifth generation (5G) wireless. The works presented in this thesis reflect our studies in two areas of central importance for bringing 5G to life: wireless security and resource allocation.

With respect to security, novel challenges emerged in 5G with the Internet of things (IoT) paradigm and device to device (D2D) low latency communications. Novel verticals, such as haptics and vehicle to everything (V2X), require low complexity and low latency security mechanisms, particularly in the context of device authentication. In the present manuscript, lightweight solutions for device authentication using physical unclonable functions (PUF) and secret key generation (SKG) at the physical layer are presented.

Furthermore, as video content is responsible for more than 70% of the global IP traffic, it is important for content delivery infrastructures to rapidly detect and respond to changes in content popularity dynamics. In this thesis, we propose a flexible edge resource allocation approach leveraging unikernel and container technologies. The allocation of the edge server resources is driven by a real-time and low-complexity content popularity detector, implemented using off-line and on-line change point analysis. Variations of these algorithms have applications in intrusion detection in wireless sensor software defined networks, discussed next.

Finally, the potential use of non-orthogonal multiple access (NOMA) in the wireless uplink is considered. Early results on the performance comparison of NOMA vs orthogonal allocation schemes in asymptotic regimes, show that the gains in using NOMA carry on to the scenario of communications under statistical delay quality of service (QoS) constrains.

Dans mon rôle de co-encadrent de thèse, je me suis efforcé de donner à mes étudiants l'occasion de travailler sur des sujets de recherche prometteurs et fondamentaux dans le cadre général de communications sans fil de cinquième génération (5G). Les œuvres présentées dans cette thèse reflètent nos études dans deux domaines d'importance centrale pour la réalisation de la 5G : la sécurité et l'allocation des ressources.

En ce qui concerne la sécurité, de nouveaux défis sont apparus en 5G avec le paradigme de l'Internet des objets (IoT) et les communications device to device (D2D) à faible latence. Les nouvelles verticales, telles que l'haptique et les communications véhiculaires (V2X), nécessitent une faible complexité et des mécanismes de sécurité à faible latence, en particulier dans le contexte de l'authentification. Dans cette thèse, des solutions d'authentification de légère complexité en utilisant des fonctions physiques inclonables (PUF) et des générations de clés secrètes (SKG) à la couche physique sont présentées.

En outre, comme le contenu vidéo est responsable de plus de 70% du trafic IP mondial, il est important que les infrastructures de diffusion de contenu détectent et répondent rapidement aux changements de la dynamique de popularité du contenu. Dans cette thèse, nous proposons une approche flexible d'allocation des ressources qui tire parti des technologies unikernel et containers. L'allocation des ressources est entraînée par un détecteur de popularité de contenu en temps réel et à faible complexité, mis en œuvre à l'aide des analyses hors ligne et en ligne des points de changement. Des variantes de ces algorithmes ont des applications dans la détection d'intrusion dans les réseaux définis par les logiciels de capteurs sans fil, qui sont discutés ensuite.

Enfin, l'utilisation potentielle d'un accès multiple non orthogonal (NOMA) dans le lien ascendant sans fil est envisagée. Les premiers résultats de la comparaison des systèmes d'allocation NOMA par rapport aux schémas orthogonaux dans les régimes asymptotiques, montrent que les gains dans l'utilisation de NOMA se poursuivent dans le scénario des communications sous des contraintes statistiques de délai de qualité de service (QoS).

Acknowledgements

I would like to take this opportunity to thank wholeheartedly the *Telecom girls*, Inbar Fijalkow, Iryna Andriyanova, Veronica Belmega, Marwa Chafii, Laura Luzzi, and, also Mylène Pischella, Marine Moguen and Aymeric Histace for all their support during the last three years.

To my family

Blessed are the cheesemakers.

- The Life of Brian (1979)

Contents

List of Tables

List of Figures

Nomenclature

List of Abbreviations

0-RTT Zero round trip time

3GPP The 3rd Generation Partnership Project

5G Fifth generation

6G Sixth generation

AE Authenticated encryption

AES GCM Advanced encryption standard Galois counter mode

ARMA Autoregressive moving average model

B5G Beyond 5G

BF-AWGN Block fading additive white Gaussian noise

CRP Challenge response pair

CSI Channel state information

DMT Detection median time

DR Detection rate

EAP-TLS Extensible authentication protocol - transport layer security

EC Effective capacity

EH Energy harvesting

FDFF False data flow forwarding

FNI False neighbour information

FNR False negative rate

FPR False positive rate

FTN Faster than Nyqist

GARCH Generalized autoregressive conditional heteroskedasticity

HMAC Hashed message authentication code

ID Identification

IFFT Inverse fast Fourier transform

IoT Internet of things

LRT Likelihood ratio test

MA Multiple access

MAC Media access control

MiM Man in the middle

mMTC massive machine type communications

NB-IoT Narrow band IoT

NOMA Non Orthogonal multiple access

OFDM Orthogonal frequency division multiplexing

OMA Orthogonal multiple access

PAM Pulse amplitude modulation

PHY Physical layer

PKE Public key encryption

PLS Physical layer security

PNC Physical layer network coding

PUF Physical unclonable function

QAM Quadrature amplitude modulation

QoS Quality of service

QoSec Quality of security

RAN Radio access network

RPL Routing protocol for low-power and lossy networks

RSA Rivest, Shamir, Adleman

RSS Received signal strength

SC Secrecy capacity

SDN Software defined networking

SDWSN Software defined wireless sensor network

SIR Signal to interference ratio

SKG Secret key generation

SLA Service level agreement

SNR Signal-to-noise ratio

STEK Session ticket encryption key

TLS Transport layer security

URLLC Ultra reliable low latency communication

V2X Vehicle-to-everything communication

WSN Wireless sensor network

Chapter 1

Activity Review

1.1 Motivation for Application for the HdR Diploma

With this thesis, I wish to submit my application for the Habilitation to Direct Research at CY - Cergy Paris Université. Currently, I am a Maître de Conférences at the Ecole Nationale Supérieure de l'Electronique et de ses Applications (ENSEA) in Cergy and in parallel I have a Visiting Research Fellow status at the Department of Electrical and Electronic Engineering of Princeton University in the USA and at the School of Computer Science and Electronic Engineering of the University of Essex in the UK.

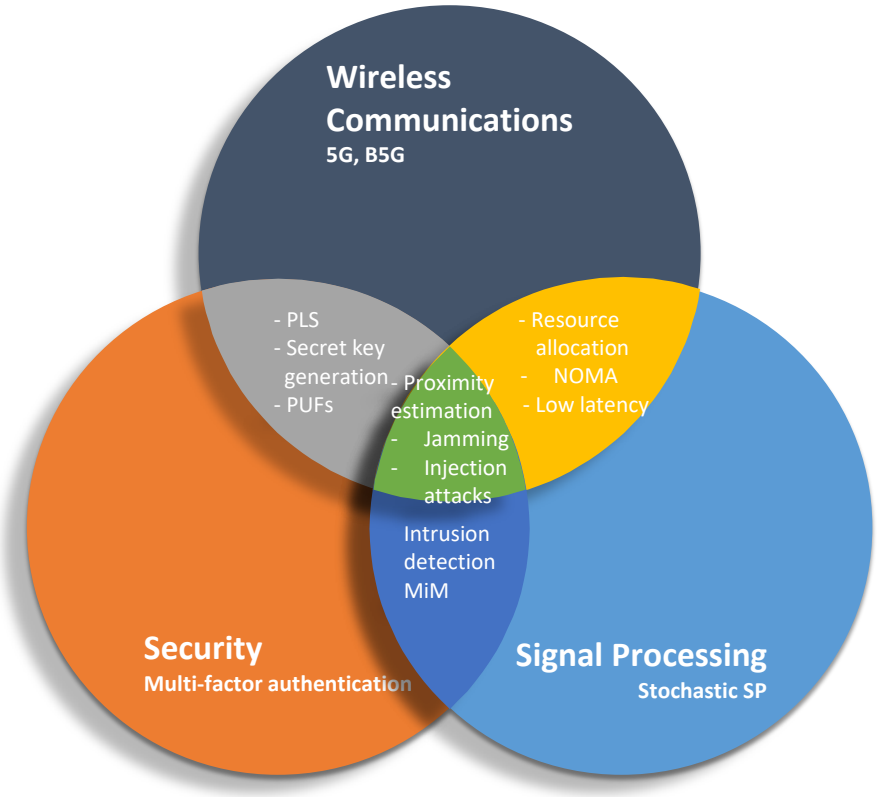


Figure 1.1: Recent research areas and topics

My current research activities relate to various topics in wireless communications and physical layer security with an emphasis on the proposal of low latency communication systems and the development of new security protocols for future generations of wireless. I actively work on topics in flexible numerology, non-orthogonal mutiple access (NOMA) and fast authentication protocols for

delay constrained systems using physical unclonable functions (PUFs) and RF fingerprinting. In this framework, with my current research team, that comprises four PhD students and two postdoctoral researchers, we investigate resource allocation in beyond fifth generation (B5G) leveraging NOMA, the efficient design of Slepian Wolf and Wyner Ziv reconciliation decoders at the short block-length, the development of zero-round-trip-time (0-RTT) authentication protocols using resumption keys generated from wireless fading coefficients, the analysis of the wireless channel secrecy capacity under statistical delay quality of service (QoS) constraints and the development of quick anomaly detection algorithms for software defined networks.

My research lies at the interface of wireless communications, signal processing and security studies, as depicted in Fig. 1.1; at this – not so-frequented – scientific crossroad, new engineering problems are encountered, a few of which will be discussed in later chapters of this thesis, along with proposed solutions.

With respect to my contribution as an academic teacher and supervisor, I have a long experience in teaching and supervising students in security, coding and wireless communications for more than 8 years in the UK and France. I have had the chance to teach a variety of courses both at the undergraduate and graduate level and contribute in teaching in the continuous education engineering track of ENSEA. I have taught to a variety of class sizes and have customarily received very positive feedback from my students, both in formal assessment and in face-to-face interaction. Furthermore, since last September I am acting as the liaison of the international mobility for ENSEA students towards the UK and have secured internships at Imperial College London, the University of York, etc.

In my academic employment I have had the opportunity to undertake a number of important administrative responsibilities. I currently head the research team ICI (information, communications, imaging) of the ETIS Lab that comprises 13 permanent faculty (3 PU, 8 MCF, 2 CNRS CR) and more than 18 research students and teaching fellows. In this role my aim is to help maintain and enhance the quality and quantity of the team's collective research output, its ability to attract research funding and good young researchers, increase further the team's visibility in the national and international level and ensure the team members work in a friendly and fertile learning environment. Furthermore, during my employment at the School of Computer Science and Electronic Engineering in the UK, I have acted in 2017 as the President of the Athena Swan Committee, steering 15 faculty and admin staff for the preparation of the department's gender equality and diversity charter.

With respect to my involvement in professional bodies, I am a member of the IEEE INGR Roadmap Security Workgroup, of the IEEE P1940 Standardization Workgroup on "Standard profiles for ISO 8583 authentication services" and have been a member of the IEEE Teaching Awards Committee for the last three years.

I feel that my overall experience is of sufficient standing to allow me to lead independent research and act as a stand-alone thesis advisor / director. A brief overview of my research and teaching activities the last 8 years is provided in Sections 1.4 and 1.5 respectively. First, I introduce myself to the reader through a detailed academic CV in Section 1.2. and the full list of my publications in Section 1.3.

1.2 Curriculum Vitae

In the following pages my full academic Curriculum Vitae is provided, including a full record of my publications in Section 1.3. The interested reader may also consult my web page and my google scholar page.

Dr. Arsenia Chorti

Address: Room 341, ENSEA, 6 Avenue du Ponceau, Cergy, FR
Telephone: +33 (0)769113367
e-mail: arsenia.chorti@ensea.fr achorti@princeton.edu

1.2.1 Current Position / Responsibilities (in chronological order)

Sep. 2017-present: **ENSEA (ETIS UMR8051) Associate Professor (MCF)** in Communications and Networking, Research Group: **4 PhD students, 2 postdocs**

Sep. 2017-Jul. 2020: **Member of** the IEEE Teaching Awards Committee

Sep. 2019 – present: **PEDR** (prime d'encadrement doctoral et recherche) – premium for excellence in supervision and research

Sep. 2019 – present: **Member of the IEEE P1940 Standardization Workgroup** on "Standard profiles for ISO 8583 authentication services"

April 2020-present: **Head of the Information, Communications and Imaging (ICI) Group of ETIS UMR 8051** (*Responsable d'équipe information, communications et imagerie*), comprising 3 Professors, 8 MCF, 2 CNRS Researchers, 2 Postdocs, 2 ATERs, 14 PhD students

Apr. 2020: Award of **CNRS delegation** (**half year travel sabbatical**) to visit Prof. H.V. Poor (Princeton University, NJ USA) and Dr. A. Barolo (Barkhausen Institute, Dresden DE) in Spring / Summer 2021

May 2020: **Member of the IEEE International Network Generations Roadmap (INGR) Security Workgroup** (pre-standardization workgroup for security in future networks)

June 2020: Elevated to **IEEE Senior Member**

1.2.2 Research Interests

My current research spans the areas of wireless security and beyond fifth generation (B5G) networks. I work on the design of security schemes for B5G, with a particular focus on physical layer security; my recent contributions concern fast authentication protocols using physical unclonable functions (PUFs) and secret key generation (SKG) from shared randomness, with proximity / localization as an extra authentication factor. Furthermore, I work on low latency communications, leveraging recent results on non-orthogonal multiple access (NOMA), investigate polynomial complexity algorithms for flexible numerology and eMBB – URLLC coexistence and joint PHY-MAC resource allocation optimization using the theory of the effective capacity. Recent contributions (since 2017) include:

- **Wireless security for B5G and Internet of things (IoT)** [J19], [J21], [C37], [C33], [S2]
- **Authentication protocols** leveraging PUFs, SKG and proximity estimation [BC3], [U1]
- **Resource allocation in 5G** using change point analysis [J17], [J20], [C32]
- **Anomaly detection** in software defined networks [C39], [J18], [S1], [U2]
- **Active attacks** in PHY [J15], [J16], [C36], [C28-C31] [BC2]
- **Low latency B5G communications**, non-orthogonal multiple access (**NOMA**), **NOMA-R**, **flexible numerology** for B5G [J22], [U3-U5]

1.2.3 Education

2000-2005 **Imperial College London:** Department of Electrical and Electronic Engineering
Ph.D. in Communications and Signal Processing
Thesis Title: *"The Impact of Circuit Nonlinearities and Noise in OFDM Receivers*", Supervisor: Mike Brookes, Scholarship awarded by I.K.Y. and Panasonic UK ltd.

1999-2000 **Université Pierre et Marie Curie – Paris VI**
MSc (D.E.A.) in Electronics
Dissertation Title: *"F.P.G.A. Implementation of Multi-Layer Perceptron Neural Network for Real-Time Applications in High Energy Physics",* Supervisor: Prof. Patrick Garda, I.K.Y. Scholar

1992-1998 **University of Patras:** Department of Electrical and Computer Science Engineering
M.Eng (Diploma) in Electrical Engineering
Dissertation Title: *"Development of User-Friendly Interface for the Testing of Nodal Cards of an Industrial Network"*, Supervisor: Prof. K. Koumbias

1.2.4 Academic Employment

- September 2017 – present: **ENSEA**, **Associate Professor in Communications and Networks**
- October 2013 – August 2017: **University of Essex**, School of Computer Science and Electrical Engineering, **Lecturer in Communications and Networks** and subsequently **Visiting Research Fellow** (ongoing)
- November 2012 – October 2013: **Foundation for Research and Technology Hellas (FORTH)**, Institute of Computer Science, **International Outgoing Fellow (IOF) Marie Curie Research Fellow**
- May 2011 – present: **Princeton University**, Dep. of Electrical Engineering, **IOF Marie Curie Fellow** and subsequently **Visiting Research Fellow**
- December 2008 – April 2011: **Middlesex University UK,** School of Engineering and Information Sciences, Dep. of Computer Communications, **Senior Lecturer in Communications and Networks**
- October 2007 – September 2009: **University College London (UCL),** Dep. of Electronic and Electrical Engineering, **Postdoctoral Research Fellow** and subsequently **Visiting Researcher**
- October 2006 – September 2007: **Technical University of Crete (TUC),** Department of Mineral Resources Engineering, Resources Detection and Identification Research Unit, **Postdoctoral Research Fellow**
- October 2005 – September 2006: **University of Southampton,** School of Electronics and Computer Science, Electronic Systems Design Group (ECS), **Postdoctoral Research Fellow**

1.2.5 Research Funding and Grants

Project proposals currently under review:

- **Principal Investigator (PI) of ANR PRCE project HERCULES** (enhancement measures in the security of beyond fifth generation networks) **2nd round AAPG 2020**, with the SME Montimage, K. Salamatian (LISTIC), I. Andriyanova (ETIS), A. Histace (ETIS) and F. Ghaffari (ETIS)
- **External collaborator** of project **LEON** (Intelligent Network Softwarization for the Internet of Things), ELIDEK, GR, with Dr. L. Mamatas
- **PI** project **PROCOPE PHC** (travel grant) to visit the Barkhausen Institute, DE in 2021-2022

Ongoing projects:

- **Co-investigator (co-I)** project **PHEBE** (Physical layer security for beyond fifth generation communications) with L. Wang (PI), L. Luzzi, M. Chafii, M. Le Treust**, Paris-Seine Excellence Initiative, 2020-2024**, 400,000€
- **PI** project **SAFEST with F. Jardel (NOKIA Bell Labs)** (Physical layer security for future generations wireless systems)**, DIM RFSI**, **2019-2021**, 27,500€
- **PI** project **eNiGMA** (Non-orthogonal multiple access techniques under security and delay constraints)**,** with I. Fijalkow, **Paris-Seine Excellence Initiative**, **2019-2021**, 110,000€
- **Co-I** project **ELIOT** (Enabling technologies for IoT), **ANR PRCI with Univ. Sao Paolo, Brazil**, with V. Belmega (PI), I. Andriyanova, I. Fijalkow, J. Lorandel, Role: **Leader of WP on IoT security**, **2019-2023**, ETIS: 390,420€ (total of 740 k€)

Past projects:

- **PI** SRV-ENSEA de l'Institut des Etudes Avancées Université Paris Seine, **2018-2019**: 3,000€
- **PI** SRV-ENSEA Institut des Etudes Avancées Université Paris Seine: **2017-2018** : 2,850€
- **PI** project **PHOTINO**, University of Essex, Research and Innovation Fund: **2014-2015**: £13,000

- **Co-I FP7 PEOPLE** Marie Curie IOF, project **APLOE with H.V. Poor (Princeton University)**, 245,448€, **2010-2013**
- PG Scholarship from the **State Scholarships Foundation of Greece–I.K.Y. 2000-2004**: £41,820

1.2.6 Teaching and Related Responsibilities at ENSEA (since 2017)

- 2019-present: **Responsible of student international mobility to the UK,** 2nd year MEng, 3rd year MEng, Erasmus programme with the UK
- 2019-present **Instructor in the MSc (M2R) module "Cryptography and Network Security", University Cergy Pontoise,** Master 2 Informatique et Ingénierie des Systèmes Complexes (IISC), specilization SIC (Signal, Information, Communications),
- 2018-present: **Responsible of the module** "**Network security**" 3rd year MEng, ENSEA
- 2018-present: **Responsible of module** "**Interconnexion réseaux**" 3rd year Cycle par Alternance, ENSEA
- 2017-present: **Responsible of the Option Internet of Things "Option IoT"**, 2nd year MEng, ENSEA
- 2017-present: **Instructor** "**IoT Security**", 2nd year MEng, ENSEA
- 2017-present: **Responsible of the module** "**Internetworking**", 3rd year MEng, ENSEA
- 2017-present: **Instructor** "**Wireless Communications**", 3rd year MEng, ENSEA
- 2017-present: **Lab instructor** in various courses, including Digital Communications, Internetworking, Signals and Systems, etc.

1.2.7 Research Supervision

Current supervision

- **PhD Student Mr. Miroslav Mitev: supervision @60%**, 25/4/2017-9/2020, "*Physical layer security for the Internet of things*", co-supervised with Dr. M. Reed, University of Essex, UK, Thesis VIVA (defence) scheduled for September 2020, publications: [J21], [C37], [C36], [C33], [P1], [U1]
- **PhD student Mr. Sotiris Skaperas: supervision @40%**, 1/9/2017-9/2020, "*Data analysis and forecasting models for flexible resource management in 5th generation networks*", co-supervised with Dr. L. Mamatas, University of Macedonia in Thessaloniki, GR, Thesis defence scheduled for September 2020, publications: [J20], [J17], [C32], [U5]
- **PhD student Mr. Gustavo Alonso Nunez Segura: supervision @35%**, 1/2/2019-projected to finish in 1/2022 (4-year thesis programme in Brazil), "Cooperative *Intrusion Detection System for Software Defined Wireless Sensor Networks*", co-supervised with Prof. Cintia Borges Margi, University of Sao Paolo, Brazil, publications: [J18], [C39], [C35], [S1], [U2]
- **PhD student Mr. Mouktar Bello: supervision @70%**, 1/11/2020-projected to finish in 10/2023, "*Meeting delay and security constraints in 6G wireless networks*", co-supervised with Prof. I. Fijalkow, ETIS/ENSEA, FR, publications: [C38], [U3, U4]
- **Postdoc Dr. Mahdi Shakiba Herfeh: supervision @100%**, 21/11/2019-20/5/2021 (fixed term 1.5 years), "*Physical layer security for IoT applications*", project ELIOT ANR PRCI, ETIS/ENSEA FR, publications: [BC3], [U1]
- **Postdoc Dr. Nasim Ferdosian: supervision @90%**, 1/1/2020-31/12/2021 (fixed term 2 years), "Non-orthogonal multiple access techniques under security and delay constraints", with Prof. I. Fijalkow, ETIS/ENSEA, FR, publications: [U5]

Past supervision

- **MSD (Master by Thesis – full year research project) student Cornelius Saiki: supervision @84%**, 1/9/2014-31/8/2015, "*A Novel Physical Layer Key Generation and Authenticated Encryption Protocol Exploiting Shared Randomness*", co-supervised with Prof. S. Walker, University of Essex, publications: [C27]

- **MSc (M2R) SIC student Gada Rezgui: supervision @50%**, 1/3/2017-31-8-2017, "*Energy Harvesting as a Means to Mitigate Jamming Attacks; a Game Theoretic Analysis"*, co-supervised with V. Belmega, ETIS/University of Cergy Pontoise, publications: [J16]
- **MSc (M2R) SIC student Rihem Nasfi: supervision @100%,** 1/11/2018-15/3/2019, Projet d'Initiation à la Recherche (PIR), *"Non-orthogonal multiple access networks under QoS delay constraints",* publications : [C34]
- **MSc (M2R) SIC student Gada Rezgui**, **supervision @50%**, 1/11/2016-15/3/2017, Projet d'Initiation à la Recherche (PIR), *"Secret Key Generation systems under Jamming Attacks via Game Theoretic Tools"*
- **MSc (M2R) IMD student Amani Gran**, **supervision @100%**, 1/11/2018-15/3/2019, Projet d'Initiation à la Recherche (PIR), *"IoT lightweight security"*
- **MSc (M2R) SIC student Fatiha Ait Larbi**, **supervision @100%**, 1/11/2018-15/3/2019, Projet d'Initiation à la Recherche (PIR), *"Cross-layer security protocol design"*
- **MSc (M2R) SIC student Mouad Nahri**, **supervision @100%**, 1/11/2019-15/3/2020, Projet d'Initiation à la Recherche (PIR), *"Flexible numerology for B5G"*
- **Other MSc/BSc supervision:** 5 MSc and 9 BSc dissertations at the University of Essex and more than 10 MSc and BSc dissertations at Middlesex University

1.2.8 Recruitment (Selection) Committees / Thesis Examiner

- May 2020: Recruitment Committee (**Comité de sélection**) **for a MCF** post at CY Cergy University on *Networks and Security*
- Sep. 2019: Recruitment Committee (**Comité de sélection**) **for a MCF** post at EISTI on *Cybersecurity*
- Jun. 2020: **Thesis Examiner (rapporteur),** A. Ben Hadj Fredj, Télécom ParisTech, supervisors Prof. G. Rekaya and Prof. J-C Belfiore, "Computations for Multiple Access Channels in Wireless Networks"
- Jan. 2019: **Thesis Reviewer,** L. Senigagliesi, Univ. Polytechnica delle Marche, supervisors Prof. L. Spalazzi and Prof. M. Baldi, "Information-theoretic security techniques for data communications and storage"
- Aug. 2014: **Thesis Examiner,** I. K. Musa, CSEE University of Essex UK, supervisor Prof. S. Walker, "Optimized Self-Service Resource Containers for Next Generation Cloud Delivery"

1.2.9 Workshop Organization / Keynotes / Tutorials

- **Tutorial** on "What Physical Layer Security Can Do for 6G", **IEEE Global Communications (GLOBECOM) 2020**, A. Chorti and V.H. Poor, December 2020, Taipei TW.
- **Tutorial** on "Statistical methods in physical layer security", **IEEE Statistical Signal Processing (SSP) Workshop**, July 2020, Rio de Janeiro, BR (*rescheduled to July 2021 due to COVID-19*)
- **Special Session Organizer**, "Selected topics on 6G security", **IEEE ISWCS**, Sep. 2020, Berlin, Germany (*rescheduled to Sep. 2021 due to COVID-19*)
- **Special Session Organizer**, "Statistical Methods for IoT", **IEEE SSP 2020**, Jul. 2020, Rio de Janeiro, Brazil (*postponed to July 2021 due to COVID-19*)
- **Training School Co-organizer** (with M. Chafii, S. Stanczak and R. Cavalcante), **"**Machine Learning for Communications", 3-4 Sep. 2020, Berlin (co-located with ISWCS, *rescheduled to Sep. 2021 due to Covid-19*)
- **Chair of the GdR ISIS Workshop** "Women in Communications, Information Theory and Signal Processing", May 19 2020 (*rescheduled to May 2021 due to Covid-19*)
- **Chair of the GdR ISIS Workshop** "Enabling ultra-reliability, low latency and massive connectivity", June 18 2020 (*virtual event due to Covid-19*)
- **Keynote IEEE PIMRC Workshop Security Public RATs**: "Practical examples of physical layer security", 4 Sep. 2016, Valencia, Spain

- **Chair** of the **workshop ACCESS** - Cutting edge topics in physical layer security, communications and distributed storage workshop, 11 May 2014, Aalborg, Denmark
- **Co-chair of** "2nd Women's Workshop on Communications and Signal Processing", 16-18 July 2014, Princeton NJ, US
- **Track chair of the IEEE Global Wireless Summit 2014**, 11-14 May 2014, Aaborg, Denmark
- **Chair** of the "Second International Conference on Communications, Connectivity, Convergence, Content and Cooperation", 11-14 May 2014, Aalborg, Denmark
- **Chair** of the "WirelessVITAE, 10-13 May 2014, Aalborg, Denmark

1.2.10 Editor / Reviewer / Selected TPCs

- 2020- present: **Associate Editor** of the **IEEE Open Journal on Signal Processing** (OJSP)
- Sep. 2019-present: **Lead Guest Editor, EURASIP JWCN Special Issue** "Physical layer security solutions for 5G-and-beyond", Editors: S. Tomasin, H.V. Poor, M. Baldi, S. El Ruayheb, X. Wang, to appear in 2020
- 2018-2019: **Executive Editor Transactions on Emerging Telecommunications Technologies** (ETT), *Wiley*
- 2017-2019: **Executive Editor** of **Internet Technology Letters** (ITL), *Wiley*
- **Reviewer**: IEEE Transactions (Trans.) on Information (Inf.) Forensics and Security, Elsevier Computers and Security, IEEE Trans. on Wireless Communications (Commun.)., IEEE Trans. Signal Processing, IEEE Trans. Vehicular Technologies, IEEE JSAC, IEEE Wireless Commun. Letters (L.), IEEE Commun. L., Trans. on Emerging Telecom Tech. (ETT), Eurasip JWCN, IEEE Trans on Commun., …
- **TPCs**: more than 30 TPCs, indicatively IEEE GLOBECOM 2015, 2016, 2017, 2018, 2019, 2020, IEEE ICC 2014, 2015, 2016, 2018, 2019, 2020, IEEE WCNC 2016, 2019 (executive member), …

1.2.11 Selected Invited Talks (after 2016)

- July **2019**, "Physical layer security in delay constrained applications", **NOKIA Bell Labs**, FR
- May **2019**, "Physical layer security in delay constrained applications", **Barkhausen Institute**, Dresden DE
- May **2019**, "Physical layer security in delay constrained applications", **ICS FORTH**, GR
- October **2017,** "Emerging security paradigms", **Thales**, FR
- March **2017,** "Physical layer security for future networks", **British Telecom**, Adastral Park, UK
- June **2016,** "Practical examples of physical layer security", **Summer Research Institute, EPFL**, CH

1.2.12 Past Administrative Responsibilities and Outreach Activities

- 2016-2017: **President** of the Committee for Gender Equality and Diversity *Athena Swan*, Univ. Essex, UK
- 2016-2017: **Vice-president** "Research Student Progress and Management Committee", Univ. Essex, UK
- 2015-present: **Fellow of the Higher Education Academy**, UK (professional title in pedagogical training)
- 2014-2015: **Organizer** of student recruitment activities "Visit Days", Univ. Essex, UK

1.3 Publication List

1.3.1 Books [B] / Book Chapters [BC]

(supervised students and postdocs appear underlined)

BC3 M. Shakiba Herfeh, **A. Chorti**, V.H. Poor, *A Review of Recent Results on Physical Layer Security*, to appear in Springer Nature 2020;

BC2 **A. Chorti**, *A Study of Injection and Jamming Attacks in Wireless Secret Sharing Systems*, (Proc. 2nd Workshop Communication Security, WCS 2017), Lect. Notes in Elect. Eng., vol 447, pp. 1-14, Springer;

BC1 **A. Chorti**, C. Hollanti, J.-C. Belfiore, H.V. Poor, *Physical Layer Security: A Paradigm Shift in Data Confidentiality*, Springer, Lecture Notes in Electrical Engineering - Physical and Data-Link Security Techniques for Future Communication Systems, vol. 358, pp. 1-15, Sep. 2015;

B **A. Chorti**, *The Impact of Circuit Nonlinearities and Noise in OFDM Receivers*, Feb. 2010, Verlag

1.3.2 Refereed International Journals [J]

(supervised students and postdocs appear underlined)

J22 M. Pischella, **A. Chorti**, I. Fijalkow, "On the Performance of NOMA-Relevant Strategies Under Statistical Delay QoS Constraints", *IEEE Wireless Commun. Letters*, in print (early access);

J21 M. Miroslav, **A. Chorti**, M.J. Reed, L. Musavian, "Authenticated Secret Key Generation in Delay Constrained Wireless Systems", *EURASIP J Wireless Com Network*, vol. 122, Jun. 2020;

J20 S. Skaperas, L. Mamatas, **A. Chorti**, "Real-Time Algorithms for the Detection of Changes in the Variance of Video Content Popularity", *IEEE Access*, vol. 8, pp: 30,445-30,457, Feb. 2020;

J19 W. Yu, **A. Chorti**, L. Musavian, V.H. Poor, Q. Ni, "Effective Secrecy Capacity for a Downlink NOMA Network", *IEEE Trans. Wireless Commun.*, vol. 18, no 12, pp: 5,673-5690, Dec. 2019;

J18 G.A. Nunez Segura, C. B. Margi, **A. Chorti**, "Understanding the Performance of Software Defined Wireless Sensor Networks Under Denial of Service Attack", *Open Journal of Internet of things (OJIOT)*, Vol.5, no 1, pp:59-68 Aug. 2019 (published in the OJIOT as a special issue);

J17 S. Skaperas, L. Mamatas, **A. Chorti**, "Real-Time Video Content Popularity Detection Based on Mean Change Point Analysis", *IEEE Access*, vol.7 pp: 142,246-142,260, Jul. 2019;

J16 G. Rezgui, E.V. Belmega, **A. Chorti**, "Mitigating Jamming Attacks Using Energy Harvesting", *IEEE Wireless Commun. Let.*, vol. 8 no 1, pp: 297-300, Feb. 2019;

J15 E.V. Belmega, **A. Chorti** "Protecting Secret Key Generation Systems against Jamming: Energy Harvesting and Channel Hopping Approaches", *IEEE Trans. Inf. Forensics Security*, vol. 12, no 11, pp: 2611-2626, Nov. 2017;

J14 D. Karpuk, **A. Chorti**, "Perfect Secrecy in Physical-Layer Network Coding Systems from Structured Interference", *IEEE Trans. Inf. Forensics Security*, vol. 11, no 8, pp. 1875-1887, Aug. 2016;

J13 **A. Chorti**, K. Papadaki, H.V. Poor, "Optimal power allocation in block fading channels with confidential messages", *IEEE Trans. Wireless Commun.*, vol. 14, no 9, pp. 4708-4719, Sep. 2015;

J12 **A. Chorti**, S. Perlaza, Z. Han, H.V. Poor, "On the resilience of wireless multiuser networks to passive and active eavesdroppers", *IEEE Journal of Selected Areas in Commun.*, vol. 31 no 9, pp. 1850-1863, Sep. 2013;

J11 **A. Chorti**, M. Brookes, "On the effect of Voigt profile oscillators on OFDM systems", *IEEE Trans. Circuits Syst. II*, vol. 58, no 11, pp. 768-772, Nov. 2011;

J10 G. Spiliopoulos, D.T. Hristopulos, M.P. Petrakis, **A. Chorti**, "A multigrid method for the estimation of geometric anisotropy in environmental data from sensor networks", Elsevier *Computers and Geosciences*, vol. 37, no 3, pp. 320-330, Mar. 2011;

J9 **A. Chorti**, M. Brookes, "Performance Analysis of COFDM and DAB Receivers in narrow-band and tonal interference", Springer *Telecommunication Systems J.*, vol. 46, no 2, pp. 181-190, 2011.

J8 Y. Kanaras, **A. Chorti**, M. Rodrigues, I. Darwazeh, "A fast constrained sphere decoder for ill conditioned communication systems", *IEEE Commun. Let.*, vol. 14, no 11, pp. 999-1001, Nov. 2010.

J7 **A. Chorti**, "How to model the near-to-the-carrier regime and the lower knee frequency of real RF oscillators", *J. Electrical Computer Eng.*, vol. 2010, article ID 537132, Oct. 2010.

J6 **A. Chorti**, D.T. Hristopulos, "Non-parametric identification of anisotropic correlations in spatially distributed data sets", *IEEE Trans. Signal Proces*, vol. 56, no 10, pp. 4738-4751, Oct. 2008.

J5 D. Karantzas, **A. Chorti**, N.M. White, C.J. Harris, "Teaching old sensors new tricks: archetypes of intelligence", *IEEE Sensors J.*, Special Issue on Intelligent Sensing", invited paper, vol. 7, no 5, pp. 868-881, May 2007.

J4 **A. Chorti**, D. Karatzas, N.M. White and C.J. Harris, "Intelligent Sensors in Software: The Use of Parametric Models for Phase Noise Analysis", *International Journal of Information Processing*, vol. 1, no. 2, June 2007.

J3 **A. Chorti**, D. Karantzas, N.M. White and C.J. Harris, "Use of the extended Kalman filter for state dependent drift estimation in weakly nonlinear sensors", *Sensors Let.*, vol. 4, no 4, pp. 377-379, Dec. 2006.

J2 **A. Chorti**, M. Brookes, "A spectral model for RF oscillators with power-law phase noise, *IEEE Trans. Circuits Syst. I*", vol. 53, no 9, pp. 1989-1999, Sep. 2006.

J1 **A. Chorti**, M. Brookes, "On the effects of memoryless nonlinearities on M-QAM and DQPSK OFDM Signals", *IEEE Trans. Microw. Theory Techn.*, vol. 54, no 8, pp. 3301-3315, Aug. 2006.

1.3.3 Refereed International Conference Proceedings [C]

(supervised students and postdocs appear underlined)

C39 G.A. Nunez Segura, S. Skaperas, **A. Chorti**, L. Mamatas, C. Borges Magri, "Denial of Service Attacks Detection in Software-Defined Wireless Sensor Networks", Proc. *IEEE Int. Conf. Commun. (ICC) Worskhop on SDN Security*, Dublin UK, 7-11 Jun. 2020;

C38 B. Mouktar, W. Yu, **A. Chorti**, L. Musavian, "Performance Analysis of NOMA Uplink Networks under Statistical QoS Delay Constraints", Proc. *IEEE Int. Conf. Commun. (ICC)*, Dublin UK, 7-11 Jun. 2020;

C37 M. Mitev, **A. Chorti**, M.J. Reed "Subcarrier Scheduling for Joint Data Transfer and Key Generation Schemes in Multicarrier Systems", Proc. *IEEE Int. Global Commun. Conf. (GLOBECOM)*, Hawaii US, 9-13 Dec. 2019;

C36 M. Mitev, **A. Chorti**, E.V. Belmega, M.J. Reed "Man-in-the-Middle and Denial of Service Attacks in Wireless Secret Key Generation", Proc. *IEEE Global Commun. (GLOBECOM)*, Hawaii US, 9-13 Dec. 2019;

C35 G.A. Nunez Segura, C. B. Margi, **A. Chorti** , "Understanding the Performance of Software Defined Wireless Sensor Networks Under Denial of Service Attack", Proc. *Int. Workshop on Very Large IoT (VLIoT)* 2019, Los Angeles, US, 30th Aug. 2019 (*invited paper);

C34 R. Nasfi, **A. Chorti**, "Performance Analysis of the Uplink of a Two User NOMA Network under QoS Delay Constraints", Proc. *IEEE Int. Conf. on Ubiquous and Future Networks (ICUFN)* 2018, Zagreb, Croatia, 2-5 July 2019;

C33 M. Mitev, **A. Chorti**, M.J. Reed "Optimal Resource Allocation in Joint Secret Key Generation and Data Transfer Schemes", Proc. *IEEE Int. Conf. Wireless Commun. Mobile Comput. (IWCMC)*, Tangiers Morocco, 24-28 June 2019;

C32 S. Skaperas, L. Mamatas, **A. Chorti**, "Early Video Content Popularity Detection with Change Point Analysis", Proc. *IEEE Int. Global Commun. (GLOBECOM)*, Abu Dhabi, UAE, 6-11 December 2018;

C31 E.V. Belmega, **A. Chorti**, "Energy Harvesting in Secret Key Generation Systems under Jamming Attacks", Proc. *IEEE Int. Conf. Commun. (ICC)*, Paris, France, May 2017;

C30 **A. Chorti**, "Secret Key Generation in Rayleigh Block Fading AWGN Channels under Jamming Attacks", Proc. *IEEE Int. Conf. Commun. (ICC)*, Paris France, May 2017;

C29 **A. Chorti**, "Optimal Signalling Strategies and Power Allocation for Wireless Secret Key Generation Systems in the Presence of a Jammer", Proc. *IEEE Int. Conf. Commun. (ICC)*, Paris, France, May 2017;

C28 **A. Chorti**, "Overcoming limitations of secret key generation in block fading channels under active attacks", Proc. *IEEE 17th Int. Workshop Signal Process. Advances Wireless Commun. (SPAWC)*, pp. 1-5, Jul. 2016 (*invited paper);

C27 C. Saiki, **A. Chorti**, "A novel authenticated encryption protocol exploiting shared randomness", Proc. *IEEE Commun. Network Security (CNS)*, 2nd Workshop on Physical Layer methods for Wireless Security, pp. 651-656, Sep. 2015;

C26 **A. Chorti**, M.M. Molu, D. Karpuk, C. Hollanti, A. Burr, "Strong secrecy in wireless network coding systems with M-QAM modulators", Proc. *IEEE Int. Conf. Commun. China (ICCC)*, pp. 181-186, Oct. 2014;

C25 **A. Chorti**, K. Papadaki, H.V. Poor, "Optimal power allocation in block fading Gaussian channels with causal CSI and secrecy constraints", Proc. *IEEE Global Commun. (GLOBECOM)*, pp. 752-757, Dec. 2014;

C24 S.M. Perlaza, **A. Chorti**, H.V. Poor, Z. Han, "On the trade-offs between networks state knowledge and secrecy", Proc. *IEEE Int. Symp. Wireless Personal Multimedia Commun. (WPMC)*, pp. 1-6, Jun. 2013;

C23 **A. Chorti**, K. Papadaki, P. Tsakalides, H.V. Poor, "The secrecy capacity of block fading multiuser wireless networks", Proc. *IEEE Int. Conf. Adv. Tech. Commun. (ATC)*, pp. 247-251, Oct. 2013, (*best paper award);

C22 S.M. Perlaza, **A. Chorti**, H.V. Poor and Z. Han, "On the impact of network-state knowledge on the feasibility of secrecy", Proc. *IEEE Int. Symp. Inf. Theory (ISIT)*, pp. 2960-2964, Istanbul, Turkey, Jul. 2013;

C21 **A. Chorti**, S. Perlaza, Z. Han, H.V. Poor, "Physical layer security in wireless networks with passive and active eavesdroppers", Proc. *IEEE Global Commun. (GLOBECOM)*, Anaheim, USA, 3-7 Dec. 2012;

C20 **A. Chorti**, "Helping interferer physical layer security strategies for M-QAM and M-PSK systems", Proc. *IEEE CISS* 2012, Princeton NJ, USA, 21-23 Mar. 2012;

C19 **A. Chorti** and V. Poor, "Achievable secrecy rates in physical layer security systems with a helping interferer", Proc. *IEEE Int. Conf. Comp. Netw. Commun. (ICNC)*, Maui, HI, Feb. 2012;

C18 **A. Chorti** and V. Poor, "Faster than Nyquist interference assisted secret communication for OFDM systems, *IEEE Asilomar*, Pacific Grove, CA, US, 4-7 Nov. 2011, (*invited paper);

C17 **A. Chorti**, "Masked M-QAM OFDM: Encryption of OFDM signals through faster than Nyquist signalling", Proc. *IEEE MCECN Global Commun. (GLOBECOM)*, Miami, US, 6-10 Dec. 2010;

C16 **A. Chorti**, Y. Kanaras, M. Rodrigues, I. Darwazeh, "Joint channel equalization and detection of spectrally efficient FDM signals", Proc. *IEEE Personal Indoor Multimedia Radio Commun. (PIMRC)*, Istanbul, Turkey, 26-29 Sep. 2010

C15 Y. Kanaras, **A. Chorti**, M. Rodrigues, I. Darwazeh, "A new quasi-optimal detection algorithm for a non-orthogonal spectrally efficient FDM system", Proc. *Int. Symp. Commun. Inf. Tech. (ISCIT)*, Incheon, Korea, 28-30 Sep. 2009;

C14 Y. Kanaras, **A. Chorti**, M. Rodrigues, I. Darwazeh, "An Overview of Optimal and sub-Optimal Detection Techniques for a Non Orthogonal Spectrally Efficient FDM", Proc. *LCS/NEMS*, London UK, 3-4 Sep. 2009;

C13 **A. Chorti**, Y. Kanaras, "Masked M-QAM OFDM: A simple approach for enhancing the security of OFDM systems", *IEEE Personal Indoor Multimedia Radio Commun. (PIMRC)*, Tokyo, Japan, 13-16 Sep. 2009;

C12 D.T. Hristopulos, M.P. Petrakis, G. Spiliopoulos, **A. Chorti**, "Non-parametric estimation of geometric anisotropy from environmental sensor network measurements", Proc. *StatGIS2009*, Milos, Greece, 17-19 Jun. 2009;

C11 Y. Kanaras, **A. Chorti**, M. Rodrigues, and I. Darwazeh, "Spectrally efficient FDM signals: bandwidth gain at the expense of receiver complexity", *IEEE Int. Conf. Commun. (ICC)*, Dresden, Germany, 13-17 Jun. 2009;

C10 Y. Kanaras, **A. Chorti**, M. Rodrigues, I. Darwazeh, "A near optimum detection for a spectrally efficient non orthogonal FDM system", Proc. *InOWo'08*, Hamburg Germany, 27-28 Aug. 2008;

C9 D.T. Hristopulos, **A. Chorti**, G. Spiliopoulos, E. Petrakis, "Systematic detection of anisotropy in spatial data obtained from environmental monitoring networks", *EGU2008*, Vienna, Austria, 13-18 Apr. 2008;

C8 Y. Kanaras, **A. Chorti**, M. Rodrigues, I. Darwazeh, "A combined MMSE-ML detection for a Gram-Schmidt orthogonalized FDM system", Proc. *IEEE BROADNETS*, London, UK, Sep. 2008;

C7 Y. Kanaras, **A. Chorti**, M. Rodrigues, I. Darwazeh, "Sub-optimum detection techniques for a bandwidth efficient multi-carrier communication system", *Mutli-Strand Conf.*, Milton, UK, 6-7 May 2008.

C6 **A. Chorti**, D.T. Hristopulos, "Automatic detection of spatial anisotropy in environmental data sets", Proc. *StatGIS2007*, Klagenfurt, Austria, Oct. 2007.

C5 A. Moustakas, **A. Chorti** and D.T. Hristopulos, "Geostatistical analysis of tree size distributions in the southern Kalahari, obtained from remotely sensed data", Proc. *SPIE Europe Remote Sensing*, Florence, Italy, 17-20 Sep. 2007.

C4 **A. Chorti** and M. Brookes, "Resolving near carrier spectral infinities due to 1/f phase noise in oscillators", Proc. *IEEE Int. Conf. Acoustics Speech Signal Process. (ICASSP)*, vol. 3, pp. III 1005-III 1008, Hawaii, USA, 15-18 Apr. 2007

C3 **A. Chorti**, D. Karatzas, N.M. White, C.J. Harris, "Intelligent sensors in software: the use of parametric models for phase noise analysis", Proc. *IEEE Int. Conf. Intelligent Sensing Inf.*, Bangalore, India, 15-18 Dec. 2006.

C2 **A. Chorti**, B. Granado, B. Denby, P. Garda, "Une architecture electronique temps reel pour les reseaux connexionnistes en physique des hautes energies", *NSI2000*, Toulouse FR, May 2000.

C1 **A. Chorti**, B. Granado, B. Denby and P. Garda, "An electronic system for the simulation of neural networks with real time constraints", Proc. *ACAT*, Chicago, U.S., Dec. 2000.

1.3.4 Posters

P2 M. Mitev, **A. Chorti**, M.J. Reed, "Physical layer security in wireless networks with active eavesdroppers", *Munich Workshop on Coding and Cryptography (MWCC) 2018*, (*invited poster), Germany, 10-11 April 2018;

P1 **A. Chorti**, "Optimal resource allocation in secure multi-carrier systems", *1st IEEE Women's Workshop Commun. Signal Proc.*, Banff, (*invited poster), Canada, 13-15 Jul. 2012.

1.3.5 In Preparation [U] / Submitted [S]

U1 M. Mitev, M. Shakiba Herfeh, **A. Chorti**, M.J. Reed, "Multi-factor lightweight authentication for the Internet of Things", IEEE Trans. Inf. Forensics Security, in preparation;

U2 G. A. Nunez Segura, **A. Chorti**, C. Borges Magri, "Multimetric centralized and decentralized intrusion detection in software defined networks", IEEE Internet of Things Journal, in preparation;

U3 M. Bello, W. Yu, M. Pischella, **A. Chorti**, I. Fijalkow, L. Musavian, "A Review of DL/UL Multiple Access Enabling Low-Latency Communications", IEEE Access, in preparation;

U4 M. Bello, **A. Chorti**, I. Fijalkow, W. Yu, L. Musavian, "Performance Analysis of NOMA Uplink Networks under Statistical QoS Delay Constraints", IEEE Trans. Communications, in preparation;

U5 N. Ferdosian, S. Skaperas, **A. Chorti**, L. Mamatas, "Unleashing the Potential of Flexible Numerology by Resolving Conflicts", IEEE Trans. Wireless Communications, in preparation;

S1 G.A. Nunez Segura, **A. Chorti**, C. Borges Magri, "Multimetric Online Intrusion Detection in Software-Defined Wireless Sensor Networks", in review IEEE Globecom 2020;

1.4 Recent Research Results

1.4.1 Motivation on Studying Physical Layer Security and Resource Allocation for 5G Systems

Physical Layer Security

The goal of physical layer security (PLS) [1–3] is to make use of the properties of the physical layer – including the wireless communication medium and / or the transceiver hardware – to enable critical security aspects. In particular, PLS can be employed to provide i) node (device) authentication, ii) message authentication, iii) message confidentiality through the use of secrecy encoders, and, iv) key management and distribution solutions through symmetric secret key generation from shared randomness. Furthermore, proposals for intrusion detection and counter-jamming at PHY have recently emerged [4]; indeed these two topics emerge as important research areas in B5G systems, particularly in the industrial Internet of things (IoT) and the mmWave era.

PLS has been explicitly mentioned in the first white paper on 6G: "The strongest security protection may be achieved at the physical layer". Importantly, it is stated as an enabling technology in the IEEE International Network Generations Roadmap 1st Edition 2019 in the Chapters on "Security" (Section 1.1 pp. 1-2) and on "Massive MIMO" (Section 4.3 pp. 8-9). The increasing interest in PLS has been stimulated by many practical needs. Notably, many critical IoT networks require ultra-low latency communications ($<$ 1msec), e.g., in autonomous driving and vehicle to everything (V2X) applications, telemedicine and haptics. However, standard authentication often requires significant processing time. We note in passing that in the Third Generation Partnership Project (3GPP) technical report "Study on the Security of URLLC" [5], all aspects related to low latency (fast) authentication remain open and no solutions have so far been standardized. An added complication is due to hardware limitations of low-end sensors and their ineptest to execute sophisticated security protocols such as the IPSec or the DTLS.

A further challenge comes from quantum computing, which has seen significant progress after massive investment by companies such as Google, Intel and IBM to build prototypes with more than 50 qubits. In October 2019 Google published in the journal "Nature" their quantum computer experiments showing they have achieved quantum supremacy for a particular set of problems [6]. In this aspect, PLS, that relies upon information-theoretic security proofs, could resist quantum computers, unlike corresponding asymmetric key schemes relying on the (unproven) intractability in polynomial time of certain algebraic problems. Even state-of-the-art elliptic curve cryptography (ECC) schemes, that require substantially shorter keys than RSA or Diffie Hellman (DH) schemes, are still considerably more intensive computationally than their PLS counterparts and are not post-quantum.

As a result, the study of novel PLS based solution for 5G and B5G security is highly pertinent. Related proposals using physical unclonable functions [7] and secret key generation from shared randomness [8] are included in this thesis.

Resource Allocation

The roll-out of fifth-generation (5G) mobile networks and the forthcoming 6G will bring about fundamental changes in the way we communicate, access services and entertainment.With respect to the latter, the multi-fold increase in the service data rates will provide users with ultra high resolution in video-streaming, multi-media and virtual reality, offering immersive experiences. To this end, it is important for Edge content delivery infrastructures to rapidly detect and respond to changes in content popularity dynamics. For flexible and highly adaptive solutions, the capability for quick resource (re-)allocation should be driven by early (*real-time*) and low-complexity content popularity detection schemes. In this thesis, we study aspects of low-complexity detection of changes in video content popularity in real-time, addressed as a statistical change point (CP) detection problem [9], breaking completely new ground compared to earlier works that relied upon prediction models [10], [11].

Furthermore, novel exciting use cases were introduced in 5G in the context of ultra-reliable low latency communications (URLLC) and massive machine type communications (mMTC); the new industrial revolution, dubbed as Industry 4.0, along with emerging verticals in telemedicine, smart agriculture, etc., will bring about automation and intelligence to levels never seen before.

As 5G is required to support a large variety of services, novel solutions to enable higher resource efficiency are sought; amongst the various possible solutions, in this thesis we study non-orthogonal multiple access (NOMA) because of its advantages over conventional orthogonal multiple access (OMA) schemes in terms of spectral efficiency [12], cell-edge throughput [13], and energy efficiency [14], rendering it an attractive solution in particular for the mMTC uplink scenario.

Additionally, to account for media access control (MAC) sub-layer latency, we use the theory of the *effective capacity* [15], which can serve in wireless networks to provide *statistical* delay guarantees. The pertinence of the theory of the effective capacity as a suitable metric results from the fact that in the wireless MAC, due to small scale fading and shadowing, it is *inherently* impossible to provide hard delay guarantees.

In the following, a brief presentation of my principal past contributions over the last 7 years is given in reverse chronological order, to emphasize more recent results. Section 1.4.2 offers an outline of recent results in the area of resource allocation using NOMA, the theory of the effective capacity and CP analysis, while results in the area of PLS are described in Section 1.4.3.

1.4.2 Results in Resource Allocation

NOMA and Effective Capacity

Related Contributions: [J22], [J19], [C38], [C34]

In our works a flexible delay quality of service (QoS) model was employed using the theory of large deviations (Gärtner-Ellis theorem [16]) that allows defining the metric of the effective capacity (EC) in block fading additive white Gaussian noise (BF-AWGN) channels. The EC denotes the maximum constant arrival rate that can be served by a given service process, while guaranteeing a required statistical delay provisioning and is closely related to the concept of the effective bandwidth [17]. In order to capture the impact of link layer (MAC) delays in the secrecy capacity of wireless BF-AWGN channels, we introduced a novel metric, referred to as the "effective secrecy rate" (ESR); the ESR represents the maximum constant arrival rate that can be *securely* served (with perfect secrecy), on the condition that the required delay constraint can be statistically satisfied.

In more detail, in [J19] a novel approach was introduced to study the achievable delay-guaranteed secrecy rate, focusing on the downlink of a NOMA network with one base station, multiple single-antenna NOMA users and an eavesdropper. Two possible eavesdropping scenarios were considered; an internal, unknown, eavesdropper in a purely antagonistic network and an external eavesdropper in a network with trustworthy peers. For a purely antagonistic network with an internal eavesdropper, the only receiver with a guaranteed positive ESR was proved to be the one with the highest channel gain. The ESR in the high signal to noise ratio (SNR) regime was shown to approach a constant value irrespective of the power coefficients, while the strongest user was shown to achieve a higher ESR when it had a distinctive advantage in terms of channel gain with respect to the second strongest user. For a trustworthy NOMA network with an external eavesdropper, a lower bound and an upper bound on the ESR were proposed and investigated for an arbitrary legitimate user. For the lower bound, a closed-form expression was derived in the high SNR regime. For the upper bound, the analysis showed that if the external eavesdropper could not attain any channel state information (CSI), the legitimate NOMA user at high SNRs would always achieve positive ESR. Simulation results numerically validated the accuracy of the derived closed-form expressions and verified the analytical results given in the theorems and lemmas.

Furthermore, in [J22], [C38], [C34], we turned our attention to NOMA uplink networks. We provided performance analyses in asymptotic regimes (low and high SNR) and also proposed a novel multiple access (MA) scheme referred to as NOMA-Relevant (NOMA-R). In NOMA-R, a flexible MA

scheme is proposed based on the requirement that any user will opt for NOMA only when there is a rate gain associated. We have shown that NOMA-R outperforms both NOMA and OMA in terms of sum rates achievable in all SNR regions. Importantly, using the theory of the effective capacity we demonstrated that the NOMA-R strategy is more favorable when the target delay-bound violation probabilities are more stringent, especially for weak NOMA users.

Resource Allocation Using Change Point Analysis

Related Contributions: [J17], [J20], [C32]

In [J17], [J20] and [C32] we developed novel algorithms for the real-time detection of changes in the mean and the variance of content popularity. Approaching the problem statistically, we efficiently combined off-line and on-line non-parametric CUSUM procedures. The use of non-parametric CUSUM allowed us to avoid making assumptions about the underlying statistics of the popularity of any particular content, with the additional benefit of reduced computational cost. For the detection of changes in the mean we divided the algorithm in two phases. The first phase was an extended retrospective (off-line) procedure with an improved binary segmentation step and was used to adjust on-line parameters, based on historical data of the particular video. The second phase integrated a modified trend indicator to the sequential (on-line) procedure, to reveal the direction of a detected change. We provided extensive simulations, using real data, that demonstrated the performance of the first phase of our algorithm. We also provided proof-of-concept results that highlighted the efficiency of the overall algorithm.

The approach of combining off-line and on-line CP algorithms was also employed in [J20] for the detection of changes in the variance. However, a major difference concerned the choice of the underlying test statistic, as unlike in the case of the mean, tracking changes in the variance is inherently a nonlinear estimation problem. To develop the test statistic we proposed three different approaches: i) a non-parametric approach, ii) a parametric approach using an autoregressive moving average (ARMA) model, and, iii) a parametric approach using a nonlinear generalised autoregressive conditional heteroskedasticity (GARCH) model. Our studies using synthetic data indicated that the ARMA parametric approach did not generalize well. Due to this fact, we only performed experiments on real data using the non-parametric and the GARCH approaches. We concluded that both can equally well identify large deviations in the variance and that in the general case the non-parametric approach can provide quicker detection of CPs in the datasets studied in this work. In the future, we will develop joint detectors for the mean and the variance of video content popularity.

1.4.3 Results in PLS

PLS for 5G

Related contributions: [J21], [BC3], [C37], [C33], [C27]

With the emergence of 5G low latency applications, such as haptics and V2X, low complexity and low latency security mechanisms are needed. Promising lightweight mechanisms include physical unclonable functions (PUF) and secret key generation (SKG) at the physical layer from wireless fading coefficients, as considered in [J21], [C37], [C33]. In this framework we proposed a zero-round-trip-time (0-RTT) authentication protocol combining PUF for fast authentication and generation of resumption keys using SKG. Furthermore, a novel authenticated encryption (AE) scheme using SKG and standard symmetric key block ciphers for encryption and message authentication – first proposed in [C27] – was enhanced in [J21]. Aiming at a fast PHY protocol we proposed the pipelining of the AE SKG process and the encrypted data transfer at PHY in order to reduce latency. Looking at various alternatives to implement the pipelining at PHY, we investigated a "parallel" SKG approach for multi-carrier systems (e.g., using orthogonal frequency division multiplexing (OFDM) as in LTE and 5G new radio). In the parallel approach a subset of the subcarriers was used for SKG and the rest for encrypted data transmission (using the keys generated on the subset of SKG subcarriers). The optimal solution

to the respective PHY resource allocation problem was identified under security, power and delay constraints, by formulating the subcarrier scheduling as a subset-sum 0-1 knapsack optimization [18] . A heuristic algorithm of linear complexity was proposed and shown to incur negligible loss with respect to the optimal dynamic programming solution [J21], [C37], [C33]. The proposed mechanisms, have the potential to pave the way for a new breed of latency aware PHY security protocols with an emphasis on URLLC and IoT emerging systems.

Finally, the main lines of application of PLS in B5G systems were reviewed in [BC3], starting with node authentication, moving to the information theoretic characterization of message integrity, and finally, discussing message confidentiality both in the SKG and from the wiretap channel point of view. The aim of this review was to provide a comprehensive roadmap on important relevant results by the authors and other contributors and discuss open issues on the applicability of PLS in 6G systems.

Anomaly Detection in Software Defined Networks

Related contributions: [C39], [J18]

Software-defined networking (SDN) is a promising technology to overcome many challenges in wireless sensor networks (WSN), particularly with respect to flexibility and reuse. Conversely, the centralization and the planes' separation turn SDNs vulnerable to new security threats in the general context of distributed denial of service (DDoS) attacks. State of-the-art approaches to identify DDoS do not always take into consideration restrictions in typical WSNs e.g., computational complexity and power constraints, while further performance improvement is always a target. The objective of the works in [J18], [C39] was to propose a lightweight but very efficient DDoS attack detection approach using CP analysis. Our approach was shown to have a high detection rate and linear complexity with respect to the observed time series length, rendering it suitable for WSNs. We demonstrated the performance of our detector in software-defined WSNs of 36 and 100 nodes with varying attack intensity (the number of attackers ranging from 5% to 20% of nodes).

We used CP detectors to monitor anomalies in two metrics: the data packets delivery rate and the control packets overhead. Our results showed that as the intensity of the attack increased, our approach could achieve a detection rate close to 100% and that, importantly, the type of the attack could also be inferred. As an extension of this work, we will look into distributed anomaly detection by allowing clusters of nodes to act on local early detection systems. A trade-off to be studied will concern the cluster size versus the speed of the detection while maintaining the ability to localize the source of the anomaly.

Shielding PLS Against Active Attacks

Related contributions: [BC2], [J16], [J15], [J12], [C36], [C31], [C30], [C29], [C28], [C24], [C22], [C21]

SKG schemes have been shown to be vulnerable to DoS attacks in the form of jamming and to man in the middle attacks implemented as injection attacks. In [BC2] and [C36], a comprehensive study on the impact of correlated and uncorrelated jamming and injection attacks in wireless SKG systems was presented. First, two optimal signaling schemes for the legitimate users were proposed and the impact of injection attacks as well as counter-measures were investigated. Finally, it was demonstrated that the jammer should inject either correlated jamming when imperfect channel state information (CSI) regarding the main channel was at their disposal, or, uncorrelated jamming when the main channel CSI was completely unknown.

As jamming attacks represent a critical vulnerability for wireless SKG systems, in [J15], [C31], [C30], [C29], [C28] two counter-jamming approaches were investigated for SKG systems: first, the employment of energy harvesting (EH) at the legitimate nodes to turn part of the jamming power into useful communication power, and, second, the use of channel hopping or power spreading in BF-AWGN channels to reduce the impact of jamming.[1] In both cases, the adversarial interaction between the pair

[1]We note in passing that spreading / hopping can be directly implemented with a standard inverse fast Fourier

of legitimate nodes and the jammer was formulated as a two-player zero-sum game and the Nash and Stackelberg equilibria (NE and SE) were characterized analytically and in closed form. In particular, in the case of EH receivers, the existence of a critical transmission power for the legitimate nodes allowed the full characterization of the game's equilibria and also enabled the complete neutralization of the jammer. In the case of channel hopping vs. power spreading techniques, it was shown that the jammer's optimal strategy was always power spreading while the legitimate nodes should only use power spreading in the high signal-to-interference ratio (SIR) regime. In the low SIR regime, when avoiding the jammer's interference becomes critical, channel hopping is optimal for the legitimate nodes. Numerical results demonstrated the efficiency of both counter-jamming measures.

Furthermore, in [J16] the novel proposal of using EH as a counter-jamming measure for point-to-point communication was investigated on the premise that part of the harmful interference could be harvested to increase the transmit power. We formulated the strategic interaction between a pair of legitimate nodes and a malicious jammer as a zero-sum game. Our analysis demonstrated that the legitimate nodes were able to neutralize the jammer. However, this policy was not necessarily a Nash equilibrium and hence was sub-optimal. Instead, harvesting the jamming interference could lead to relative gains of up to 95%, on average, in terms of Shannon capacity, when the jamming interference was high.

Finally, in our earlier works [J12], [C24], [C22], [C21] the resilience of wireless multiuser networks to passive (interception of the broadcast channel) and active (interception of the broadcast channel and false feedback) eavesdroppers was investigated. Stochastic characterizations of the secrecy capacity (SC) were obtained in scenarios involving a single transmitter (base station) and multiple destinations. The expected values and variances of the SC along with the probabilities of secrecy outage were evaluated in the following cases: (i) in the presence of passive eavesdroppers without any side information; (ii) in the presence of passive eavesdroppers with side information about the number of eavesdroppers; and (iii) in the presence of a single active eavesdropper with side information about the behavior of the eavesdropper. This investigation demonstrated that substantial secrecy rates are attainable on average in the presence of passive eavesdroppers as long as minimal side information is available. On the other hand, it was further found that active eavesdroppers could potentially compromise such networks unless statistical inference was employed to restrict their ability to attack. Interestingly, in the high SNR regime, multiuser networks were shown to become insensitive to the activeness or passiveness of the attack.

PLS Encoders and Secrecy Enhancement in Collaborative Networks

Related contributions: [J14], [J13], [C23], [C19], [C18]

Physical layer network coding (PNC) has been proposed for future generations of wireless networks. In [J14], we investigated PNC schemes with embedded perfect secrecy by exploiting structured interference in relay networks with two users and a single relay. In a practical scenario where both users employed finite and uniform signal input distributions, we established upper bounds (UB) on the achievable perfect secrecy rates and made these explicit when pulse amplitude modulation (PAM) modems were used, while our results extend straightforwardly to quadrature amplitude modulation (QAM) modems. We then described two simple, explicit encoders that could achieve perfect secrecy rates close to these UBs with respect to an untrustworthy relay in the single antenna and single relay setting. Lastly, we generalized our system to a MIMO relay channel where the relay had more antennas than the users and studied optimal precoding matrices which satisfied a required secrecy constraint. Our results established that the design of PNC transmission schemes with enhanced throughput and guaranteed data confidentiality was feasible.

Finally, in [J13] the optimal power allocation that maximizes the SC of BF-AWGN networks with causal CSI, M-block delay tolerance and a frame based power constraint was examined. In particular,

transform (IFFT) transmitter employed in OFDM systems; spreading requires no change in the canonical OFDM transmitter, while hopping requires setting all but one of the IFFT inputs to zero.

the SC maximization was formulated as a dynamic program. First, the SC maximization without any information on the CSI was studied; in this case the SC was shown to be maximized by equidistribution of the power budget, denoted as the "blind policy". Next, extending earlier results on the capacity maximization of BF-AWGN channels without secrecy constraints, transmission policies for the low SNR and the high SNR regimes were proposed. When the available power resources were very low the optimal strategy was a "threshold policy". On the other hand, when the available power budget was very large a "constant power policy" was shown to maximize the frame secrecy capacity. Subsequently, a novel universal transmission policy was introduced, denoted as the "blind horizon approximation" (BHA), by imposing a blind policy in the horizon of unknown events. Through numerical results, the novel BHA policy was shown to outperform both the threshold and constant power policies as long as the mean channel gain of the legitimate user was distinctively greater than the mean channel gain of the eavesdropper. Furthermore, the secrecy rates achieved by the BHA compared well with the secrecy rates of the secure waterfilling policy in the case of acausal CSI feedback to the transmitter.

1.5 Recent Teaching Activities

I have had the opportunity to teach for over 7 consecutive years in France and the UK, a variety of courses from cryptography and network security, to networking and wireless communications. A detailed description of my teaching record is presented in reverse chronological order in the following Sections.

1.5.1 Overview of Teaching Activities in France (ENSEA)

Since September 2017 I have been engaged with teaching at ENSEA, giving courses both in French and in English. I have taught in the second and third year of the engineering track of ENSEA, as well as in the continuing education track (cycle par alternance). Furthermore, since September 2019 I am responsible of the students' international mobility towards the UK.

Engineering Track (teaching in English)

I have been teaching in the third year specialization "Networks and Telecommunications" the modules of "Network Security" (module responsible), "Internetworking" (module responsible) and "Wireless Communications". In parallel I am teaching Cryptography in the M2 MSc of ETIS SIT (Systèmes, Information, Télécommunications), whose syllabus mirrors in great extent that of "Network Security". A brief presentation of the courses is given below:

(1) Network Security / Cryptography: 10 hours of lectures. Topics covered include:

- Data Confidentiality: perfect secrecy, semantic security, block ciphers, DES, 3DES, AES;
- Data Integrity: message authentication codes (MAC), authenticated encryption;
- Key management using a trusted third party;
- Public key encryption, Diffie Hellman, El Gamal, RSA;
- Digital signatures, digital certificates, public key infrastructure;
- SSL / TLS.

(2) Internetworking: 16 hours of lectures, 24 hours of lab work (on GNS3), 6 hours of seminars (classes)

- IP protocol, DHCP, ARP, ICMP, NAT;

- Routing protocols: RIP, OSPF, BGP, Mobile IP, Dynamic Source Routing, Reverse Path Forwarding, Multicasting;
- Quality of Service: Integrated Services, Differentiated Services, MPLS;
- Congestion control, TCP Tahoe, TCP Reno, TCP Vegas, Fast-TCP.

(3) Wireless Communications: 6 hours of lectures, 4 hours of lab work, 4 hours of seminars (classes)

- Signal space, maximum a posteriori detection, maximum likelihood detection;
- Design of communication system, power / bandwidth limited systems, digital modulations;
- Narrowband fading channel models and channel capacity;
- Waterfilling algorithm, adaptive QAM.

Furthermore, immediately after my recruitment at ENSEA I was tasked with developing the second year option on "Internet of things" (IoT Option: 36 hours of lectures, 28 hours of lab work in total). I have engaged with the FIT IoT-lab of INRIA in Saclay and secured three related lab sessions with remote access to the FIT-IoT lab. In the IoT option, typically 2 instructors from the industry (Nokia, Huawei or Orange) give a number of lectures on topics related to low power wide area networks (LPWAN), 3GPP standards (NB-IoT, MTC), vehicular IoT, wireless sensor networks. In the IoT option I give 6 hours of lectures on
(4) IoT security and 4 hours of lab work, covering the following topics:

- Background concepts, introduction to DTLS and IPSec;
- Introduction to blockchains for IoT;
- RFID authentication;
- Jamming attacks (primarily through lab work).

Student satisfaction in the IoT Option has been strong with an average 4/5 in the first year, bringing it amongst the best ranked second year options with a consistently high demand.

Continuous Education Track (teaching in French)

Additionally, I am teaching at the final year of the "cycle par alternace" of ENSEA in the specialization "Réseaux et Télécommunications" the module "Interconnexion et Administration des Réseaux", mirroring a reduced syllabus of the topics covered in the engineering track module "Internetworking". The module consists of 10 hours of lectures, 16 hours of lab work and 10 hours of seminars (classes).

1.5.2 Overview of Teaching Activities in the UK

Since July 2015 I have been a Fellow of the Higher Education Academy (FHEA) of the UK. FHEA is a professional title in higher education that is recognized (and currently required) by academic institutions in the United Kingdom. To become FHEA, I followed the courses offered at the University of Essex as part of the CADENZA program, between October 2014 and March 2015. Then, I prepared my teaching portfolio which included: (i) factual aspects concerning my teaching experience, (ii) in-depth familiarization with recognized teaching theories, showcased in a pedagogical thesis. The evaluation of my portfolio by the HEA took place in May 2015 and I obtained the title of a FHEA the following July. In addition, student satisfaction in the courses I have taught has been particularly strong. Notably, in the Student Evaluation of Teaching (SET) for the year 2014-2015 I scored a perfect 5/5 for my teaching of the course CE702 Digital Communications at the University of Essex.

University of Essex

I served as a Lecturer at the University of Essex, School of Computer Science and Electronic Engineering between 2013-2017. From 2014 to 2017 I was responsible for the module **"CE702 - Digital Communications"** of the MSc in Advanced Communication Systems. The 12-week module included weekly lectures and seminars (classes). Throughout the semester, 2 different assignments were given. Topics covered include:

- Systems and signals, channel coding, modulation, OFDM, MIMO;
- Multiple access methods: TDD, FDD, TDMA, FDMA, CDMA, OFDMA;
- Wireless multipath channels, equalization;
- Antennas, satellite communication;
- Optical networks, wavelength division multiplexing (WDM), dense WDM (DWDM).

In January 2015 I became the responsible of the module **"CE823 – Network Security and Cryptographic Principles"** of the MSc in Computer Networks and Security and of the optional third-year BSc module CE324 (with the same description). The 12 weeks long module included weekly lectures and lab sessions. Throughout the semester, 2 different assignments were given. The course syllabus is described below:

- Data Confidentiality: perfect secrecy, semantic security, stream ciphers, block ciphers, DES, 3DES, AES;
- Data Integrity: message authentication codes (MAC), authenticated encryption;
- Key management using a trusted third party, Kerberos protocol;
- Public key encryption, Diffie Hellman, El Gamal, RSA;
- Digital signatures, digital certificates, public key infrastructure;
- SSL / TLS, HTTPS, SSH, IPSec, DNSSec;
- Denial of service (DoS), intrusion detection, firewalls;
- Security of wireless networks, WEP, WPA, WPA2.

After my maternity leave (Sep. 2015-Jun. 2016) I became responsible for the reorganization of the module **"CE740 Mobile Communications"** of the MSc Computer Networks and Security and of the MSc in Electronic Engineering. The 12-weeks long module consisted of weekly lectures. Throughout the semester, 3 different assignments are given. Topics covered included:

- Routing: routing for static networks (Dijkstra algorithm), dynamic source routing for ad-hoc networks, clustering, data aggregation;
- MAC: Static access methods (TDMA, FDMA, CDMA, OFDMA), random access (Aloha, Slotted Aloha, CSMA, CSMA / CD, CSMA / CA), MACA protocol, scatternets, piconets, master-slave protocols, management power management / wake-up patterns, infrastructure networks and ad-hoc networks, 802.11;
- Physical layer: wireless channel, capacity, waterfilling, diversity, modulation, OFDM, Direct Sequence Spread Spectrum, Frequency Hopping Spread Spectrum, MIMO, 5G.

In addition, while at the University of Essex I supervised 3 MSc students in their projects, one of which obtained an MSc by research dissertation (year-long research project, related conference paper [C27]). Finally, I supervised 8 BSc projects, one of which was awarded the best departmental project award on "Securing DNS on Android".

Middlesex University

Between January 2009 and April 2011, I held the position of Senior Lecturer at Middlesex University, Department of Computer Communications. I was in charge of the course **"CCM4820 - Digital Transmission Techniques"** of the MSc in Telecommunications Engineering. In this context, I designed and developed my own course on digital communications. Teaching at the master's level for the first time was a great experience for me. I was able to explore all aspects of the management of teaching a course: the creation of the syllabus, the choice of the textbook, the employment and supervision of teaching assistants, the development of tutorials, additional exams and materials, preparation and presentation of courses. The course lasted 12 weeks during one semester, and included 2 hour lectures per week, weakly seminars and weakly lab sessions. Throughout the semester, 3 different assignments were given. The typical class size ranged from 25 to 80 students, and gradually increased over the years. Topics covered included:

- Stochastic signals, systems and processes, spectrum;
- Source coding: entropy, Huffman encoders, Lempel-Ziv encoders;
- Channel coding: block and convolution encoders, introduction to Turbo encoders;
- Digital modulation, OFDM systems;
- Multiple access techniques, TDMA, FDMA, CDMA, OFDMA;
- Introduction to MIMO systems, multi-path wireless channels, equalization;
- Introduction to optical systems.

In addition, I have supervised more than 10 master students in their projects, in a variety of subjects and areas of research, including physical layer security, network security, detection of anomalies in networks.

1.6 Research Supervision

My supervision activities at the PhD level include 2 students that have scheduled theses' defences for September 2020 and two further that are ongoing. In detail:

1.6.1 PhD Theses to be Defended in September 2020

PhD Student Mr. Miroslav Mitev

Supervision @60% for the period 25/4/2017-9/2020

Thesis title: "Physical layer security for the Internet of things".

Student co-supervised with Dr. M. Reed, Senior Lecturer at University of Essex, UK.

Thesis VIVA (defence) scheduled for September 2020.

Publications from thesis: [J21], [C37], [C36], [C33], [P1], [U1].

M. Mitev is registered at the Ecole Doctorale of CY University. I acted as his sole thesis director at the Doctoral School of CSEE between April-August 2017, before joining ENSEA.

PhD student Mr. Sotiris Skaperas

Supervision @40% for the period 1/9/2017-9/2020

Thesis title: "Data analysis and forecasting models for flexible resource management in 5th generation networks".

Co-supervised with Dr. L. Mamatas, Assistant professor at the University of Macedonia, GR.

Thesis defence scheduled for September 2020.

Publications from thesis: [J20], [J17], [C32], [U5].

1.6.2 Ongoing Theses

PhD student Mr. Gustavo Alonso Nunez Segura
Supervision @35% started on 1/2/2019.
Thesis title: "Cooperative Intrusion Detection System for Software Defined Wireless Sensor Networks".
Co-supervised with Dr. Cintia Borges Margi, Associate Professor at the University of Sao Paolo, BR.
Publications from thesis: [J18], [C39], [C35], [S1], [U2].

PhD student Mr. Mouktar Bello
Supervision @70% started on 1/11/2020.
Title: "Meeting delay and security constraints in 6G wireless networks".
Co-supervised with Prof. I. Fijalkow, ETIS/ENSEA, FR.
Publications from thesis: [C38], [U3, U4].

1.6.3 Current Postdoctoral Students

- Postdoc Dr. Mahdi Shakiba Herfeh: supervision @100%, 21/11/2019-20/5/2021 (fixed term 1.5 years), "Physical layer security for IoT applications", project ELIOT ANR PRCI, ETIS/ENSEA FR, publications: [BC3], [U1].
- Postdoc Dr. Nasim Ferdosian: supervision @90%, 1/1/2020-31/12/2021 (fixed term 2 years), "Non-orthogonal multiple access techniques under security and delay constraints", with Prof. I. Fijalkow, ETIS/ENSEA, FR, publications: [U5].

1.7 Structure of the Rest of the Thesis

This thesis is structured around my most recent publications (dating within the last two years) with the PhD students I supervise.

In Chapter 2, novel authentication protocols using PUFs and SKG proposed by Miroslav Mitev, myself and Martin Reed are presented. This Chapter focuses on works presented in [J21], [C33] and [C37] and includes contributions by Leila Musavian.

Next, in Chapter 3 a novel, real-time and non-parametric detector for changes in the mean value of content popularity is discussed, reflecting [J17] and [C32] with Sotiris Skaperas and Lefteris Mamatas. Additionally, the application of the same detector for intrusion detection in a software defined network is demonstrated, showcasing part of our results with Gustavo Nunez and Cintia Borges Magri in [J18] and [C35].

Chapter 4 includes some of our early results with Mouktar Bello, Wenjuan Wu and Leila Musavian on the performance analysis of NOMA uplink networks under statistical delay constraints, published in [C38].

Finally, my perspectives for future research in 6G technologies are presented in Chapter 5 and selected publications are included in the Appendix.

References

[1] A. Chorti, K. Papadaki, and H. V. Poor. Optimal power allocation in block fading channels with confidential messages. *IEEE Trans. Wireless Commun.*, 14(9):4708–4719, Sep. 2015.

[2] A. Chorti, S. M. Perlaza, Z. Han, and H. V. Poor. On the resilience of wireless multiuser networks to passive and active eavesdroppers. *IEEE J. Sel. Areas Commun.*, 31(9):1850–1863, Sep. 2013.

[3] Arsenia Chorti, Camilla Hollanti, Jean-Claude Belfiore, and H. Vincent Poor. Physical layer security: A paradigm shift in data confidentiality. *Lecture Notes in Electrical Engineering*, 358, 01 2016.

[4] G. Rezgui, E.V. Belmega, and A. Chorti. Mitigating jamming attacks using energy harvesting. *IEEE Wireless Commun. Lett.*, 8:297–300, 2019.

[5] Third Generation Partnership Project (3GPP). TR 33.825 Study on the Security of URLLC, 2019.

[6] F. Arute, K. Babbush, and *et al.* Quantum supremacy using a programmable superconducting processor. *Nature*, 574:505–510, 2019.

[7] G. E. Suh and S. Devadas. Physical unclonable functions for device authentication and secret key generation. In *2007 44th ACM/IEEE Design Automation Conference*, pages 9–14, June 2007.

[8] E. V. Belmega and A. Chorti. Protecting secret key generation systems against jamming: Energy harvesting and channel hopping approaches. *IEEE Trans. Inf. Forensics Security*, 12(11):2611–2626, Nov 2017.

[9] Michèle Basseville, Igor V Nikiforov, et al. *Detection of abrupt changes: theory and application*, volume 104. Prentice Hall Englewood Cliffs, 1993.

[10] Alexandru Tatar, Marcelo Dias De Amorim, Serge Fdida, and Panayotis Antoniadis. A survey on predicting the popularity of web content. *J. Internet Services Appl.*, 5(1):8, Dec. 2014.

[11] Gabor Szabo and Bernardo A Huberman. Predicting the popularity of online content. *Commun. ACM*, 53(8):80–88, Aug. 2010.

[12] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, C. I, and H. V. Poor. Application of non-orthogonal multiple access in LTE and 5G networks. *IEEE Commun. Mag.*, 55(2):185–191, Feb. 2017.

[13] Z. Ding, M. Peng, and H. V. Poor. Cooperative non-orthogonal multiple access in 5G systems. *IEEE Commun. Lett.*, 19(8):1462–1465, Aug. 2015.

[14] F. Fang, H. Zhang, J. Cheng, and V. C. M. Leung. Energy efficiency of resource scheduling for non-orthogonal multiple access wireless network. In *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016.

[15] D. Wu and R. Negi. Effective capacity: a wireless link model for support of quality of service. *IEEE Trans. Wireless Commun.*, 2(4):630–643, 2003.

[16] Po-Ning Chen. Generalization of Gärtner-Ellis theorem. *IEEE Trans. Inf. Theory*, 46:2752–2760.

[17] D. N. C. Tse and S. V. Hanly. Linear multiuser receivers: effective interference, effective bandwidth and user capacity. *IEEE Trans. Inf. Theory*, 45:641–657.

[18] D.P. Williamson and D.B. Shmoys. *Approximation Algorithms.* Cambridge University Press, 2011.

Chapter 2

Security Protocols for Internet of Things Applications

2.1 Introduction

With the emergence of 5G low latency applications, such as haptics and V2X, low complexity and low latency security mechanisms are needed. Promising lightweight mechanisms include physical unclonable functions (PUF) and secret key generation (SKG) at the physical layer, as considered in this Chapter. In this framework we propose i) a zero-round-trip-time (0-RTT) resumption authentication protocol combining PUF and SKG processes; ii) a novel authenticated encryption (AE) using SKG; iii) pipelining of the AE SKG and the encrypted data transfer in order to reduce latency. Implementing the pipelining at PHY, we investigate a *parallel* SKG approach for multi-carrier systems, where a subset of the subcarriers are used for SKG and the rest for data transmission. The optimal solution to this PHY resource allocation problem is identified under security, power and delay constraints, by formulating the subcarrier scheduling as a subset-sum $0-1$ knapsack optimization. A heuristic algorithm of linear complexity is proposed and shown to incur negligible loss with respect to the optimal dynamic programming solution. All of the proposed mechanisms, have the potential to pave the way for a new breed of latency aware security protocols.

2.2 Contributions and Chapter Organization

Many standard cryptographic schemes, particularly those in the realm of public key encryption (PKE), are computationally intensive, incurring considerable overheads and can rapidly drain the battery of power constrained devices [1], [2], notably in Internet of things (IoT) applications [3]. For example, a 3GPP report on the security of ultra reliable low latency communication (URLLC) systems notes that authentication for URLLC is still an open problem [4]. Additionally, traditional public key generation schemes are not *quantum secure* – in that when sufficiently capable quantum computers will be available they will be able to break current known PKE schemes – unless the key sizes increase to impractical lengths.

In the past years, physical layer security (PLS) [5–9] has been studied as a possible alternative to classic, complexity based, cryptography. As an example, signal properties as in [10], can be exploited to generate opportunities for confidential data transmission [11, 12]. Notably, PLS is explicitly mentioned as a 6G enabling technology in the first white paper on 6G [13]: "The strongest security protection may be achieved at the physical layer." In this work, we propose to move some of the security core functions down to the physical layer, exploiting both the communication radio channel and the hardware, as unique entropy sources.

Since the wireless channel is reciprocal, time-variant and random in nature, it offers a valid, inherently secure source that may be used in a key agreement protocol between two communicating

parties. The principle of secret key generation (SKG) from correlated observations was first studied in [14] and [15]. A straightforward SKG approach can be built by exploiting the reciprocity of the wireless fading coefficients between two terminals within the channel coherence time [16] and the contributions in this Chapter build upon this mechanism. This is pertinent to many forthcoming B5G applications that will require a strong, but nevertheless, lightweight security key agreement; in this direction, PLS may offer such a solution, or, complement existing algorithms. With respect to authentication, physical unclonable functions (PUFs), firstly introduced in [17] (based on the idea of physical one-way functions [18], [19]), could also enhance authentication and key agreement in demanding scenarios, including (but not limited to) device to device (D2D) and tactile Internet. We note that others also point to using physical layer security to reduce the resource overhead in URLLC [20].

A further advantage of PLS is that it is information-theoretic secure [21], *i.e.*, it is not open to attack by future quantum computers, and, it requires lower computation costs. In this work, we will discuss how SKG from shared randomness [22] is a promising alternative to PKE for key agreement. However, unauthenticated key generation is vulnerable to man in the middle (MiM) attacks. In this sense, PUFs, can be used in *conjunction* with SKG to provide authenticated secret key agreement. As summarised in [19], the employment of PUFs can decrease the computational cost and play a pivotal role in reducing the authentication latency in constrained devices.

In this study we introduce the joint use of PUF authentication and SKG in a zero-round-trip-time (0-RTT) [23, 24] approach, allowing to build quick authentication mechanisms with forward security. Further, we develop an authenticated encryption (AE) primitive [25–27] based on standard SKG schemes. To investigate a fast implementation of the AE SKG we propose a pipelined (*parallel*) scheduling method for optimal resource allocation at the physical layer (PHY) (*i.e.*, by optimal allocation of the subcarriers in 5G resource blocks).

Next, we extend the analysis to account for statistical delay quality of service (QoS) guarantees, a pertinent scenario in B5G. The support of different QoS guarantee levels is a challenging task. In fact, in time-varying channels, such as in wireless networks, determining the exact delay-bound depending on the users' requirements, is impossible. However, a practical approach, namely the effective capacity [28], can provide statistical QoS guarantees, and, can give delay-bounds with a small violation probability. In our work, we employ the effective capacity as the metric of interest and investigate how the proposed pipelined AE SKG scheme performs in a delay-constrained scenario.

The system model introduced in this work assumes that a block fading additive white Gaussian noise (BF-AWGN) channel is used with multiple orthogonal subcarriers. In our *parallel* scheme a subset of the subcarriers is used for SKG and the rest for encrypted data transfer. The findings of this study are supported by numerical results, and the efficiency of the proposed *parallel* scheme is shown to be greater or similar to the efficiency of an alternative approach in which SKG and encrypted data transfer are sequentially performed.

To summarize, the contributions of this Chapter are as follows:

1. We combine PUF authentication and SKG for resumption key agreement in a single 0-RTT protocol.

2. We develop an AE SKG scheme.

3. We propose a fast implementation of the AE SKG based on pipelining of key generation and encrypted data transfer. This *parallel* approach is achieved by allocation of the PHY resources, *i.e.*, by optimal scheduling of the subcarriers in BF-AWGN channels.

4. We propose a heuristic algorithm of linear complexity that finds the optimal subcarrier allocation with negligible loss in terms of efficiency.

5. We numerically compare the efficiency of our *parallel* approach with a *sequential* approach where SKG and data transfer are performed sequentially. This comparison is performed in two delay

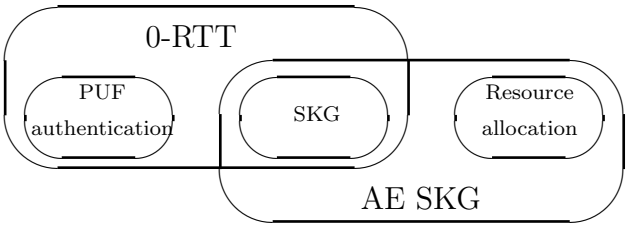


Figure 2.1: Roadmap of contributions.

scenarios:

- When a relaxed QoS delay constraint is in place;
- When a stringent QoS delay constraint is in place.

A roadmap of the Chapter's contributions is shown in Fig. 2.1.

2.2.1 Threat Model

In this work we assume a commonly used adversarial model with an active man-in-the-middle attacker (Eve) and a pair of legitimate users (Alice and Bob). For simplicity, we assume a rich Rayleigh multipath environment where the adversary is more than a few wavelengths away from each of the legitimate parties. This forms the basis of our hypothesis that the measurements of Alice and Bob are uncorrelated to the Eve's measurements.

2.2.2 Notation

Random variables are denoted in italic font, e.g., x, vectors and matrices are denoted with lower and upper case bold characters, e.g., $\mathbf{x}$ and $\mathbf{X}$, respectively. Functions are printed in a fixed-width teletype font, e.g., `F`. All sets of vectors are given with calligraphic font $\mathcal{X}$ and the elements within a set are given in curly brackets e.g. $\{\mathbf{x}, \mathbf{y}\}$, the cardinality of a vector or set is defined by vertical lines e.g., $|\mathbf{x}|$ or $|\mathcal{X}|$. Concatenation and bit-wise XORing are represented as $[\mathbf{x}||\mathbf{y}]$ and $\mathbf{x} \oplus \mathbf{y}$, respectively. We use H to denote entropy, I mutual information, $\mathbb{E}$ expectation and $\mathbb{C}$ the set of complex numbers.

2.2.3 Chapter Organization

The rest of the Chapter is organized as follows: related work is discussed in Section 2.3 followed by the general system model introduced in Section 2.4. The use of PUF authentication is illustrated in Section 2.4.1, the baseline SKG in Section 2.4.2; next, in Sections 2.4.3 and 2.4.4 we present an AE scheme using SKG and a resumption scheme to build a 0-RTT protocol. Subsequently, we evaluate the optimal power and subcarrier allocation at PHY considering both the long term average rate in Section 2.5 and the effective rate in Section 2.6. In Section 2.7, the efficiency of the proposed approach is evaluated against that of a sequential approach, while conclusions are presented in Section 2.8.

2.3 Related Work

This work assumes the use of PUF-based authentication with SKG. PUFs are hardware entities based on the physically unclonable variations that occur during the production process of silicon. These unique and unpredictable variations allow the extraction of uniformly distributed binary sequences. Due to their unclonability and simplicity, PUFs are seen as lightweight security primitives that can offer alternatives to today's authentication mechanisms. Furthermore, employing PUFs can eliminate

the need of non-volatile memory, which reduces cost and complexity [29]. Common ways of extracting secret bit sequences are through measuring jitter on oscillators, delays on gates, or, observing the power up behavior of a silicon.

Numerous PUF architectures have been proposed for IoT applications in the literature. A few of these architectures are: arbiter PUF [30], ring oscillator PUF [17], transient effect ring oscillator PUF [31], static random-access memory PUF [32], hardware embedded delay PUF [33] and more [34]. Utilising these basic properties, many PUF-based authentication protocols have been proposed, both for unilateral authentication [35, 36] and mutual authentication [29, 36–38]. A comprehensive survey on lightweight PUF authentication schemes is presented by Delvaux *et al.* [39].

On the other hand, due to the nature of propagation in the shared, free-space medium, wireless communications remain vulnerable to different types of attacks. Passive attacks such as eavesdropping or traffic analysis can be performed by anyone in the vicinity of the communicating parties; to ensure confidentiality, data encryption is vital for communication security. The required keys can be agreed at PHY using SKG. In this case, all pilot exchanges need to take place over the coherence time of the channel[1], during which Alice and Bob can observe highly correlated channel states that can be used to generate a shared secret key between them. SKG has been implemented and studied for different applications such as vehicular communications [42, 43], underwater communications [44], optical fiber [45], visible light communication [46] and more as summarized in [47]. The key conclusion from these studies is that SKG shows promise as an important alternative to current key agreement schemes.

Widely used sources of shared randomness used for SKG are the received signal strength (RSS) and the full channel state information (CSI) [48]. In either case, it is important to build a suitable pre-processing unit to decorrelate the signals in the time / frequency and space domains. As an example, some recent works have shown that the widely adopted assumption [49] that a distance equal to *half* of the wavelength (which at 2.4 GHz is approximately 6 cm [50]) is enough for two channels to decorrelate, may not hold in reality [40]. Other works show that the mobility can highly increase the entropy of the generated key [51, 52] while an important issue with the RSS-based schemes is that they are open to predictable channel attacks [40, 53]. These important issues need to be explicitly accounted for in actual implementations, but fall outside the scope of the present Chapter. We note in passing that pilot randomization can be employed to overcome limitations related to channel predictability [54].

2.4 Node Authentication Using PUFs and SKG

In this Section we present a joint physical layer SKG and PUF authentication scheme. To the best of our' knowledge this is the first work that proposes the utilization of the two schemes in conjunction. As discussed in Section 2.3, many PUF authentication protocols have been proposed in the literature, with even a few commercially available [55, 56]. We do not look into developing a new PUF architecture or a new PUF authentication protocol, instead, we look at combining existing PUF mechanisms with SKG. In addition, we develop an AE scheme that can prevent tampering attacks. To further develop our hybrid crypto-system we propose a resumption type of authentication protocol, inspired by the 0-RTT authentication mode in the transport layer security (TLS) 1.3 protocol. The resumption protocol is important as it significantly reduces the use of the PUF to the initial authentication, thus, overcoming the limitation of a PUFs' challenge response space [34, 57].

[1]The coherence time corresponds to the interval during which the multipath properties of wireless channels (channel gains, signal phase, delay) remain stable [40–42]. It is inversely proportional to the Doppler spread, which on the other hand, is a dispersion metric that accounts for the spectral broadening caused by the user's mobility (for more details and derivation please see [41]).

2.4.1 Node Authentication Using PUFs

As discussed in Section 3.9.3, for security against MiM attacks, the SKG needs to be protected through authentication. While existing techniques, such as the extensible authentication protocol-transport layer security (EAP-TLS), could be used as the authentication mechanism, these are computationally intensive and can lead to significant latency [58, 59].

This leads to the motivation to seek lightweight authentication mechanisms that can be used in conjunction with SKG. Such a mechanism that is achieving note within the research community uses a PUF. A typical PUF-based authentication protocol consists of two main phases, namely *enrolment phase* and *authentication phase* [60–64]. During the *enrolment phase* each node runs a set of challenges on its PUF and characterizes the variance of the measurement noise in order to generate side information. Next, a verifier creates and stores a database of all challenge-response pairs (CRPs) for each node's PUF within its network. A CRP pair in essence consists of an authentication key and related side information. Within the database, each CRP is associated with the ID of the corresponding node.

Later, during the *authentication phase* a node sends its ID to the verifier requesting to start a communication. Receiving the request, the verifier checks if the received ID exists in its database. If it does, the verifier chooses a random challenge that corresponds to this ID and sends it to the node. The node computes the response by running the challenge on its PUF and sends it to the verifier. However, the PUF measurements at the node are never exactly the same due to measurement noise, therefore, the verifier uses the new PUF measurement and the side information stored during the enrollment to re-generate the authentication key. Finally, the verifier compares the re-generated key to the one in the CRP and if they are identical the authentication of the node is successful. A simple approach to prevent replay attacks consists in deleting a CRP from the verifier database once it is used, but more elaborate schemes can also be built.

In summary, the motivation for using a PUF authentication scheme in conjunction with SKG is to exclude all of the computationally intensive operations required by EAP-TLS, which use modulo arithmetic in large fields. Measurements performed on current public key operations within EAP-TLS on common devices (such as IoT) give average authentication and key generation times of approximately 160 ms in static environments and this can reach up to 336 ms in high mobility conditions [65].

On the other hand, PUF authentication protocols have very low computational overhead and require overall authentication times that can be less than 10 ms [61, 66]. Furthermore, our key generation scheme, proposed in Section 2.4.2, requires just a hashing operation and (syndrome) decoding. Hashing mechanisms such as SHA256 performed on an IoT device require less than 0.3ms [66, 67]. Regarding the decoding, if we assume the usage of standard LDPC or BCH error correcting mechanisms, even in the worst-case scenario with calculations carried out as software operations, the computation is trivial compared to the hashing and requires less computational overhead [68].

2.4.2 SKG Procedure

The SKG system model is shown in Fig. 2.2. This assumes that two legitimate parties, Alice and Bob, wish to establish a symmetric secret key using the wireless fading coefficients as a source of shared randomness. Throughout our work a rich Rayleigh multipath environment is assumed, such that the fading coefficients rapidly decorrelate over short distances [16]. Furthermore, Alice and Bob communicate over a BF-AWGN channel that comprises N orthogonal subcarriers. The fading coefficients $\mathbf{h} = [h_1, \ldots, h_N]$, are assumed to be independent and identically distributed (i.i.d), complex circularly symmetric zero-mean Gaussian random variables $h_j \sim \mathcal{CN}(0, \sigma^2), j = 1, \ldots, N$. Although in actual multicarrier systems neighbouring subcarriers will typically experience correlated fading, in the present work this effect is neglected as its impact on SKG has been treated in numerous contributions in the past [69–71] and will not enhance the problem formulation in the following Sections.

The SKG procedure encompasses three phases: *advantage distillation*, *information reconciliation*, and *privacy amplification* [14], [15] as described below:

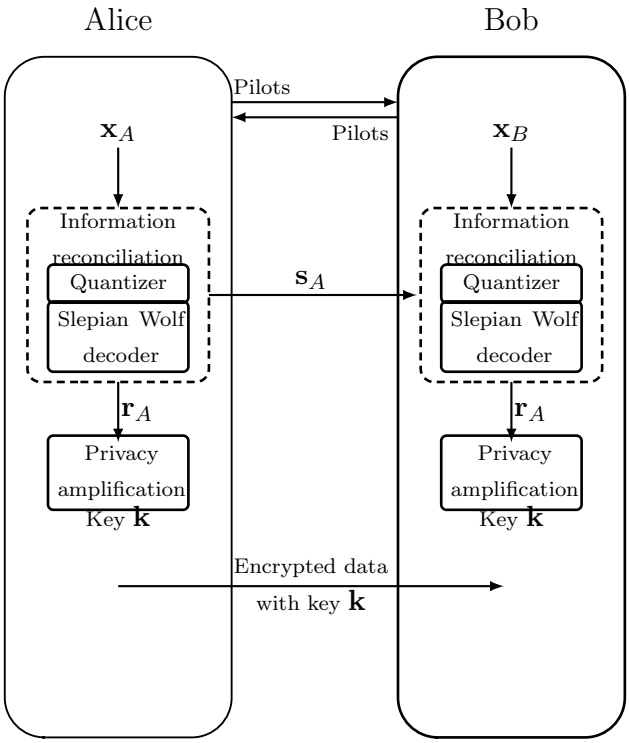


Figure 2.2: Secret key generation between Alice and Bob.

1) Advantage distillation: This phase takes place during the coherence time of the channel. The legitimate nodes sequentially exchange constant probe signals with power P on all subcarriers[2], to obtain estimates of their reciprocal CSI. We note in passing that the pilot exchange phase can be made robust with respect to injection type of attacks (that fall in the general category of MiM) as analyzed in [54]. Commonly, the received signal strength (RSS) has been used as the source of shared randomness for generating the shared key, but it is possible to use the full CSI [72]. At the end of this phase, Alice and Bob obtain observation vectors $\mathbf{x}_A = [x_{A,1}, \ldots, x_{A,N}], \mathbf{x}_B = [x_{B,1}, \ldots, x_{B,N}]$, respectively, so that:

$$\mathbf{x}_A = \sqrt{P}\mathbf{h} + \mathbf{z}_A, \tag{2.1}$$

$$\mathbf{x}_B = \sqrt{P}\mathbf{h} + \mathbf{z}_B, \tag{2.2}$$

where $\mathbf{z}_A$ and $\mathbf{z}_B$ denote zero-mean, unit variance circularly symmetric complex AWGN random vectors, such that $(\mathbf{z}_A, \mathbf{z}_B) \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{2N})$. On the other hand, Eve observes $\mathbf{x}_E = [x_{E,1}, \ldots, x_{E,N}]$ with:

$$\mathbf{x}_E = \sqrt{P}\mathbf{h}_E + \mathbf{z}_E. \tag{2.3}$$

Due to the rich Rayleigh multipath environment, Eve's channel measurement $\mathbf{h}_E$ is assumed uncorrelated to $\mathbf{h}$ and $\mathbf{z}_E$ denotes a zero-mean, unit variance circularly symmetric complex AWGN random vector $\mathbf{z}_E \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_N)$.

2) Information reconciliation: At the beginning of this phase the observations $x_{A,j}, x_{B,j}$ are quantized to binary vectors[3] $\mathbf{r}_{A,j}$, $\mathbf{r}_{B,j}$ $j = 1, \ldots, N$ [73–75], so that Alice and Bob distill $\mathbf{r}_A = [\mathbf{r}_{A,1}|| \ldots ||\mathbf{r}_{A,N}]$ and $\mathbf{r}_B = [\mathbf{r}_{B,1}|| \ldots ||\mathbf{r}_{B,N}]$, respectively. Due to the presence of noise, $\mathbf{r}_A$ and $\mathbf{r}_B$ will differ. To reconcile discrepancies in the quantizer local outputs, side information needs to be exchanged via a public channel. Using the principles of Slepian Wolf decoding, the distilled binary vectors can be

[2] An explanation of the optimality of this choice under different attack scenarios is discussed in [22].

[3] Note that each observation can generate a multi-bit vector at the output of the quantizer.

expressed as

$$\mathbf{r}_A = \mathbf{d} + \mathbf{e}_A, \tag{2.4}$$
$$\mathbf{r}_B = \mathbf{d} + \mathbf{e}_B, \tag{2.5}$$

where $\mathbf{e}_A, \mathbf{e}_B$ are error vectors that represent the distance from the common observed (codeword) vector $\mathbf{d}$ at Alice and Bob, respectively.

Numerous practical information reconciliation approaches using standard forward error correction codes (e.g., LDPC, BCH, etc.,) have been proposed [16], [72]. As an example, if a block encoder is used, then the error vectors can be recovered from the syndromes $\mathbf{s}_A$ and $\mathbf{s}_B$ of $\mathbf{r}_A$ and $\mathbf{r}_B$, respectively. Alice transmits her corresponding syndrome to Bob so that he can reconcile $\mathbf{r}_B$ to $\mathbf{r}_A$. It has been shown that the length of the syndrome $|\mathbf{s}_A|$ is lower bounded by $|\mathbf{s}_A| \geq H(\mathbf{x}_A|\mathbf{x}_B) = H(\mathbf{x}_A, \mathbf{x}_B) - H(\mathbf{x}_B)$ [15]. This has been numerically evaluated for different scenarios and coding techniques [74,76–78]. Following that, the achievable SKG rate is upper bounded by $I(\mathbf{x}_A; \mathbf{x}_B|\mathbf{x}_E)$.

3) Privacy amplification: The secret key is generated by passing $\mathbf{r}_A$ through a one-way collision resistant *compression* function i.e., by hashing. Note that this final step of privacy amplification, is executed locally without any further information exchange. The need for privacy amplification arises in order to suppress the entropy revealed due to the public transmission of the syndrome $\mathbf{s}_A$. Privacy amplification produces a key of length strictly shorter than $|\mathbf{r}_A|$, at least by $|\mathbf{s}_A|$. At the same time, the goal is for the key to be uniform, i.e., to have maximum entropy. In brief, privacy amplification *reduces the overall output entropy* while at the same time *increases the entropy per bit* – compared to the input.

The privacy amplification is typically performed by applying either cryptographic hash functions such as those built using the Merkle-Damgard construction, or universal hash functions and has been proven to be secure, in an information theoretic sense, through the leftover hash lemma [79]. As an example, [40, 80] use a 2-universal hash family to achieve privacy amplification. Summarizing, the maximum key size after privacy amplification is:

$$|\mathbf{k}| \leq H(\mathbf{x}_A) - I(\mathbf{x}_A; \mathbf{x}_\mathbf{E}) - H(\mathbf{x}_A|\mathbf{x}_B) - r_0, \tag{2.6}$$

where $H(\mathbf{x}_A)$ represents the entropy of the measurement, $I(\mathbf{x}_A; \mathbf{x}_\mathbf{E})$ represents the mutual information between Alice's and Eve's observations, $H(\mathbf{x}_A|\mathbf{x}_B)$ represents the entropy revealed during information reconciliation and $r_0 > 0$ is an extra security parameter that ensures uncertainty on the key at Eve's side. For details and estimation of these parameters in a practical scenario please see [81].

As shown in this Section the SKG procedure requires only a few simple operations such as quantization, syndrome calculation and hashing. In future work we will examine the real possibilities of implementing such a mechanism in practical systems.

2.4.3 AE Using SKG

To develop a hybrid cryptosystem that can withstand tampering attacks, SKG can be introduced in standard AE schemes in conjunction with standard block ciphers in counter mode (to reduce latency), e.g., the advanced encryption standard (AES) in Galois counter mode (GCM). As a sketch of such a primitive, let us assume a system with three parties: Alice who wishes to transmit a secret message $\mathbf{m}$ with size $|\mathbf{m}|$, to Bob with confidentiality and integrity, and Eve, that can act as a passive and active attacker. The following algorithms are employed:

- The SKG scheme denoted by $\mathtt{G} : \mathbb{C} \to \mathcal{K} \times \mathcal{S}$, accepting as input the fading coefficients (modelled as complex numbers), and generating as outputs binary vectors $\mathbf{k}$ and $\mathbf{s}_A$ in the key and syndrome spaces, of sizes $|\mathbf{k}|$ and $|\mathbf{s}_A|$, respectively,

$$\mathtt{G}(\mathbf{h}) = (\mathbf{k}, \mathbf{s}_A)\,, \tag{2.7}$$

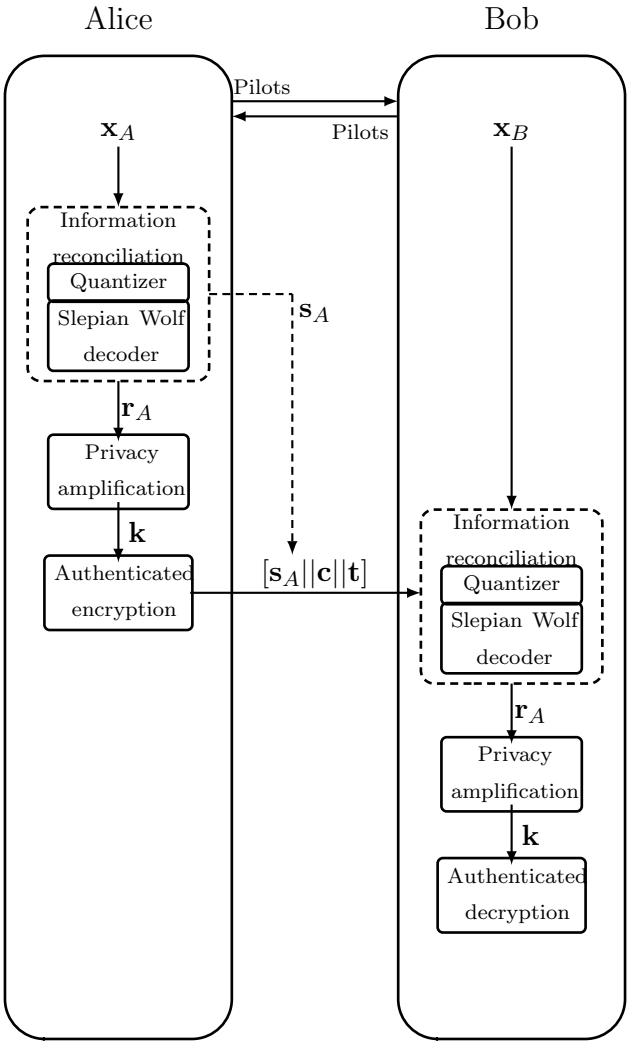


Figure 2.3: Pipelined SKG and encrypted data transfer between Alice and Bob.

where $\mathbf{k} \in \mathcal{K}$ denotes the key obtained from $\mathbf{h}$ after privacy amplification and $\mathbf{s}_A$ is Alice's syndrome.

- A symmetric encryption algorithm, e.g., AES GCM, denoted by $\texttt{Es} : \mathcal{K} \times \mathcal{M} \rightarrow \mathcal{C}_{\mathcal{T}}$ where $\mathcal{C}_{\mathcal{T}}$ denotes the ciphertext space with corresponding decryption $\texttt{Ds} : \mathcal{K} \times \mathcal{C}_{\mathcal{T}} \rightarrow \mathcal{M}$, such that

$$\texttt{Es}(\mathbf{k}, \mathbf{m}) = \mathbf{c}, \tag{2.8}$$
$$\texttt{Ds}(\mathbf{k}, \mathbf{c}) = \mathbf{m}, \tag{2.9}$$

  for $\mathbf{m} \in \mathcal{M}$, $\mathbf{c} \in \mathcal{C}_{\mathcal{T}}$.

- A pair of message authentication code (MAC) algorithms, e.g., in hashed-MAC (HMAC) mode, denoted by $\texttt{Sign} : \mathcal{K} \times \mathcal{M} \rightarrow \mathcal{T}$, with a corresponding verification algorithm $\texttt{Ver} : \mathcal{K} \times \mathcal{M} \times \mathcal{T} \rightarrow (yes, no)$, such that

$$\texttt{Sign}(\mathbf{k}, \mathbf{m}) = \mathbf{t}, \tag{2.10}$$
$$\texttt{Ver}(\mathbf{k}, \mathbf{m}, \mathbf{t}) = \begin{cases} yes, & \text{if integrity verified} \\ no, & \text{if integrity not verified} \end{cases} \tag{2.11}$$

A hybrid crypto-PLS system for AE SKG can be built as follows:

1. The SKG procedure is launched between Alice and Bob generating a key and a syndrome $\texttt{G}(\mathbf{h}) = (\mathbf{k}, \mathbf{s}_A)$.

2. Alice breaks her key into two parts $\mathbf{k} = \{\mathbf{k}_e, \mathbf{k}_i\}$ and uses the first to encrypt the message as $\mathbf{c} = \texttt{Es}(\mathbf{k}_e, \mathbf{m})$. Subsequently, using the second part of the key she signs the ciphertext using the signing algorithm $\mathbf{t} = \texttt{Sign}(\mathbf{k}_i, \mathbf{c})$ and transmits to Bob the extended ciphertext $[\mathbf{s}_A \| \mathbf{c} \| \mathbf{t}]$, as it is depicted in Fig. 2.3.

3. Bob checks first the integrity of the received ciphertext as follows: from $\mathbf{s}_A$ and his own observation he evaluates $\mathbf{k} = \{\mathbf{k}_e, \mathbf{k}_i\}$ and computes $\texttt{Ver}(\mathbf{k}_i, \mathbf{c}, \mathbf{t})$. The integrity test will fail if any part of the extended ciphertext was modified, including the syndrome (that is sent as plaintext); for example, if the syndrome was modified during the transmission, then Bob would not have evaluated the correct key and the integrity test would have failed.

4. If the integrity test is successful then Bob decrypts $\mathbf{m}=\texttt{Ds}(\mathbf{k}_e, \mathbf{c})$.

2.4.4 Resumption Protocol

In Section 2.4.1 we discussed that using PUF authentication can greatly reduce the computational overhead of a system. Authentication of new keys is required at the start of communication and at each key renegotiation. However, the number of challenges that can be applied to a single PUF is limited. Due to that we present a solution that is inspired by the 0-RTT authentication mode introduced in the 1.3 version of the TLS [23]. The use of 0-RTT obviates the need of performing a challenge for every re-authentication through the use of a resumption secret $\mathbf{r}_s$, thus reducing latency. Another strong motivation for using this mechanism is that it is forward secure in the scenario we are using here [24]. We first briefly describe the TLS 0-RTT mechanism before describing a similarly inspired 0-RTT mechanism applied to the information reconciliation phase of our SKG mechanism.

The TLS 1.3 0−RTT handshake works as follows: In the very first connection between client and server a regular TLS handshake is used. During this step the server sends to the client a look-up identifier $\mathbf{k}_l$ for a corresponding entry in session caches or it sends a session ticket. Then both parties derive a resumption secret $\mathbf{r}_s$ using their shared key and the parameters of the session. Finally, the client stores the resumption secret $\mathbf{r}_s$ and uses it when reconnecting to the same server which also retrieves it during the re-connection.

If session tickets are used the server encrypts the resumption secret using long-term symmetric encryption key, called a session ticket encryption key (STEK), resulting in a session ticket. The session ticket is then stored by the client and included in subsequent connections, allowing the server to retrieve the resumption secret. Using this approach the same STEK is used for many sessions and clients. On one hand, this property highly reduces the required storage of the server, however, on the other hand, it makes it vulnerable to replay attacks and not forward secure. Due to these vulnerabilities, in this work we focus on the session cache mechanism described next.

When using session caches the server stores all resumption secrets and issues a unique look-up identifier $\mathbf{k}_l$ for each client. When a client tries to reconnect to that server it includes its look-up identifier $\mathbf{k}_l$ in the 0-RTT message, which allows the server to retrieve the resumption secret $\mathbf{r}_s$. Storing a unique resumption secret $\mathbf{r}_s$ for each client requires server storage for each client but it provides forward security and resilience against replay attacks, when combined with a key generation mechanisms such as Diffie Hellman (or the SKG proposed in the present) which are important goals for security protocols [24]. In our physical layer 0-RTT, given that a node identifier state would be required for link-layer purposes, the session cache places little comparative load and thus is the mechanism proposed here for (re-)authentication.

The physical layer resumption protocol modifies the information reconciliation phase of Section 2.4.2 following initial authentication to provide a re-authentication mechanism between Alice and Bob. At the first establishment of communication we assume initial authentication is established, such as the mechanism shown in Section 2.4.1. During that Alice sends to Bob a look-up identifier $\mathbf{k}_l$. Then, both derive a resumption secret $\mathbf{r}_s$ that is identified by $\mathbf{k}_l$. Note, $\mathbf{r}_s$ and the session key have the same length $|\mathbf{k}|$. Then referring to the notation and steps in Section 4.1-4.3:

1. Advantage distillation phase is carried out as before (See section 2.4.2), where both parties obtain channel observations and obtain the vectors $\mathbf{r}_A$ and $\mathbf{r}_B$.

2. During the information reconciliation phase both Alice and Bob exclusive-or the resumption secret $\mathbf{r}_s$ with their observations $\mathbf{r}_A$ and $\mathbf{r}_B$, obtaining syndromes $\mathbf{s}'_A$ and $\mathbf{s}'_B$ with which both parties can carry out reconciliation to obtain the same shared value which is now $\mathbf{r}_A \oplus \mathbf{r}_s$.

3. The privacy amplification step in Section 4.2 is carried out as before, but now the hashing takes place on $\mathbf{r}_A \oplus \mathbf{r}_s$ to produce the final shared key $\mathbf{k}'$ that is a result of both the shared wireless randomness and the resumption secret.

Note that the key $\mathbf{k}'$ can only be obtained if both the physical layer generated key and the resumption key are valid and this method can be shown to be forward secure [24].

2.5 Pipelined SKG and Encrypted Data Transfer

As explained in the previous Section, if Alice and Bob follow the standard sequential SKG process they can exchange encrypted data only after both of them have distilled the key at the end of the privacy amplification step. In this Section, we propose a method to pipeline the SKG and encrypted data transfer. Alice can unilaterally extract the secret key from her observation and use it to encrypt data transmitted in the same "extended" ciphertext that contains the side information (see Fig. 2.3). Subsequently, using the side information, Bob can distill the same key $\mathbf{k}$ and decrypt the received data in one single step.

We have discussed in Section 2.4.2 how Alice and Bob can distill secret keys from estimates of the fading coefficients in their wireless link and in Section 2.4.3 how these can be used to develop an AE SKG primitive. At the same time CSI estimates are prerequisite in order to optimally allocate power across the subcarriers and achieve high data rates[4]. As a result, a question that naturally arises is

[4]As an example, despite the extra overhead, in URLLC systems advanced CSI estimation techniques are employed in order to be able to satisfy the strict reliability requirements.

whether the CSI estimates (obtained at the end of the pilot exchange phase), should be used towards the generation of secret keys or towards the reliable data transfer, and, furthermore, whether the SKG and the data transfer can be inter-woven using the AE SKG principle.

In this study, we are interested in answering this question and shed light into whether following the exchange of pilots, Alice should transmit reconciliation information on all subcarriers, so that she and Bob can generate (potentially) a long sequence of key bits, or, alternatively, perform information reconciliation only over a subset of the subcarriers and transmit encrypted data over the rest, exploiting the idea of the AE SKG primitive. Note here that the data can be already encrypted with the key generated at Alice, the sender of the side information, so that the proposed pipelining does not require storing keys for future use. We will call the former approach a *sequential* scheme, while we will refer to the latter as a *parallel* scheme. The two will be compared in terms of their efficiency with respect to the achievable data rates.

A simplified version of this problem, where the reconciliation rate is roughly approximated to the SKG rate, was investigated in [82]. In this study it was shown that in order to maximize the data rates in the *parallel* approach Alice and Bob should use the strongest subcarriers – in terms of SNR – for data transmission and the worst for SKG. Under this simplified formulation, the optimal power allocation for the data transfer has been shown to be a *modified* waterfilling solution.

Here, we explicitly account for the rate of transmitting reconciliation information and differentiate it from the SKG rate. We confirm whether the policy of using the strongest subcarriers for data transmission and not for reconciliation, is still optimal when the full optimization problem is considered, including the communication cost for reconciliation.

As discussed in Section 2.4.2, our physical layer system model assumes Alice and Bob exchange data over a Rayleigh BF-AWGN channel with N orthogonal subcarriers. Without loss of generality the variance of the AWGN in all links is assumed to be unity. During channel probing, constant pilots are sent across all subcarriers [16, 83] with power P. Using the observations (2.1), Alice estimates the channel coefficients as

$$\hat{h}_j = h_j + \tilde{h}_j, \tag{2.12}$$

for $j = 1, \ldots, N$ where $\tilde{h}_j$ denotes an estimation error that can be assumed to be Gaussian, $\tilde{h}_j \sim \mathcal{CN}(0, \sigma_e^2)$ [84]. Under this model, the following rate is achievable on the j-th subcarrier from Alice to Bob when the transmit power during data transmission is p_j [84]:

$$R_j = \log_2\left(1 + \frac{g_j p_j}{\sigma_e^2 P + 1}\right) = \log_2(1 + \hat{g}_j p_j), \tag{2.13}$$

where we use $\hat{g}_i = \frac{g_i}{\sigma_{i,e}^2 P + 1}$, to denote the estimated channel gains. As a result, the channel capacity $C = \sum_{j=1}^{N} R_j$ under the short term power constraint

$$\sum_{j=1}^{N} p_j \leq NP, \;\; p_j \geq 0, \; \forall j \in \{1, \ldots, N\}, \tag{2.14}$$

is achieved with the well known waterfilling power allocation policy $p_j = \left[\frac{1}{\lambda} - \frac{1}{\hat{g}_j}\right]^+$, where the water-level λ is estimated from the constraint (2.14). In the following, the estimated channel gains $\hat{g}_j$ are – without loss of generality – assumed ordered in descending order, so that:

$$\hat{g}_1 \geq \hat{g}_2 \geq \ldots \geq \hat{g}_N. \tag{2.15}$$

As mentioned above, the advantage distillation phase of the SKG process consists of the two-way exchange of pilot signals during the coherence time of the channel to obtain $\mathbf{r}_{A,j}, \mathbf{r}_{B,j}, j = 1, \ldots, N$.

On the other hand, the CSI estimation phase can be used to estimate the reciprocal channel gains in order to optimize data transmission using the waterfilling algorithm. In the former case, the shared parameter is used for generating symmetric keys, in the latter for deriving the optimal power allocation. In the parallel approach the idea is to inter-weave the two procedures and investigate whether a joint encrypted data transfer and key generation scheme as in the AE SKG in Section 4.3 could bear any advantages with respect to the system efficiency. While in the sequential approach the CSI across all subcarriers will be treated as a source of shared randomness between Alice and Bob, in the parallel approach it plays a dual role.

2.5.1 Parallel Approach

In the parallel approach, after the channel estimation phase, the legitimate users decide on which subcarrier to send the reconciliation information (e.g., the syndromes as discussed in Section 2.4.2) and on which data (*i.e.*, the SKG process here is not performed on all of the subcarriers). The total capacity has now to be distributed between data and reconciliation information bearing subcarriers. As a result, the overall set of orthogonal subcarriers comprises two subsets; a subset $\mathcal{D}$ that is used for encrypted data transmission with cardinality $|\mathcal{D}| = D$ and a subset $\breve{\mathcal{D}}$ with cardinality $|\breve{\mathcal{D}}| = N - D$ used for reconciliation such that, $\mathcal{D} \cup \breve{\mathcal{D}} = \{1, \ldots, N\}$. Over $\mathcal{D}$ the achievable sum data transfer rate, denoted by C_D is given by

$$C_D = \sum_{j \in \mathcal{D}} \log_2(1 + \hat{g}_j p_j), \tag{2.16}$$

while on the subset $\breve{\mathcal{D}}$, Alice and Bob exchange reconciliation information at rate

$$C_R = \sum_{j \in \breve{\mathcal{D}}} \log_2(1 + \hat{g}_j p_j). \tag{2.17}$$

As stated in Section 2.4.2 the fading coefficients are assumed to be zero-mean circularly-symmetric complex Gaussian random variables. Using the theory of order statistics, the distribution of the ordered channel gains of the SKG subcarriers, $j \in \breve{\mathcal{D}}$, can be expressed as [85]:

$$f(g_j) = \frac{N!}{\sigma^2 (N-j)!(j-1)!} \left(1 - e^{-\frac{\hat{g}_j}{\sigma^2}}\right)^{N-j} \left(e^{-\frac{\hat{g}_j}{\sigma^2}}\right)^{j}, \tag{2.18}$$

where σ^2 is the variance of the channel gains. As a result of ordering the subcarriers, the variance of each of the subcarriers, is now given by:

$$\sigma_j^2 = \sigma^2 \sum_{q=j}^{N} \frac{1}{q^2}, \quad j \in \{D+1, \ldots, N\}. \tag{2.19}$$

Thus, we can now write the SKG rate as (note that the noise variances are here normalized to unity for simplicity) [16, 83]:

$$C_{SKG} = \sum_{j \in \breve{\mathcal{D}}} \log_2 \left(1 + \frac{P\sigma_j^2}{2 + \frac{1}{P\sigma_j^2}}\right). \tag{2.20}$$

The minimum rate necessary for reconciliation was discussed in Section 4.2. Here, alternatively, we employ a practical design approach in which the rate of the encoder used is explicitly taken into account. Note that in a rate $\frac{k}{n}$ block encoder the side information is $n - k$ bits long, i.e., the rate of syndrome to output key bits after privacy amplification is $\frac{n-k}{k}$. However, in each key session a 0-RTT look-up identifier of length k is also sent. Therefore, we define the parameter $\kappa = \frac{n-k}{k} + 1 = \frac{n}{k}$, i.e., the inverse of the encoder rate, that reflects the ratio of the reconciliation and 0-RTT transmission

rate to the SKG rate. For example, for a rate $\frac{k}{n} = \frac{1}{2}$ encoder, $\kappa = 2$, etc. Based on this discussion, we capture the minimum requirement for the reconciliation rate through the following expression:

$$C_R \quad \geq \quad \kappa C_{SKG}. \tag{2.21}$$

Furthermore, to identify the necessary key rate, we note that depending on the exact choices of the cryptographic suites to be employed, it is possible to reuse the same key for the encryption of multiple blocks of data, e.g., as in the AES GCM, that is being considered for employment in the security protocols for URLLC systems [4]. In practical systems, a single key of length 128 to 256 bits can be used to encrypt up to gigabytes of data. As a result, we will assume that for a particular application it is possible to identify the ratio of key to data bits, which in the following we will denote by β. Specifically, we assume that the following security constraint should be met

$$C_{SKG} \geq \beta C_D, \;\; 0 < \beta \leq 1, \tag{2.22}$$

where, depending on the application, the necessary minimum value of β can be identified. We note in passing that the case $\beta = 1$ would correspond to a one-time-pad, *i.e.*, the generated keys could be simply x-ored with the data to achieve perfect secrecy without the need of any cryptographic suites.

Accounting for the reconciliation rate and security constraints in (2.21) and (2.22) we formulate the following maximization problem:

$$\max_{p_j, j\in\mathcal{D}} \sum_{j\in\mathcal{D}} R_j \tag{2.23}$$

$$\text{s.t. } (2.14), (2.21), (2.22),$$

$$\sum_{j\in\mathcal{D}} R_j + \sum_{j\in\check{\mathcal{D}}} R_j \leq C. \tag{2.24}$$

(2.22) can be integrated with (2.21) to the combined constraint

$$\sum_{j\in\mathcal{D}} R_j \quad \leq \quad \frac{\sum\limits_{j\in\check{\mathcal{D}}} R_j}{\kappa\beta}. \tag{2.25}$$

The optimization problem at hand is a mixed-integer convex optimization problem with unknowns both the sets $\mathcal{D}, \check{\mathcal{D}}$, as well as the power allocation policy $p_j, j \in \{1, \ldots, N\}$. These problems are typically NP hard and addressed with the use of branch and bound algorithms and heuristics.

In this work, we propose a simple heuristic to make the problem more tractable by reducing the number of free variables. In the proposed approach, we assume that the constraint (2.24) is satisfied with equality. The only power allocation that allows this is the waterfilling approach that uniquely determines the power allocation p_j and also requires that the constraint (2.14) is also satisfied with equality. Thus, if we follow that approach, we determine the power allocation vector uniquely and can combine the remanining constraints (2.24) and (2.25) into a single one as:

$$\sum_{j\in\mathcal{D}} R_j \leq \frac{C}{\kappa\beta + 1}. \tag{2.26}$$

Algorithm 1: Heuristic Greedy Algorithm for (2.27)-(2.28)

procedure HEURISTIC(start, end, R_j)
  $j \leftarrow 1, R_0 \leftarrow 0, R_{N+1} \leftarrow 0$
  while $j \leq N-1$ and $\sum_{j=1}^{N} R_j x_j \leq \frac{C}{1+\kappa\beta}$ **do**
    $\sum_{j=1}^{N} R_j x_j \leftarrow \sum_{j=1}^{N} R_{j-1} x_{j-1} + R_j x_j$
    if $\sum_{j=1}^{N} R_j x_j \leq \frac{C}{1+\kappa\beta}$ **then**
      $x_j \leftarrow 1; j \leftarrow j+1$
    else do $x_j \leftarrow 0; j \leftarrow j+1$
    end if
  end while
end procedure

The new optimization problem can be re-written as

$$\max_{x_j \in \{0,1\}} \sum_{j=1}^{N} R_j x_j \tag{2.27}$$

$$\text{s.t.} \sum_{j=1}^{N} R_j x_j \leq \frac{C}{1+\kappa\beta}. \tag{2.28}$$

The problem in (2.27)-(2.28) is a subset-sum problem from the family of $0-1$ knapsack problems, that is known to be NP hard [86]. However, these type of problems are solvable optimally using dynamic programming techniques in pseudo-polynomial time [86, 87]. Furthermore, it is known that greedy heuristic approaches are bounded away from the optimal solution by half [88].

We propose a simple greedy heuristic algorithm of *linear complexity,* as follows.[5] The data subcarriers are selected starting from the best – in terms of SNR – until (2.28) is not satisfied. Once this situation occurs the last subcarrier added to set $\mathcal{D}$ is removed and the next one is added. This continues either to the last index N or until (2.28) is satisfied with equality. The algorithm is described in *Algorithm 1.*

The efficiency of the proposed parallel method – measured as the ratio of the long-term data rate versus the average capacity – is evaluated as:

$$\eta_{\text{parallel}} = \frac{\mathbb{E}\left[\sum_{j \in \mathcal{D}} R_j\right]}{\mathbb{E}[C]}. \tag{2.29}$$

This efficiency quantifies the expected back-off in terms of data rates when part of the resources (power and frequency) are used to enable the generation of secret keys at the physical layer. In future work, we will compare the efficiency achieved to that of actual approaches currently used in 5G by accounting for the actual delays incurred due to the PKE key agreement operations [20].

2.5.2 Sequential Approach

In the sequential approach encrypted data transfer and secret key generation are two separate events; first, the secret keys are generated over the whole set of subcarriers, leading to a sum SKG rate given

[5] Without loss of generality, the algorithm assumes that the channel gains are ordered in decreasing order as in (2.15), and, consequently, the rates R_j are also ordered in descending order. The ordering is a $\mathcal{O}(N \log N)$ operation and required in common power allocation schemes such as the waterfilling, and, therefore does not come at any additional cost.

as

$$C_{SKG} = N \log_2 \left(1 + \frac{P\sigma^2}{2 + \frac{1}{P\sigma^2}} \right). \tag{2.30}$$

To estimate the efficiency of the scheme, we further need to identify the necessary resources for the exchange of the reconciliation information. We can obtain an estimate of the number of transmission frames that will be required for the transmission of the syndromes, as the expected value of the reconciliation rate (*i.e.*, it's long-term value) $\mathbb{E}[C_R]$. The average number of frames needed for reconciliation is then computed as:

$$M = \left\lceil \frac{\kappa C_{SKG}}{\mathbb{E}[C_R]} \right\rceil, \tag{2.31}$$

where $\lceil x \rceil$ denotes the smallest integer that is larger than x.

The average number of the frames that can be sent while respecting the secrecy constraint is:

$$L = \left\lfloor \frac{C_{SKG}}{\beta \mathbb{E}[C]} \right\rfloor, \tag{2.32}$$

where $\lfloor x \rfloor$ denotes the largest interger that is smaller than x. The efficiency of the sequential method is then calculated as:

$$\eta_{\text{sequential}} = \frac{L}{L + M}. \tag{2.33}$$

2.6 Effective Data Rate Taking into Account Statistical Delay QoS Requirements

In the previous section, we investigated the optimal power and subcarrier allocations strategy of Alice and Bob in order to maximize their long-term average data rate and proposed a greedy heuristic algorithm of linear complexity. Here, we extend our work from Section 2.5 by taking into account delay requirements. In detail, we investigate the optimal resource allocation for Alice and Bob, when their communication has to satisfy specific delay constraints. To this end, we use the theory of *effective capacity* [28] which gives a limit for the maximum arrival rate under delay-bounds with a specified violation probability.

We study the *effective data rate* for the proposed pipelined SKG and encrypted data transfer scheme; the effective rate is a data-link layer metric that captures the impact of statistical delay QoS constraints on the transmission rates. As background, we refer to [89] which showed that the probability of a steady-state queue length process $Q(t)$ exceeding a certain queue-overflow threshold x converges to a random variable $Q(\infty)$ as:

$$\lim_{x \to \infty} \frac{\ln(\Pr[Q(\infty) > x])}{x} = -\theta, \tag{2.34}$$

where θ indicates the asymptotic exponential decay-rate of the overflow probability. For a large threshold x, (2.34) can be represented as $Pr[Q(\infty) > x] \approx e^{-\theta x}$. Furthermore, the delay-outage probability can be approximated by [28] :

$$\Pr{}^{\text{out}}_{\text{delay}} = \Pr[\text{Delay} > D_{\max}] \approx \Pr[Q(\infty) > 0] e^{-\theta \zeta D_{\max}}, \tag{2.35}$$

where $D_{\max}$ is the maximum tolerable delay, $\Pr[Q(\infty) > 0]$ is the probability of a non-empty buffer, which can be estimated from the ratio of the constant arrival rate to the averaged service rate, ζ is the upper bound for the constant arrival rate when the statistical delay metrics are satisfied.

Using the delay exponent θ and the probability of non-empty buffer, the effective capacity, that

denotes the maximum arrival rate, can be formulated as [28]:

$$E_C(\theta) = -\lim_{t\to\infty} \frac{1}{\theta} \ln \mathbb{E}\left[e^{-\theta S[t]}\right], \tag{2.36}$$

where $S[t] = \sum_{i=1}^{t} s[i]$ denotes the time-accumulated service process, and $s[i], i = 1, 2, ...$ denotes the discrete-time stationary and ergodic stochastic service process. Therefore, the delay exponent θ indicates how strict the delay requirements are, *i.e.*, $\theta \to 0$ corresponds to looser delay requirements, while $\theta \to \infty$ implies exceptionally stringent delay constraints. Assuming a Rayleigh block fading system, with frame duration T_f and total bandwidth B, we have $s[i] = T_f B\tilde{R}_i$, with $\tilde{R}_i$ representing the instantaneous service rate achieved during the duration of the ith frame. In the context of the investigated data and reconciliation information transfer, $\tilde{R}_i$, is given by:

$$\tilde{R}_i = \frac{1}{F} \sum_{i\in\mathcal{D}} \log_2(1 + p_i\hat{g}_i), \tag{2.37}$$

where F is the equivalent frame duration, *i.e.*, the total number of subcarriers used for data transmission, so that for the parallel approach we have $F = |D|$ while for the sequential approach $F = N(L+M)L^{-1}$.

Under this formulation and assuming that Gärtner-Ellis theorem [90, 91] is satisfied, the *effective data rate*[6] $E_C(\theta)$ is given as:

$$E_{C,\mathcal{D}}(\theta) = -\frac{1}{\theta T_f B} \ln\left(\mathbb{E}\left[e^{-\theta T_f B\tilde{R}_i}\right]\right). \tag{2.38}$$

We set $\alpha = \frac{\theta T_f B}{\ln(2)}$. By inserting (2.37) into (2.38) we get:

$$\begin{aligned} E_{C,\mathcal{D}}(\theta) &= -\frac{1}{\ln(2)\alpha} \ln\left(\mathbb{E}\left[e^{-\ln(2)\alpha F^{-1}\sum_{i\in\mathcal{D}}\log_2(1+p_i\hat{g}_i)}\right]\right), \\ E_{C,\mathcal{D}}(\theta) &= -\frac{1}{\alpha}\log_2\left(\mathbb{E}\left[\prod_{i\in\mathcal{D}}(1+p_i\hat{g}_i)^{-\alpha F^{-1}}\right]\right). \end{aligned} \tag{2.39}$$

Assuming i.i.d. channel gains, by using the distributive property of the mathematical expectation, (2.39) becomes [92]:

$$E_{C,\mathcal{D}}(\theta) = -\frac{1}{\alpha}\log_2\left(\prod_{i\in\mathcal{D}}\mathbb{E}\left[(1+p_i\hat{g}_i)^{-\alpha F^{-1}}\right]\right). \tag{2.40}$$

We further manipulate by using the log-product rule to obtain:

$$E_{C,D}(\theta) = -\frac{1}{\alpha}\sum_{i\in\mathcal{D}}\log_2\left(\mathbb{E}\left[(1+p_i\hat{g}_i)^{-\alpha F^{-1}}\right]\right). \tag{2.41}$$

Similarly, the *effective syndrome rate* can be written as:

$$E_{C,\breve{D}}(\theta) = -\frac{1}{\alpha}\sum_{i\in\breve{\mathcal{D}}}\log_2\left(\mathbb{E}\left[(1+p_i\hat{g}_i)^{-\alpha\breve{F}^{-1}}\right]\right), \tag{2.42}$$

where the size of $\breve{F}$ here is $|N - D|$.

[6]Since part of the transmission rate is used for reconciliation information, and part for data transmission the terms "*effective syndrome rate*" and "*effective data rate*" are introduced instead of the term "effective capacity", for rigour. We note that we assume the information data and reconciliation information are accumulated in separate independent buffers within the transmitter.

Using that, we now reformulate the maximization problem given in (2.23) by adding a delay constraint. The reformulated problem can be expressed as follows:

$$\max_{p_j, j\in\mathcal{D}} E_{C,\mathcal{D}}(\theta), \tag{2.43}$$

$$\text{s.t. } (2.14), (2.25),$$

$$E_{C,\mathcal{D}}(\theta) + E_{C,\check{\mathcal{D}}}(\theta) \leq E_C^{\text{opt}}(\theta), \tag{2.44}$$

where $E_C^{\text{opt}}(\theta)$ represents the maximum achievable effective capacity for both key and data transmission for a given value of θ over N subcarriers:

$$E_C^{\text{opt}}(\theta) = \max_{p_i, i=1,2,\ldots N} \left\{ -\frac{1}{\alpha} \log_2 \left(\mathbb{E}\left[\prod_{i=1}^{N} (1 + p_i \hat{g}_i)^{-\alpha N^{-1}} \right] \right) \right\}. \tag{2.45}$$

In the proposed approach, we assume that the constraint (2.44) is satisfied with equality. Given that, the optimization problem in (2.43) can be evaluated as two sub-optimization problems: i) finding the optimal long term power allocation from (2.14) and (2.45); ii) finding the optimal subcarrier allocation that satisfies (2.25). We solve the first problem that gives the optimal power allocation using convex optimization tools. Next, as in Section 2.5 we use two methods to solve subcarrier allocation problem, i.e., by formulating a subset-sum 0 – 1 knapsack optimization problem or through a variation of *Algorithm 1.* The efficiency of both methods is compared numerically to the sequential method in Section 3.10.

Now, following the same steps as in (2.39)-(2.41) and using the fact that maximizing $E_C(\theta)$ is equivalent to minimizing $-E_C(\theta)$ (this is due to $\log(\cdot)$ being a monotonically increasing concave function for any $\theta > 0$) we formulate the following minimization problem:

$$\min_{p_i, i=1,2,\ldots N} \sum_{i=1}^{N} \left(\mathbb{E}\left[(1 + p_i \hat{g}_i)^{-\alpha N^{-1}} \right] \right), \tag{2.46}$$

$$\text{s.t. } (2.14).$$

where $F = N$ in this case as the full set of subcarriers is concerned. We form the Lagrangian function $\mathcal{L}$ as:

$$\mathcal{L} = \left(\mathbb{E}\left[(1 + p_i \hat{g}_i)^{-\alpha N^{-1}} \right] \right) + \lambda \left(\sum_{i=1}^{N} p_i - NP \right). \tag{2.47}$$

By differentiating (2.47) w.r.t. p_i and setting the derivative equal to zero [93] we get:

$$\frac{\partial \mathcal{L}}{\partial p_i} = \lambda - \frac{\alpha \hat{g}_i}{N} \left(\hat{g}_i p_i + 1 \right)^{-\frac{\alpha}{N} - 1} = 0. \tag{2.48}$$

Solving (2.48) gives the optimal power allocation policy:

$$p_i^* = \frac{1}{g_0^{\frac{N}{\alpha+N}} \hat{g}_i^{\frac{\alpha}{\alpha+N}}} - \frac{1}{\hat{g}_i}, \tag{2.49}$$

where $g_0 = \frac{N\lambda}{\alpha}$ is the cutoff value which can be found from the power constraint. By inserting p_i^* in

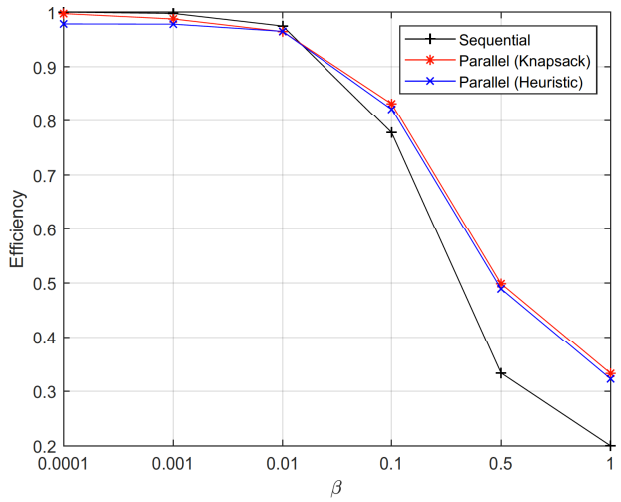


Figure 2.4: **a)** Efficiency comparison for $N = 12$, SNR=10 dB and $\kappa = 2$.

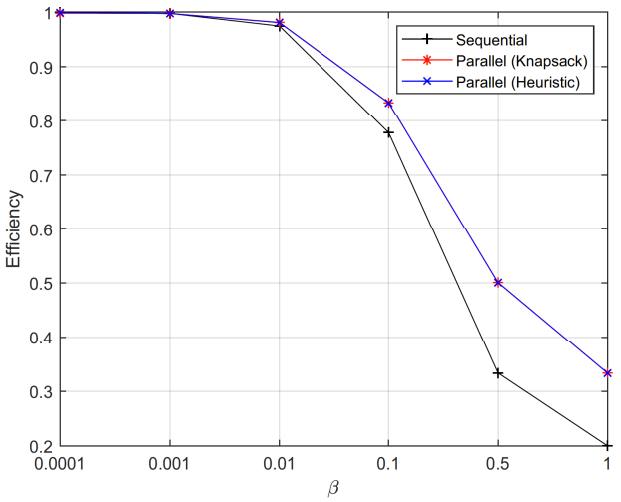


Figure 2.4: **b)** Efficiency comparison for $N = 64$, SNR=10 dB and $\kappa = 2$.

$E_C(\theta)$ we obtain the expression for $E_C^{\text{opt}}(\theta)$:

$$E_C^{\text{opt}}(\theta) = -\frac{1}{\alpha}\sum_{i=1}^{N}\log_2\left(\mathbb{E}\left[\left(\frac{\hat{g}_i}{g_0}\right)^{-\frac{\alpha}{\alpha+N}}\right]\right) \tag{2.50}$$

When $\theta \to 0$ the optimal power allocation is equivalent to water-filling and when $\theta \to \infty$ the optimal power allocation transforms to total channel inversion.

Now, fixing the power allocation as in (2.49) we can easily find the optimal subcarrier allocation that satisfies (2.25). As in Section 2.5 to do that we first formulate a subset-sum $0-1$ knapsack optimization problem that we solve using the standard dynamic programming approach. Furthermore we evaluate the performance of the heuristic algorithm presented in *Algorithm 1.*

2.7 Results and Discussion

In this Section we provide numerical evaluations of the efficiency that can be achieved with the presented methods (*i.e.*, sequential and parallel) for different values of the main parameters. With respect to the parallel approach, we provide numerical results of the optimal dynamic programming solution of the subset-sum $0-1$ knapsack problem, as well as of the greedy heuristic approach presented in *Algorithm 1.* For the case of the long term average data rate C_D (2.16), we compare the two methods through their efficiencies, *i.e.* $\eta_{\text{sequential}}$ and η_{parallel} given in (2.33) and (2.29), respectively. Next, to compare the two methods in the case of *effective data rate* we evaluate $E_{C,D}(\theta)$ given in (2.41). For better illustration of each case they are separated into different subsections.

2.7.1 Numerical results for the Long Term Average C_D

Figures 2.4a and 2.4b show the efficiency of the methods for $N = 12$, and $N = 64$, respectively, while $\kappa = 2$ and $P = 10$. We note that the proposed heuristic algorithm has a near-optimal performance (almost indistinguishable from the red curves achieved with dynamic programming). Due to this fact (which was tested across all scenarios that follow) only the heuristic approach is shown in subsequent figures for clarity in the graphs.

We see that when there are a small number of subcarriers (N=12, typical for NB-IoT) and small β the efficiency of both the parallel η_{parallel} and the sequential $\eta_{\text{sequential}}$ approaches are very close to

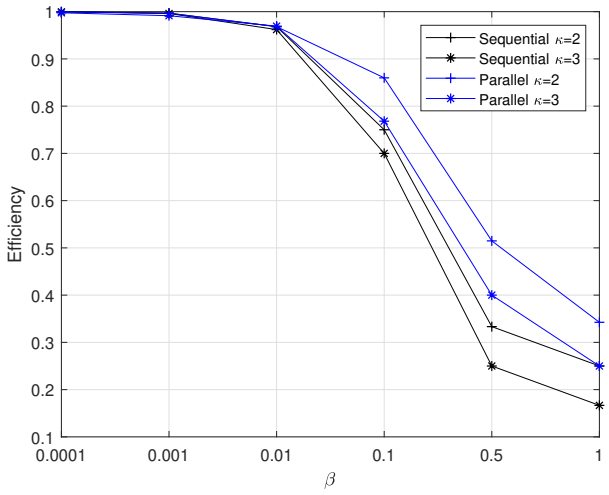


Figure 2.5: Efficiency vs κ, for $N = 24$, SNR=10 dB.

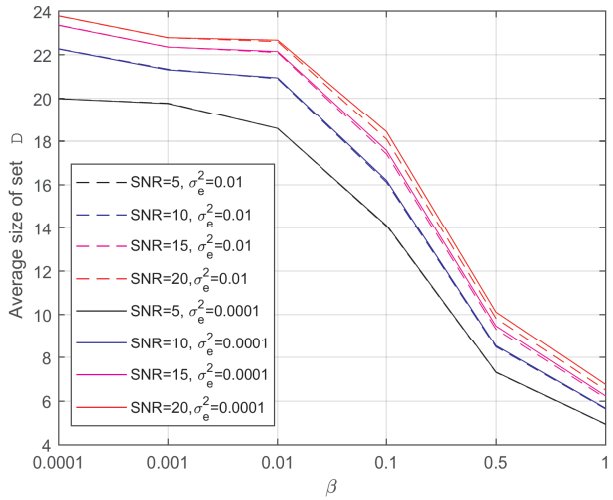


Figure 2.6: **a)** Size of set $\mathcal{D}$ for different SNR levels and σ_e^2 when $N = 24$.

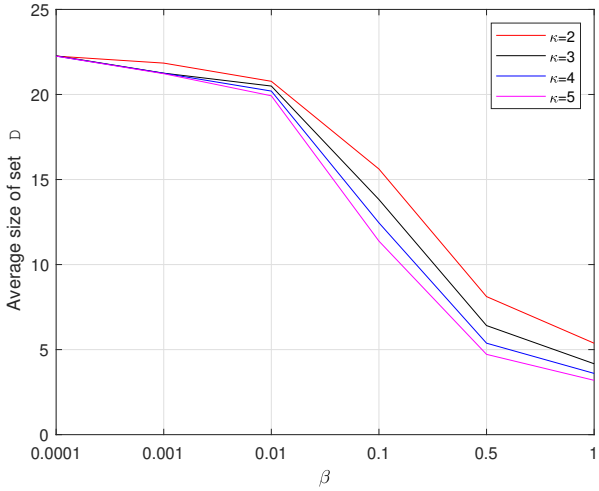


Figure 2.6: **b)** Size of set $\mathcal{D}$ for different values of κ when $N = 24$.

unity, a trend that holds for increasing N. With increasing β, due to the fact that more frames are needed for reconciliation in the sequential approach (*i.e.*, M increases), regardless of the total number of subcarriers, the parallel method proves more efficient than the sequential. While the efficiency of the sequential and parallel methods coincide almost until around $\beta = 0.01$ for $N = 12$, for $N = 64$ the crossing point of the curves moves to the left and the efficiency of the two methods coincide until around $\beta = 0.001$. This trend was found to be consistent across many values of N, only two of which are shown here for compactness of presentation.

Next, in Fig. 2.5 the efficiency of the parallel η_{parallel} and the sequential $\eta_{\text{sequential}}$ methods are shown for two different values of $\kappa \in \{2, 3\}$ for SNR = 10 dB and $N = 24$. It is straightforward to see that they both follow similar trends and when κ increases the efficiency decreases. On the other hand, regardless of the value of κ they both perform identically until around $\beta = 0.001$.

Finally, in Fig. 2.6, focusing on the parallel method, the average size of set $\mathcal{D}$ is shown for different values of σ_e^2 and SNR levels (Fig. 2.6a) and κ (Fig. 2.6b), for $N = 24$. As expected, in Fig. 2.6a we see when the SNR increases the size of the set increases, too. This is due to the fact that more power is used on any single subcarrier and consequently a higher reconcilliation rate can be sustained.

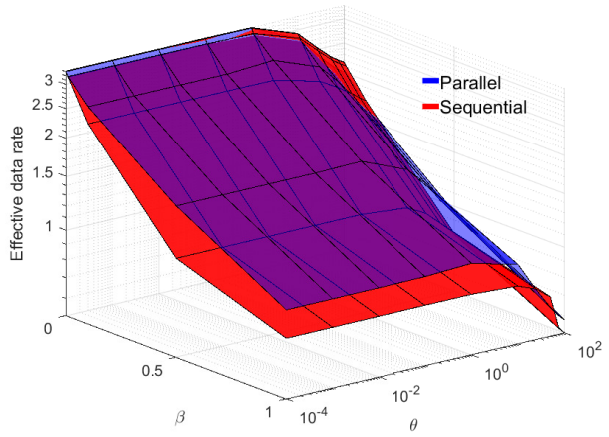


Figure 2.7: **a)** Effective data rate achieved by the parallel heuristic approach and the sequential approach when $N = 12$, SNR= 10 dB and $\kappa = 2$.

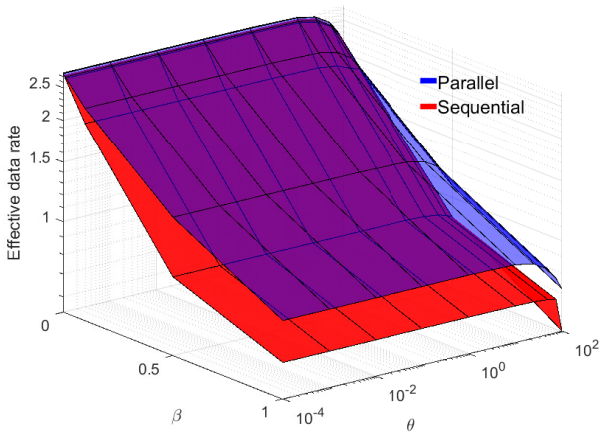


Figure 2.7: **b)** Effective data rate achieved by the parallel heuristic approach and the sequential approach when $N = 64$, SNR= 10 dB and $\kappa = 2$.

Regarding the estimation error σ_e^2 of the CSI, it only slightly affects the performance at high SNR levels. Hence more subcarriers have to be used for reconciliation, and fewer for data. The SNR level in Fig. 2.6b is set to 10 dB. The figure shows that when increasing κ the size of set $\mathcal{D}$ decreases. This result can be easily predicted from inequality (2.21), meaning, when κ increases more reconciliation data has to be sent, hence fewer subcarriers can be used for data. In both Fig. 2.6a and Fig. 2.6b when β increases the size of set $\mathcal{D}$ decreases; this effect is a consequence of constraint (2.28) as the data rate is decreasing with β.

2.7.2 Numerical Results for the Effective Data Rate

Inspired by the good performance of *Algorithm 1*, in the case where long-term average rate is the metric of interest, here, we continue our investigation with a variation of *Algorithm 1*, with the following differences: at lines 3 and 5 instead of (2.26) we use the constraint (2.25), the power allocation is fixed as in (2.49). The performance of our system is again compared with a sequential method and the metric of interest here is the *effective data rate*. The comparison is performed by taking into account the following parameters: signal to noise ration (SNR); number of subcarriers N; ratio of the reconciliation and $0-$RTT transmission rate to the SKG rate κ; delay exponent θ; and, the ratio of key bits to data bits β.

In Fig. 2.7 we give a three-dimensional plot showing the dependence of the achievable *effective data rate* $E_{C,D}(\theta)$ on β and θ. Figures 2.7a and 2.7b compare the parallel heuristic approach and the sequential approach for high SNR levels, whereas Fig. 2.8a and 2.8b compare their performance at low SNRs. In Fig. 2.7a and 2.8a we have $N = 12$ while in Fig. 2.7b and 2.8b the total number of subcarriers is $N = 64$. All graphs compare the performance of the heuristic parallel approach and the sequential approach for $\kappa = 2$.

As discussed in Section 2.6, when the delay exponent θ increases, the optimal power allocation transforms from waterfilling to total channel inversion. Consequently, the rate achieved on all subcarriers converges to the same value, hence when we a have small number of subcarriers (such as $N = 12$) and small values of β then using a single subcarrier for reconciliation data will use more capacity than needed and most of the rate on this subcarrier is wasted. Devoting a whole subcarrier for sending the reconciliation data for the case of $N = 12$ and $\beta = 0.0001$ is almost equivalent of losing $1/12$ of the achievable rate.

This can be seen in Fig. 2.7a and 2.8a where $N = 12$. When the SNR is high (See Fig. 2.7a),

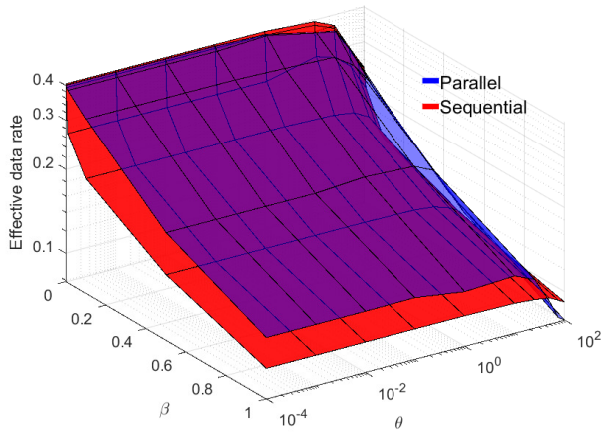


Figure 2.8: **a)** Effective data rate achieved by the parallel heuristic approach and the sequential approach when $N = 12$, SNR= 0.2 dB and $\kappa = 2$.

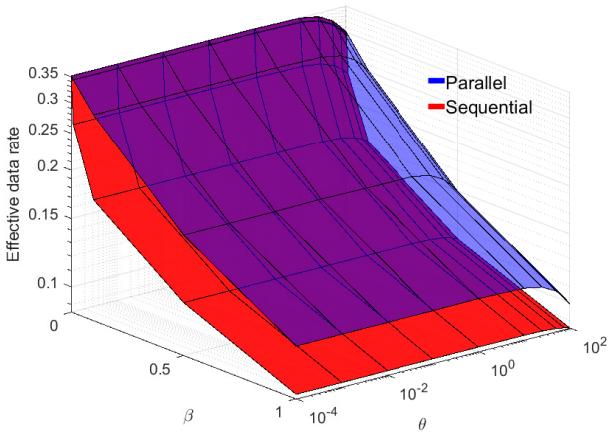


Figure 2.8: **b)** Effective data rate achieved by the parallel heuristic approach and the sequential approach when $N = 64$, SNR= 0.2 dB and $\kappa = 2$.

as discussed, this effect is mostly noticeable for large values of θ and small values of β[7], whereas for small values of β and θ both algorithms perform nearly identically. A similar trend can be seen at the low SNR regime in Fig. 2.8a. However, at a low SNR the sequential approach has a lower effective data rate. This happens because at high SNR levels each reconciliation frame will contain more information and hence more data frames will follow. Therefore, at the low SNR regime, the reconciliation information received will decrease, hence less data can be sent afterwards. This does not affect the parallel approach. However, in both scenarios high SNR Fig. 2.7a and low SNR Fig. 2.8a, when β increases regardless of the value of θ the parallel approach always achieves higher *effective data rate* $E_{C,D}(\theta)$.

In the next case, when the total number of subcarriers is $N = 64$, illustrated in Fig. 2.7b and 2.8b, we see that the penalty of devoting a high part of the achievable effective capacity $E_C^{\text{opt}}(\theta)$ to reconciliation disappears and the heuristic parallel approach always achieves higher or identical *effective data rate* $E_{C,D}(\theta)$ compared to the sequential approach. This trend repeats for high and low SNR levels as given in Fig. 2.7b and 2.8b, respectively.

Now, we take a closer look and transform some specific cases from the 3D plots to two-dimensional graphs. In Fig. 2.9 we see the achieved *effective data rate* $E_{C,D}(\theta)$ given in (2.41), for different values of N and θ while the SNR=5 dB and $\kappa = 2$. Fig. 2.9a gives the achieved effective rate on set $\mathcal{D}$ for $N = 12$ and $\theta = 0.0001$ (relaxed delay constraint). Similarly to the case of long term average value of C_D we see that for small values of β the sequential approach achieves slightly higher effective data rate. As before, the increase of β results in more reconciliation frames M required in the sequential case. This effect is not seen in the parallel case and for high values of β it performs better.

Fig. 2.9b illustrates the case when $N = 12$ and $\theta = 100$ (very stringent delay constraint). Similarly to before, we can see that for small values of β the sequential approach performs better than the parallel. As discussed, the efficiency loss is caused by the fact that the devoted part of the total achievable effective capacity $E_C^{\text{opt}}(\theta)$ to reconciliation (syndrome communication) is more than what is required. However, a higher β leads to an increase in the reconciliation information that needs to be sent, and the rate of the subcarriers in set $\breve{\mathcal{D}}$ will be fully or almost fully utilised and the parallel approach shows better performance for these values.

In the next two Fig.: 2.9c and 2.9d we show the performance of the two algorithms for higher value of $N = 64$. It is easy to see that regardless of the value of θ and β both algorithms perform identical

[7] *i.e.,* that the ratio of reconciliation information to data is small as seen from (2.25))

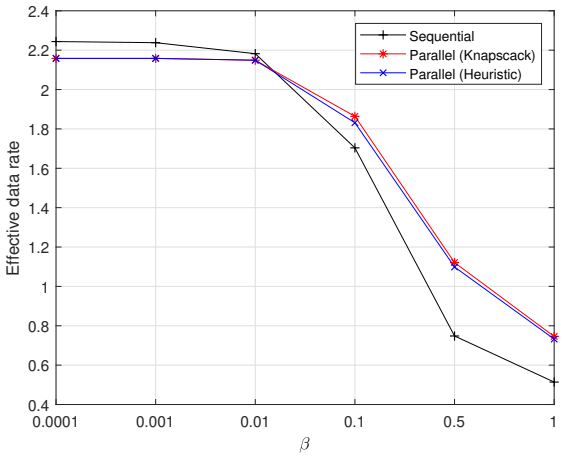


Figure 2.9: **a)** Effective data rate achieved by parallel and sequential approaches when $N = 12$, SNR= 5dB, $\theta = 0.0001, \kappa = 2$.

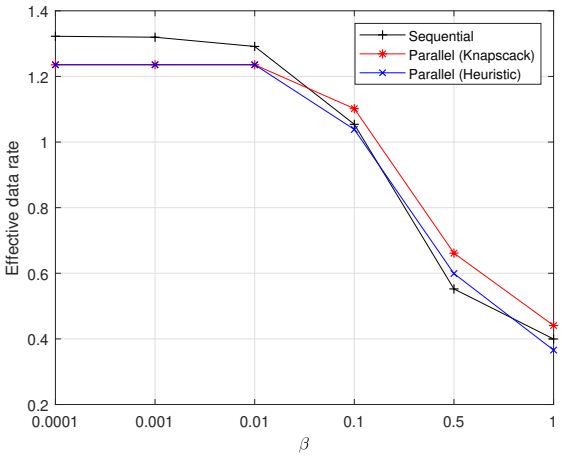


Figure 2.9: **b)** Effective data rate achieved by parallel and sequential approaches when $N = 12$, SNR= 5dB, $\theta = 100, \kappa = 2$.

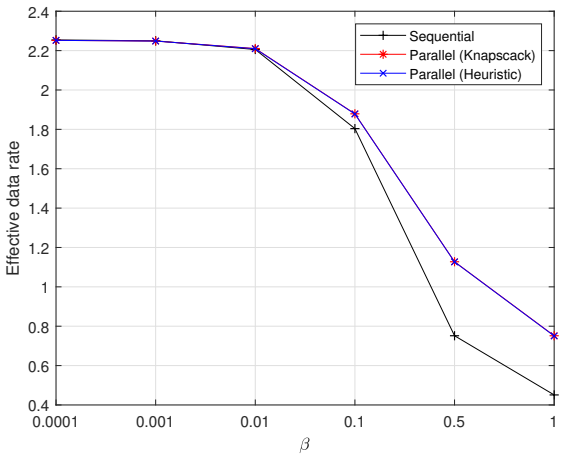


Figure 2.9: **c)** Effective data rate achieved by parallel and sequential approaches when $N = 64$, SNR= 5dB, $\theta = 0.0001, \kappa = 2$.

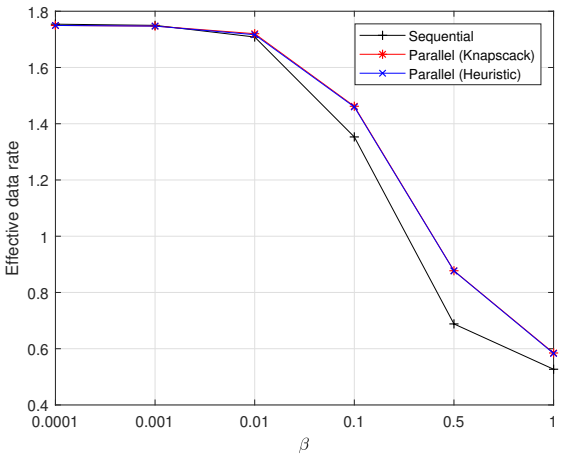


Figure 2.9: **d)** Effective data rate achieved by parallel and sequential approaches when $N = 64$, SNR= 5dB, $\theta = 100, \kappa = 2$.

or the parallel is better. In the previous case of $N = 12$ increasing θ might reduce the effectiveness of the parallel approach, however when $N = 64$ increasing θ does not incur such a penalty and the parallel is either identical to the sequential or outperforms it.

Another interesting fact from Fig. 2.9 is that looking at the parallel approach, it can easily be seen that in all cases the heuristic approach almost always performs as well as the optimal knapsack solution. The case of small values of θ is similar to the one when we work with long term average rate and choosing the best subcarriers for data transmission works as well as the optimal knapsack solution. Interestingly, *Algorithm 1* works well for high values of θ, too. This can be explained by the fact that when θ increases the rate on all of the subcarriers becomes similar and switching the subcarriers in set $\mathcal{D}$ does not incur high penalty.

2.8 Conclusions

In this work we discussed the possibility of using SKG in conjunction with PUF authentication protocols, illustrating this can greatly reduce the authentication and key generation latency compared to traditional mechanisms. Furthermore, we presented an AE scheme using SKG and a resumption protocol which further contribute to the system's security and latency reduction, respectively.

In addition, we explored the possibility of pipelining encrypted data transfer and SKG in a Rayleigh BF-AWGN environment. We investigated the maximization of the data transfer rate in parallel to performing SKG. We took into account imperfect CSI measurements and the effect of order statistics on the channel variance. Two scenarios were differentiated in our study: i) the optimal data transfer rate was found under power and security constraints, represented by the system parameters β and κ, which represent the minimum ratio of SKG rate to data rate and the maximum ratio of SKG rate to reconciliation rate; ii) by adding a delay constraint, represented by parameter θ, to the security and power constraint we found the optimal *effective data rate*.

To finalise our study we illustrated through numerical comparisons the efficiency of the proposed parallel method, in which SKG and data transfer are inter-weaved, to a sequential method where the two operations are done separately. The results of the two scenarios showed that in most of the cases the performance of both methods, parallel and sequential, is either equal or the parallel performs better. As the possible advantage of using the sequential is small and only applies in particular scenarios, we recommend the parallel scheme as a universal mechanism for general protocol design, when latency is an issue. Furthermore, a significant result is that although the optimal subcarrier scheduling is an NP hard $0-1$ knapsack problem, it can be solved in linear time using a simple heuristic algorithm with virtually no loss in performance.

References

[1] A. Mukherjee. Physical-layer security in the internet of things: Sensing and communication confidentiality under resource constraints. *Proceedings of the IEEE*, 103(10):1747–1761, Oct 2015.

[2] A. Yener and S. Ulukus. Wireless physical-layer security: Lessons learned from information theory. *Proceedings of the IEEE*, 103(10):1814–1825, Oct 2015.

[3] D. Karatzas, A. Chorti, N. M. White, and C. J. Harris. Teaching old sensors new tricks: Archetypes of intelligence. *IEEE Sensors Journal*, 7(5):868–881, May 2007.

[4] 3GPP TR 33.825 V0.3.0, Study on the Security for 5G URLLC (Release 16). 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects, available online https://www.3gpp.org/ftp/Specs/archive/33_series/33.825/.

[5] Arsenia Chorti, Camilla Hollanti, Jean-Claude Belfiore, and H. Vincent Poor. Physical layer security: A paradigm shift in data confidentiality. *Lecture Notes in Electrical Engineering*, 358, 01 2016.

[6] A. Chorti, K. Papadaki, and H. V. Poor. Optimal power allocation in block fading channels with confidential messages. *IEEE Transactions on Wireless Communications*, 14(9):4708–4719, Sep. 2015.

[7] A. Chorti, S. M. Perlaza, Z. Han, and H. V. Poor. On the resilience of wireless multiuser networks to passive and active eavesdroppers. *IEEE Journal on Selected Areas in Communications*, 31(9):1850–1863, Sep. 2013.

[8] A. Chorti and H. V. Poor. Achievable secrecy rates in physical layer secure systems with a helping interferer. In *2012 International Conference on Computing, Networking and Communications (ICNC)*, pages 18–22, Jan 2012.

[9] M. Mitev, A. Chorti, and M. Reed. Subcarrier scheduling for joint data transfer and key generation schemes in multicarrier systems. In *2019 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6, Dec 2019.

[10] M. Rodrigues I. Darwazeh Y. Kanaras, A. Chorti. An optimum detection for a spectrally efficient non orthogonal FDM system. In *Proc. 13th Int. OFDM WS*, pages 65–68, Aug 2008.

[11] A. Chorti and H. V. Poor. Faster than Nyquist interference assisted secret communication for OFDM systems. In *2011 Asilomar Conf. Signals, Systems and Computers (ASILOMAR)*, pages 183–187, Nov 2011.

[12] A. Chorti. Helping interferer physical layer security strategies for M-QAM and M-PSK systems. In *2012 46th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6, March 2012.

[13] Matti Latvaaho and Kari Leppänen. Key drivers and research challenges for 6G ubiquitous wireless intelligence. Published online by the University of Oulu, October 2019.

[14] U. M. Maurer. Secret key agreement by public discussion from common information. *IEEE Transactions on Information Theory*, 39(3):733–742, May 1993.

[15] R. Ahlswede and I. Csiszar. Common randomness in information theory and cryptography. i. secret sharing. *IEEE Transactions on Information Theory*, 39(4):1121–1132, July 1993.

[16] C. Ye, A. Reznik, and Y. Shah. Extracting secrecy from jointly Gaussian random variables. In *2006 IEEE International Symposium on Information Theory*, pages 2593–2597, July 2006.

[17] Blaise Gassend, Dwaine Clarke, Marten van Dijk, and Srinivas Devadas. Silicon physical random functions. In *Proceedings of the 9th ACM Conference on Computer and Communications Security*, CCS '02, page 148–160, New York, NY, USA, 2002. Association for Computing Machinery.

[18] Ravikanth Pappu, Ben Recht, Jason Taylor, and Neil Gershenfeld. Physical one-way functions. *Science*, 297(5589):2026–2030, 2002.

[19] Roel Maes and Ingrid Verbauwhede. *Physically Unclonable Functions: A Study on the State of the Art and Future Research Directions*, pages 3–37. 10 2010.

[20] H. Schotten A. Weinand, M. Karrenbauer. Security solutions for local wireless networks in control applications based on physical layer security. *IFAC-PapersOnLine*, 51:32–39, 2018.

[21] A. Mukherjee, S. A. A. Fakoorian, J. Huang, and A. L. Swindlehurst. Principles of physical layer security in multiuser wireless networks: A survey. *IEEE Communications Surveys Tutorials*, 16(3):1550–1573, Third 2014.

[22] Arsenia Chorti. A study of injection and jamming attacks in wireless secret sharing systems. In *in Proc. Workshop on Communication Security (WCS)*, 03 2017.

[23] The Transport Layer Security (TLS) Protocol Version 1.3. RFC 8446 (2018). Rescorla, E., available online https://rfc-editor.org/rfc/rfc8446.txt.

[24] Nimrod Aviram, Kai Gellert, and Tibor Jager. Session resumption protocols and efficient forward security for TLS 1.3 0-RTT. Cryptology ePrint Archive, Report 2019/228, 2019. https://eprint.iacr.org/2019/228.

[25] Mihir Bellare and Chanathip Namprempre. Authenticated encryption: Relations among notions and analysis of the generic composition paradigm. *J. Cryptol.*, 21(4):469–491, September 2008.

[26] Ted Krovetz and Phillip Rogaway. The software performance of authenticated-encryption modes. In *FSE, Lecture Notes in Computer Science*, 2011.

[27] S. Koteshwara and A. Das. Comparative study of authenticated encryption targeting lightweight IoT applications. *IEEE Design Test*, 34(4):26–33, Aug 2017.

[28] Dapeng Wu and R. Negi. Effective capacity: a wireless link model for support of quality of service. *IEEE Transactions on Wireless Communications*, 2(4):630–643, July 2003.

[29] Wenjie Che, Mitchell Martin, Goutham Pocklassery, Venkata K. Kajuluri, Fareena Saqib, and James F. Plusquellic. A privacy-preserving, mutual puf-based authentication protocol. *Cryptography*, 1:3, 2016.

[30] Blaise Gassend, Dwaine Clarke, Marten van Dijk, and Srinivas Devadas. Silicon physical random functions. In *Proceedings of the 9th ACM Conference on Computer and Communications Security*, CCS '02, page 148–160, New York, NY, USA, 2002. Association for Computing Machinery.

[31] C. Marchand, L. Bossuet, U. Mureddu, N. Bochard, A. Cherkaoui, and V. Fischer. Implementation and characterization of a physical unclonable function for IoT: A case study with the TERO-PUF. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 37(1):97–109, Jan 2018.

[32] Jorge Guajardo, Sandeep S. Kumar, Geert-Jan Schrijen, and Pim Tuyls. FPGA intrinsic PUFs and their use for IP protection. In *Proceedings of the 9th International Workshop on Cryptographic Hardware and Embedded Systems*, CHES '07, page 63–80, Berlin, Heidelberg, 2007. Springer-Verlag.

[33] J. Aarestad, P. Ortiz, D. Acharyya, and J. Plusquellic. HELP: A hardware-embedded delay PUF. *IEEE Design Test*, 30(2):17–25, April 2013.

[34] Armin Babaei and Gregor Schiele. Physical unclonable functions in the internet of things: State of the art and open challenges. In *Sensors*, 2019.

[35] Pramod Maurya and Satya Bagchi. A secure PUF-based unilateral authentication scheme for RFID system. *Wireless Personal Communications*, 103, 05 2018.

[36] M. Yu, M. Hiller, J. Delvaux, R. Sowell, S. Devadas, and I. Verbauwhede. A lockdown technique to prevent machine learning on PUFs for lightweight authentication. *IEEE Transactions on Multi-Scale Computing Systems*, 2(3):146–159, July 2016.

[37] Jeff Calhoun, Cyrus Minwalla, Charles Helmich, Fareena Saqib, Wenjie Che, and J. Plusquellic. Physical unclonable function (PUF)-based e-cash transaction protocol (PUF-Cash). *Cryptography*, 3:18, 07 2019.

[38] M. N. Aman, K. C. Chua, and B. Sikdar. Mutual authentication in IoT systems using physical unclonable functions. *IEEE Internet of Things Journal*, 4(5):1327–1340, Oct 2017.

[39] Jeroen Delvaux, Roel Peeters, Dawu Gu, and Ingrid Verbauwhede. A survey on lightweight entity authentication with strong PUFs. *ACM Comput. Surv.*, 48(2), October 2015.

[40] Suman Jana, Sriram Nandha Premnath, Mike Clark, Sneha K. Kasera, Neal Patwari, and Srikanth V. Krishnamurthy. On the effectiveness of secret key extraction from wireless signal strength in real environments. In *Proceedings of the 15th Annual International Conference on Mobile Computing and Networking*, MobiCom '09, page 321–332, New York, NY, USA, 2009. Association for Computing Machinery.

[41] Theodore Rappaport. *Wireless Communications: Principles and Practice*. Prentice Hall PTR, USA, 2nd edition, 2001.

[42] J. Wan, A. B. Lopez, and M. A. Al Faruque. Exploiting wireless channel randomness to generate keys for automotive cyber-physical system security. In *2016 ACM/IEEE 7th International Conference on Cyber-Physical Systems (ICCPS)*, pages 1–10, April 2016.

[43] B. Zan, M. Gruteser, and F. Hu. Key agreement algorithms for vehicular communication networks based on reciprocity and diversity theorems. *IEEE Transactions on Vehicular Technology*, 62(8):4020–4027, Oct 2013.

[44] Yicong Liu, Jiwu Jing, and Jun Yang. Secure underwater acoustic communication based on a robust key generation scheme. In *2008 9th International Conference on Signal Processing*, pages 1838–1841, Oct 2008.

[45] I. U. Zaman, A. B. Lopez, M. A. A. Faruque, and O. Boyraz. Physical layer cryptographic key generation by exploiting PMD of an optical fiber link. *Journal of Lightwave Technology*, 36(24):5903–5911, Dec 2018.

[46] D. Tian, W. Zhang, J. Sun, and C. Wang. Physical-layer security of visible light communications with jamming. In *2019 IEEE/CIC International Conference on Communications in China (ICCC)*, pages 512–517, Aug 2019.

[47] J. Zhang, T. Q. Duong, A. Marshall, and R. Woods. Key generation from wireless channels: A review. *IEEE Access*, 4:614–626, 2016.

[48] J. K. Tugnait, Lang Tong, and Zhi ding. Single-user channel estimation and equalization. *IEEE Signal Processing Magazine*, 17(3):17–28, May 2000.

[49] William C. Jakes and Donald C. Cox. *Microwave Mobile Communications*. Wiley-IEEE Press, 1994.

[50] H. Liu, Y. Wang, J. Yang, and Y. Chen. Fast and practical secret key extraction by exploiting channel response. In *2013 Proceedings IEEE INFOCOM*, pages 3048–3056, April 2013.

[51] Suhas Mathur, Wade Trappe, Narayan Mandayam, Chunxuan Ye, and Alex Reznik. Radio-telepathy: Extracting a secret key from an unauthenticated wireless channel. In *Proceedings of the 14th ACM International Conference on Mobile Computing and Networking*, MobiCom '08, page 128–139, New York, NY, USA, 2008. Association for Computing Machinery.

[52] S. T. Ali, V. Sivaraman, and D. Ostry. Eliminating reconciliation cost in secret key generation for body-worn health monitoring devices. *IEEE Transactions on Mobile Computing*, 13(12):2763–2776, Dec 2014.

[53] Suhas Mathur, Robert Miller, Alexander Varshavsky, Wade Trappe, and Narayan Mandayam. Proximate: Proximity-based secure pairing using ambient wireless signals. In *Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services*, MobiSys '11, pages 211–224, New York, NY, USA, 2011. Association for Computing Machinery.

[54] M. Mitev, A. Chorti, E.V. Belmega, and M.J. Reed. Man-in-the-middle and denial of service attacks in wireless secret key generation. In *Proc. IEEE Global Commun. (GLOBECOM)*, Big Island, HI, 2019.

[55] Intrinsic-id company. https://www.intrinsic-id.com/sram-puf.

[56] ICTK holdings corporation. https://ictk-puf.com/puf-technology.

[57] A. Maiti, I. Kim, and P. Schaumont. A robust physical unclonable function with enhanced challenge-response set. *IEEE Transactions on Information Forensics and Security*, 7(1):333–345, Feb 2012.

[58] Monis Akhlaq, Baber Aslam, Muzammil A. Khan, and M. Noman Jafri. Comparative analysis of IEEE 802.1x authentication methods. In *Proceedings of the 11th Conference on 11th WSEAS International Conference on Communications - Volume 11*, ICCOM'07, page 1–6, Stevens Point, Wisconsin, USA, 2007. World Scientific and Engineering Academy and Society (WSEAS).

[59] A. Chiornită, L. Gheorghe, and D. Rosner. A practical analysis of EAP authentication methods. In *9th RoEduNet IEEE International Conference*, pages 31–35, June 2010.

[60] Urbi Chatterjee, Rajat Chakraborty, and Debdeep Mukhopadhyay. A PUF-based secure communication protocol for IoT. *ACM Transactions on Embedded Computing Systems*, 16:1–25, 04 2017.

[61] M. N. Aman, M. H. Basheer, and B. Sikdar. Two-factor authentication for IoT with location information. *IEEE Internet of Things Journal*, 6(2):3335–3351, April 2019.

[62] M. H. Mahalat, S. Saha, A. Mondal, and B. Sen. A PUF based light weight protocol for secure WiFi authentication of IoT devices. In *2018 8th International Symposium on Embedded Computing and System Design (ISED)*, pages 183–187, Dec 2018.

[63] An Braeken. Puf based authentication protocol for IoT. *Symmetry*, 10:352, 08 2018.

[64] Y. Yilmaz, S. R. Gunn, and B. Halak. Lightweight PUF-based authentication protocol for IoT devices. In *2018 IEEE 3rd International Verification and Security Workshop (IVSW)*, pages 38–43, July 2018.

[65] S. Ahmad, A. H. Mir, and G. R. Beigh. Latency evaluation of extensible authentication protocols in WLANs. In *2011 Fifth IEEE International Conference on Advanced Telecommunication Systems and Networks (ANTS)*, pages 1–5, Dec 2011.

[66] P. Gope and B. Sikdar. Lightweight and privacy-preserving two-factor authentication scheme for IoT devices. *IEEE Internet of Things Journal*, 6(1):580–589, Feb 2019.

[67] A. Ometov, P. Masek, L. Malina, R. Florea, J. Hosek, S. Andreev, J. Hajny, J. Niutanen, and Y. Koucheryavy. Feasibility characterization of cryptographic primitives for constrained (wearable) IoT devices. In *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, pages 1–6, March 2016.

[68] J. Cho and W. Sung. Efficient software-based encoding and decoding of BCH codes. *IEEE Transactions on Computers*, 58(7):878–889, July 2009.

[69] C. Chen and M. A. Jensen. Secret key establishment using temporally and spatially correlated wireless channel coefficients. *IEEE Transactions on Mobile Computing*, 10(2):205–215, Feb 2011.

[70] J. Zhang, A. Marshall, R. Woods, and T. Q. Duong. Efficient key generation by exploiting randomness from channel responses of individual OFDM subcarriers. *IEEE Transactions on Communications*, 64(6):2578–2588, June 2016.

[71] J. Zhang, B. He, T. Q. Duong, and R. Woods. On the key generation from correlated wireless channels. *IEEE Communications Letters*, 21(4):961–964, April 2017.

[72] C. Saiki and A. Chorti. A novel physical layer authenticated encryption protocol exploiting shared randomness. In *2015 IEEE Conference on Communications and Network Security (CNS)*, pages 113–118, Sep. 2015.

[73] Q. Wang, H. Su, K. Ren, and K. Kim. Fast and scalable secret key generation exploiting channel phase randomness in wireless networks. In *2011 Proceedings IEEE INFOCOM*, pages 1422–1430, April 2011.

[74] C. Ye, S. Mathur, A. Reznik, Y. Shah, W. Trappe, and N. B. Mandayam. Information-theoretically secret key generation for fading wireless channels. *IEEE Transactions on Information Forensics and Security*, 5(2):240–254, June 2010.

[75] Christopher Huth, Ren Guillaume, Thomas Strohm, Paul Duplys, Irin Ann Samuel, and Tim Gneysu. Information reconciliation schemes in physical-layer security. *Comput. Netw.*, 109(P1):84–104, November 2016.

[76] Li Guyue, Zheying Zhang, Yi Yu, and Aiqun Hu. A hybrid information reconciliation method for physical layer key generation. *Entropy*, 21:688, 07 2019.

[77] P. Treeviriyanupab, P. Sangwongngam, K. Sripimanwat, and O. Sangaroon. BCH-based Slepian-Wolf coding with feedback syndrome decoding for quantum key reconciliation. In *2012 9th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, pages 1–4, May 2012.

[78] J. Etesami and W. Henkel. LDPC code construction for wireless physical-layer key reconciliation. In *2012 1st IEEE International Conference on Communications in China (ICCC)*, pages 208–213, Aug 2012.

[79] C. H. Bennett, G. Brassard, C. Crepeau, and U. M. Maurer. Generalized privacy amplification. *IEEE Transactions on Information Theory*, 41(6):1915–1923, Nov 1995.

[80] Furui Zhan and Nianmin Yao. On the using of discrete wavelet transform for physical layer key generation. *Ad Hoc Networks*, 64:22 – 31, 2017.

[81] M. Bloch, J. Barros, M. R. D. Rodrigues, and S. W. McLaughlin. Wireless information-theoretic security. *IEEE Transactions on Information Theory*, 54(6):2515–2534, June 2008.

[82] M. Mitev, A. Chorti, and M. Reed. Optimal resource allocation in joint secret key generation and data transfer schemes. In *2019 15th International Wireless Communications Mobile Computing Conference (IWCMC)*, pages 360–365, June 2019.

[83] E. V. Belmega and A. Chorti. Protecting secret key generation systems against jamming: Energy harvesting and channel hopping approaches. *IEEE Transactions on Information Forensics and Security*, 12(11):2611–2626, Nov 2017.

[84] M. Medard. The effect upon channel capacity in wireless communications of perfect and imperfect knowledge of the channel. *IEEE Transactions on Information Theory*, 46(3):933–946, May 2000.

[85] Hong-Chuan Yang and Mohamed-Slim Alouini. *Order Statistics in Wireless Communications: Diversity, Adaptation, and Scheduling in MIMO and OFDM Systems*. Cambridge University Press, USA, 1st edition, 2011.

[86] Silvano Martello and Paolo Toth. *Knapsack Problems: Algorithms and Computer Implementations*. John Wiley and Sons, Inc., USA, 1990.

[87] H. Kellerer, U. Pferschy, and D. Pisinger. *Knapsack Problems*. Springer, Berlin, Germany, 2004.

[88] Vijay V. Vazirani. *Approximation Algorithms*. Springer-Verlag, Berlin, Heidelberg, 2001.

[89] Cheng-Shang Chang. Stability, queue length, and delay of deterministic and stochastic queueing networks. *IEEE Transactions on Automatic Control*, 39(5):913–931, May 1994.

[90] J. Gärtner. On large deviation from invariant measure. *Theory Prob. Appl.*, 22:24–39, 1977.

[91] Richard Ellis. Large deviations for a general class of random vectors. *The Annals of Probability*, 12, 02 1984.

[92] T. Abrao, S. Yang, L. D. H. Sampaio, P. J. E. Jeszensky, and L. Hanzo. Achieving maximum effective capacity in OFDMA networks operating under statistical delay guarantee. *IEEE Access*, 5:14333–14346, 2017.

[93] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, USA, 2004.

Chapter 3

Application of Change Point Analysis in Edge Resource Allocation and Intrusion Detection

3.1 Introduction

Edge computing emerges as a critical element in future networks, provisioning storage and computation resources in the proximity of end devices to provide low latency services. The joint allocation and management of communication, computing and storage resources will improve the quality of service (QoS) and user experience, especially for various delay-sensitive applications. At the same time, software-defined networking (SDN) is a technology that can help bridge the gap when combining Edge computing and traditional Clouds. For example, the SDN controller can make decisions on whether tasks should be uploaded and processed in the Cloud or at the Edge. The advancement of Edge computing poses many challenges, e.g., in the deployment and management of distributed resources; in parallel, SDNs are prone to new security threats due to the separation of the control and data planes. In this Chapter, we first focus on content distribution at the Edge using change point (CP) analysis. Next, motivated by the good performance of the developed algorithms, we investigate their application in intrusion detection in software defined wireless sensor networks (SDWSNs), showcasing that the wide range of applications that can be covered.

Beginning with resource allocation at the Edge servers, we propose a novel and flexible approach exploiting popular virtualization technologies, such as unikernels [1] or containers; as a use case we consider video content distribution, which accounts for more than 70% of the global IP traffic. The core idea in our proposal is that virtual servers could hold individual video content and could be "live" for as long as there is corresponding demand; in case of an increase in demand, more replicas of the virtual servers could be put up, or alternatively put down if the demand dies off. We note that bringing up or down a unikernel is typically very fast, with reported numbers for the boot time as little as 20 milliseconds [2].

In this context, due to high volatility in the respective demand, it is important for video content delivery infrastructures to rapidly detect and respond to changes in "content popularity" dynamics. We explore the employment of on-line CP analysis to implement real-time, autonomous and low-complexity video content popularity detection. Our proposal, denoted as *real-time change point detector (RCPD)*, estimates the existence, the number and the direction of changes on the average number of video visits by combining: (i) off-line and on-line CP detection algorithms; (ii) an improved time-series segmentation heuristic for the reliable detection of multiple CPs; and (iii) two algorithms for the identification of the direction of changes. The proposed detector is validated against synthetic data, as well as a large database of real YouTube video visits.

Finally, we note that customarily CP analysis is employed in the detection of anomalies in times

series. Therefore, as a natural extension of this work, we further consider the application of the RCPD for anomaly detection in SDWSNs. The centralization and the planes' separation turn SDNs vulnerable to new security threats in the general context of distributed denial of service (DDoS) attacks, which carry over to SDWSNs. State-of-the-art approaches to identify DDoS do not always take into consideration restrictions in typical WSNs, e.g., computational complexity and power constraints, while further performance improvement is always a target. Our objective in this study is to propose a lightweight but very efficient DDoS attack detection approach using the RCPD.

3.2 Contributions and Chapter Organization

3.2.1 CP Analysis in Resource Allocation

Video content is projected to account for 82% of the global Internet traffic by the end of 2020, significantly increased from 72% in 2016 [3]. In parallel, novel emerging networking, Cloud and Edge computing paradigms with significant elasticity capabilities appeared recently, e.g., SDNs [4], Cloud orchestration proposals [5] and content distribution networks (CDNs) [6]. These advances offer the means to respond quickly to changes in content popularity dynamics with appropriate adaptations, e.g., in terms of efficient server resource allocation schemes, load balancing or content caching. As a result, the early detection of changes in content popularity [7], [8] is proving a highly important topic and can have a significant impact on the network traffic and the utilization of servers.

So far, the vast majority of research efforts have focused on the *prediction* of content popularity dynamics, as opposed to their *real time detection*, which is the focus of this study. There is a multitude of reasons as to why the precision of even state-of-the-art prediction algorithms can be impaired. A variety of factors – both from the digital and the physical world – can influence the users' Internet surfing behavior, e.g., [7]: (i) the quality, type (e.g., commercial or user-provided) and life-time of content; (ii) its relevance to users and physical events; (iii) the social interactions between users; and (iv) the content promotion strategies involved. Importantly, mid-term and long-term content popularity prediction [9] – and corresponding adaptations in the network or cloud environment – can prove highly inaccurate [10] and thus result in sub-optimal service planning, provisioning, and utilization of resources or violation of service level agreements.

In this work, to address the aforementioned shortcomings of the commonly employed prediction algorithms, we propose a corresponding detector, referred to as the "real-time change point detector" (RCPD). The RCPD is compatible with modern, flexible networking and Cloud approaches, that are highly adaptive and can respond to short-term network dynamics. With accurate, on-line content popularity detection, discrepancies between inaccurate predictions and actual changes can be alleviated. The RCPD is real-time, lightweight, accurate and is parameterized autonomously by analyzing historical data.

In the RCPD, we employ the CP detection theory and algorithms; their suitability is confirmed against a large number of synthetic as well as real YouTube video datasets. In this contribution, the early detection of changes in the average content popularity is addressed with a novel CP detection methodology, consisting of a training phase, using historical data, and, an on-line phase. In the training phase, we employ a modified off-line CP detection scheme to configure the on-line (sequential) algorithm's parameters. This approach is shown to greatly improve the accuracy of the on-line detector, as in essence, the algorithm parameterization is not arbitrary but rather extracted from corresponding historical data. To the best of our knowledge, this was the first proposal in the literature on combining retrospective (off-line) and sequential (on-line) CP detection schemes in a single algorithm operating autonomously (i.e., without manual configuration of parameters).

Besides that, our approach complements the off-line scheme with an improved time-series segmentation heuristic for the detection of multiple CPs. Furthermore, we propose two possible variations for the on-line CP algorithm, the first based on the standard cumulative sum (CUSUM) procedure [11]

and the second on the ratio-type CUSUM procedure [12][1]. Additionally, we introduce two alternative indicators to detect the direction of changes: the first one is directly derived from the statistical test of the on-line CP procedure, while the second is based on a modified exponential moving average filter, extensively used in econometrics. As discussed in Sections 3.4 and 3.5, the RCPD combines all the above mentioned algorithmic elements, and is based on sufficiently general and convenient assumptions. Moreover, unlike other approaches e.g., [13], we employ methods that allow dependence between observations (in the form of $t-$dependence), leading to more realistic assumptions for the statistical structure of the content visits.

We evaluate the proposed detector and its individual algorithmic components (i.e., the off-line / on-line test statistics, the time-series segmentation algorithm and the trend indicator), over synthetic and real YouTube content views data. Our experiments using synthetic data, generated by an autoregressive moving average (ARMA) filter, demonstrate:

- The superior performance of the proposed time-series segmentation heuristic over the standard approach, improving the true alarm rates by up to 43%.
- The ability of the two proposed trend indicators to identify the direction of estimated changes, with successful identification rates exceeding 99%, in all cases.
- The RCPD performance; the true alarm rates surpass 94% for medium / large changes in the mean number of content views, while the corresponding CP identification lag ranges between 10 to 20 instances, confirming the real-time operation of the detector. On the other hand, the RCPD achieves very small false alarm rates, well within the limits of the statistical error specified by the chosen significance level of the CP algorithms.

Furthermore, our tests on real YouTube content views datasets show that:

- YouTube video views match the underlying assumptions of the RCPD, i.e., the content popularity time-series datasets can be modeled as t-dependent.
- The RCPD can detect CPs in more than 70% of the videos in our dataset, implying a sufficiently high number of content popularity changes and the suitability of the CP theory framework for content popularity detection.
- The successful CP direction identifications exceed 91%, i.e., the proposed trend indicators work for real data.
- The average dynamic time warping (DTW) distance [14], [15] between the identified CPs and a benchmark off-line algorithm was estimated to be 52 time instances on average, showcasing the rapid responsiveness of the RCPD.
- The overall processing cost of the RCPD is very low; notably, it took less than one second to process 882 videos on a typical personal computer (PC).

As a proof-of-concept, we demonstrate the applicability of the proposed algorithm in a real load balancing scenario. We provide a set of measurements showcasing improvements in terms of the clients' connectivity time to download specific content, without a significant impact on the utilization of the content servers. This is achieved due to the deployment of additional content caches, an event triggered by the output of the proposed RCPD detector.

[1]The advantage of ratio-type CUSUM is that it does not require the estimation of long-run covariance (variance) matrices, which is the case for the standard CUSUM method.

3.2.2 CP Analysis for Anomaly Detection in SDWSNs

Next, we explore the application of the RCPD for anomaly (intrusion) detection in SDWSNs. The SDN paradigm was devised to simplify network management, avoid configuration errors and automate infrastructure sharing in wired networks [16]. The aforementioned benefits motivated the discussion of combining SDN and WSNs as a solution to many WSN challenges, in particular concerning flexibility and resource reuse [17]. This combination is referred to as SDWSN. The SDWSN approach decouples the control plane from the data plane and centralizes the control decisions; its main characteristic is the ability to program the network operation dynamically [18]. Recent results show that SDWSNs can perform as well as the IPv6 routing protocol for low-power and lossy networks (RPL) [19].

On the other hand, the SDN centralization and the planes' separation turn the network vulnerable to new security threats (explained in Section 3.9.1), a property that is inadvertently passed on to SDWSNs. Shielding SDNs from these vulnerabilities has already attracted a lot of attention in the literature with proposals to implement attack detection in IoT networks using SDN. Overall, in the case of SDWSNs, due to the resource constraints of the nodes, most of the security mechanisms designed for non-resource constrained SDNs have to be adapted or redesigned. This is one of the major challenges for SDWSN security.

Considering the limitations of previous works, our main objective is to propose a mechanism for DDoS detection with, i) a high detection rate, and, ii) low complexity, so that it would be suitable for "restricted" networks. To this end, we propose the employment of the RCPD [20] [21]. We study two DDoS attacks: a false data flow forwarding (FDFF) attack, and a false neighbor information (FNI) attack, chosen to illustrate the proposed algorithm's capabilities in the case of specific SDWSN vulnerabilities that exhibit largely different behavior. Both attacks are explained in Section **??**. We have tested our approach on the IT-SDN framework[2] [19] and our results show that we can detect these attacks with a detection rate close to 100%, improving the state of the art; importantly, it is further possible to gain insight regarding the *type of the attack*, based on the metric that provides the quickest detection, a feature, that to the best of our knowledge, breaks new ground in the domain of DDoS analysis for SDWSNs.

3.2.3 Chapter Organization

The rest of the Chapter is organized as follows. In Section 3.3 we provide a comprehensive literature review of related topics. In Section 3.4, we present the off-line (training) phase of the RCPD algorithm, while the on-line phase is discussed in Section 3.5. In Section 3.6, we present four experiments over synthetic video content data, providing an extensive validation of the RCPD and its subroutines, while in Section 3.7, we discuss corresponding experiments using a database of real YouTube video views. In Section 3.8, we demonstrate the load balancing gains achieved through the use of the RCPD, in a realistic content provisioning scenario.

Moving to intrusion detection in SDWSNs, Section **??** illustrates the FDFF and FNI attacks and their impact on the network performance. Experimental methods are presented in Section 3.9.3 and results on intrusion detection using the RCPD are presented in Section 3.10.

Finally, Section 3.11 concludes the Chapter.

3.3 Related Works

In this Section, we discuss how this work relates to the literature of video content popularity prediction, and, anomaly detection in general and in SDNs and SDWSNs in particular.

The topic of content popularity attracted a lot of attention in recent years, because of its importance in a number of applications, such as network dimensioning (e.g., capacity planning or scaling of resources), on-line marketing (e.g., advertising, recommendation systems) or real-world outcome

[2] http://www.larc.usp.br/users/cbmargi/www/it-sdn/

prediction (e.g., analysis of economical trends) [7]. The main approaches used for content popularity estimation can be categorized as: (i) cumulative growth studies, estimating the "amount of attention" from the publication instance to the prediction moment [8]; (ii) temporal analysis approaches, i.e., how content visits evolve over time [22]; and (iii) clustering methods of content with similar popularity trends [9]. We note that many content popularity studies consider the aggregate behavior of a particular content, e.g., [8], [22], whereas we study the real-time behavior of video views time-series. In addition, studies using clustering methods [9] are based on content popularity prediction and adopt parametric models, unlike the RCPD algorithm that is non-parametric.

To the best of our knowledge, our conference paper [23] is the first in the literature proposing CP techniques [24] for content popularity detection. The RCPD algorithm falls into the general category of anomaly detection [25]; in essence, we assume that no changes in popularity constitutes the normal behavior of video content and search for deviations from this behavior. Non-parametric anomaly detection has typically been considered for the detection of abnormalities in the network traffic. As an example, in [26] an algorithm was proposed based on the Shiryaev-Roberts procedure for anomaly detection in computer network traffic. In [27] and [28], CUSUM based approaches were introduced for the detection of SYN attacks.

As opposed to previous content popularity prediction works, in this Chapter we introduce a novel CP detection methodology that provides accurate, lightweight, autonomous and on-line CP detection of content popularity. We formulate the detection of a change in the average content popularity as a statistical hypothesis test and employ non-parametric procedures to avoid a particular distribution assumption (such as a specific copula model). This context ensures low convergence time since it avoids estimating a large number of model parameters and restrictive assumptions that may not match the structure of the time-series. Furthermore, we avoid problems of parametric models that require parameters' fitting and selection, which become challenging as new data become available. In the proposed RCPD algorithm, an off-line phase specifies important parameters for the on-line phase; these parameters are re-evaluated dynamically after a detected CP. Our load-balancing experiments, elaborated in [6], demonstrate the RCPD's behavior in a real test-bed deployment.

Up to now there are only a handful of proposals addressing the challenges of new flexible networking and Cloud architectures accounting for content popularity. Exceptions include [29] in which a machine learning approach to content popularity prediction is applied for a Fog radio access network (RAN) environment, and, our recent papers [6] and [23]. In [6], the algorithm – outlined in [23] and presented extensively here – is integrated into an elastic content distribution network (CDN) framework based on lightweight Cloud capabilities using Unikernels. [6] focuses on the platform details rather than on the CP algorithm; it confirms experimentally the suitability of the latter for relevant flexible network and cloud architectures. A detailed description of the proposed CP detection algorithm is presented in the following Sections, along with a rich set of validation results.

Further examples of parametric anomaly detection methods include [30], in which a bivariate sequential generalized likelihood ratio test (LRT) was proposed, accounting for the packet rate – assumed to follow a Poisson distribution – and the packet size – assumed to follow a normal distribution. Other parametric anomaly detection approaches assume a particular underlying process for the normal behavior and search for anomalies on the residuals of the process. For example, in [31], Kalman filtering is combined with several CP methods, such as CUSUM and LRT, to detect anomalies in origin-destination flows. In [32], traffic flows (in the form of TCP's finite state machine), are modeled using Markov chains and an anomaly detection mechanism based on the generalized LRT algorithm is developed.

On the other hand, looking at existing literature in SDN anomaly detection, the authors in [33] proposed *softhings*, an SDN-based IoT framework with security support. The framework was developed for OpenFlow [18], which, however, can be a limiting factor for its use in networks composed of low-end nodes. The use of support vector machines (SVM) was proposed to detect control plane attacks; it was shown that a detection rate of around 96% and 98% could be achieved. The algorithm was tested in Mininet, simulating scenarios with only five nodes and considering one node as attacker. Furthermore

Yin *et al.* [34] developed the framework SD-IoT, which included a security system for DDoS attacks detection, based on the difference of packets received by the controller. The difference was calculated using the *cosine similarity* method. This mechanism was devised for networks where all the nodes had periodic communication with the controller, which could be not optimal for very "restricted" networks with low-end nodes. The authors tested their proposal through simulations using Mininet. The network size was not explicitly specified, but can be inferred to be around 50 to 60 nodes.

Furthermore, Wang *et al.* [35] proposed an SDWSN trust management and routing mechanism. They compared their proposal to SDN-WISE when both networks were under attack. The focus of the work was on the selective forwarding attacks and new flow requests. The first attack applied to any type of WSNs, while the second was specific to SDNs. The mechanism was tested in simulations with 100 nodes, varying the number of attackers between 5 and 20. Their results showed an attack detection rate between 90% and 96% when 5 nodes were attackers, and between 60% and 79% when 20 nodes were attackers. Compared to these previous works, our proposal for the employment of the RCPD is SDWSN anomaly detection has the advantages of being i) lightweight, ii) fast and iii) highly accurate as will be demonstrated in Sections 3.9 to 3.10.

We begin the description of the RCPD by first elaborating on the off-line and on-line phases in Sections 3.4 and 3.5 respectively, where we also provide the corresponding pseudo-code.

3.4 Training (Off-line) Phase

In this Section, the training phase of the algorithm is discussed and the fundamental components of the off-line scheme are presented. We note that standard off-line CP schemes can only detect a single CP. To address the issue of detection of multiple CPs, we modify the basic algorithm with a novel time-series segmentation heuristic, that belongs to the family of binary segmentation algorithms.

3.4.1 Basic Off-line Approach

Let $\{X_n : n \in \mathbb{N}\}$ be a sequence of r- dimensional random vectors (r.v.). The first dimension represents the number of views for a specific video content within a time period $n \in \{1, \ldots, N\}$, while the other dimensions could be optionally used to represent other content popularity features, such as likes, comments, etc. We assume that $X_1, ..., X_N$ can be written as,

$$X_n = \mu_n + Y_n, \quad 1 \leqslant n \leqslant N \tag{3.1}$$

where $\{\mu_n : n \in \mathbb{N}\}$ is the mean value of video visits, $\{Y_n : n \in \mathbb{N}\}$ a random component with zero mean $\mathbb{E}[Y_n] = 0$ and positive definite covariance matrix, $\mathbb{E}\left[Y_n Y_n^T\right] = \Sigma$, while $\mathbb{E}[\cdot]$ denotes expectation. We further assume that the time-series is t-dependent, implying that for $t_1, t_2, t \in \mathbb{N}$, Y_{t_1} is independent of Y_{t_2} if $|t_1 - t_2| > t$.

The model in (3.34) and the underlying assumption of $t-$dependence are in agreement with statistical characterizations of the distribution of visits, which have been shown in numerous analyses to follow either a Zipf [36] or a Zipf-Mandelbrot [37] distribution for both commercial and user-generated content. Furthermore, it is confirmed in the real YouTube datasets used in the present work through the evaluation of the time-series's Hurst exponents, as will be discussed in Section 3.7.1.

The off-line analysis tests the constancy (or not) of the mean values up to the current time N. Hence, we define the following null hypothesis of constant mean,

$$H_0 : \quad \mu_1 = \ldots = \mu_N,$$

against the alternative,

$$H_1 : \quad \mu_1 = \ldots = \mu_{k^*_{off}} \neq \mu_{k^*_{off}+1} = \ldots = \mu_N,$$

indicating that the mean value changed at the unknown (time) point $k^*_{off} \in \{1, \ldots, N\}$.

Considering (3.34) and the corresponding assumptions for the stochastic process X_n, we develop a non-parametric CUSUM test statistic following [38]. The test statistic TS_{off}, can be viewed as a max-type procedure,

$$TS_{off} = \max_{1 \leqslant n \leqslant N} C_n^T \widehat{\Omega}_N^{-1} C_n, \tag{3.2}$$

where the parameter C_n is the retrospective CUSUM detector,

$$C_n = \frac{1}{\sqrt{N}} \left(\sum_{i=1}^{n} X_i - n\overline{X}_{1,N} \right), \tag{3.3}$$

while $\overline{X}_{1,N} = \frac{1}{N} \sum_{i=1}^{N} X_i$ denotes the sample mean. $\widehat{\Omega}_N$ represents a suitable estimator of the long-run covariance Ω, where

$$\Omega = \sum_{i=-\infty}^{\infty} \mathbf{Cov} \left(X_n X_{n-i} \right). \tag{3.4}$$

The estimator should satisfy,

$$\widehat{\Omega}_N \xrightarrow{P} \Omega \tag{3.5}$$

where $\xrightarrow{P}$ denotes convergence in probability.

Several estimators have been proposed in the literature that satisfy (3.5), including kernel-based [39], bootstrap-based [40], etc. Considering our requirement for real-time detection (low computational time), a kernel-based estimator is more suitable; in this context, we employ the Bartlett estimator, so that

$$\widehat{\Omega}_N = \widehat{\Sigma}_0 + \sum_{w=1}^{W} k_{BT} \left(\frac{w}{W+1} \right) \left(\widehat{\Sigma}_w + \widehat{\Sigma}_w^T \right), \tag{3.6}$$

which satisfies (3.5), while the function $k_{BT}(.)$ corresponds to the Bartlett weight,

$$k_{BT}(x) = \begin{cases} 1 - |x|, & \text{for } |x| \leqslant 1 \\ 0, & \text{otherwise} \end{cases}, \tag{3.7}$$

and $\widehat{\Sigma}_w$ denotes the empirical auto-covariance matrix for lag w,

$$\widehat{\Sigma}_w = \frac{1}{N} \sum_{n=w+1}^{N} \left(X_n - \overline{X} \right) \left(X_{n-w} - \overline{X} \right)^T. \tag{3.8}$$

Finally, we chose $W = \log_{10}(N)$ as in [39].

The long-run covariance is involved in the test statistic to incorporate the dependence structure of the r.v. into the statistical analysis, through the integration of second order statistical properties. This approach is suitable for the targeted context since we avoid a restrictive assumption for the dependence structure of the observations.

Going back to the basic question of rejecting or not H_0, we need to obtain critical values, denoted by cv_{off}, for the test statistic. We approach this issue by considering the asymptotic distribution of the test statistic under H_0,

$$TS_{off} \xrightarrow{D} cv_{off} = \sup_{0 \leqslant t \leqslant 1} \sum_{j=1}^{r} B_j^2(t) \quad (N \to \infty), \tag{3.9}$$

where $\xrightarrow{D}$ denotes convergence in distribution, $(B_j(t) : t \in [0,1])$, $1 \leqslant j \leqslant r$, are independent standard Brownian bridges $B(t) = W(t) - tW(1)$, and $W(t)$ denotes the standard Brownian motion with mean

0 and variance t. The critical values for several significance levels α can be computed using Monte Carlo simulations that approximate the paths of the Brownian bridge on a fine grid. The last step is to estimate the unknown CP, defined previously as k^*_{off}, under H_1, given by:

$$\hat{k}^*_{off} = \frac{1}{N} \underset{1 \leqslant n \leqslant N}{\operatorname{argmax}} TS_{off}. \tag{3.10}$$

3.4.2 Extended Off-line Approach

The above hypothesis test identifies the existence of at most one CP and does not ensure that the sample remains statistically stationary in either direction of the detection. In particular, by construction (see (3.2)), the off-line test statistic detects the CP with the highest magnitude. Therefore, for the detection of multiple CPs we need to rephrase the hypothesis test H_1, as follows:

$$H_1: \quad \mu_1 = \ldots = \mu_{k_1} \neq \mu_{k_1+1} = \ldots = \mu_{k_2} \neq \ldots \cdots \neq \mu_{k_{\tau-1}+1} = \ldots = \mu_{k_\tau} \neq \mu_{k_\tau+1} = \ldots = \mu_N.$$

A greedy technique to identify multiple CPs is the binary segmentation (BS) algorithm. The standard BS algorithm relies on the general concept of binary segmentation and is an extension of the single CP estimator. First, a single CP is searched for in the time-series. In case of no change, the procedure stops and H_0 is accepted. Otherwise, the detected CP is used to divide the time-series into two segments in which new searches are performed. The procedure is iterated until no more CPs are detected. The BS algorithm is lightweight (computational time $O(N\log N)$), while its conceptual simplicity leads to efficient implementations. On the other hand, it has been shown in the literature [41], [42], that the standard BS algorithm tends to overestimate the number of CPs, as it does not cross-validate them after their detection.

In the extended off-line approach, we propose the modification of the standard BS with a cross-validation step of the estimated CPs. The cross-validation step is similar to that used in the iterative cumulative sum of squares (ICSS) segmentation algorithm [43], which is used to search for CPs on the marginal variance of independent and identically distributed (i.i.d.) r.vs. In the extended off-line algorithm we consider the CPs estimated from the standard BS in pairs and check if H_0 is rejected in the segment delimited by each pair. If H_0 is not rejected in a particular segment, then no change can be detected in it; as a result, all CPs that fall in the respective segment are eliminated. The improvement, in terms of accuracy, is shown through simulation results in Section IV.

3.5 On-line Phase

In this Section, we describe the on-line scheme that includes: (i) two alternative CUSUM-type approaches for the detection of a change in the mean; and (ii) two alternative approaches to estimate the direction of a change.

3.5.1 On-line Analysis

We rewrite equation (1) in the form,

$$X_n = \begin{cases} \mu + Y_n, & n = 1, \ldots, m + k^* - 1 \\ \mu + Y_n + I, & n = m + k^*, \ldots \end{cases} \tag{3.11}$$

where μ, $I \in \mathbb{R}^r$ represents the mean parameters before and after the unknown time of possible change $k^* \in \mathbb{N}^*$ respectively. As a reminder, the first dimension of the time-series represents the video views; the rest could be likes, comments, etc., and $\{Y_n : n \in \mathbb{N}\}$ is a random component. The term $m \in \mathbb{N}$ denotes the length of the training period, i.e., an interval of length m over the historical period during

which the mean is assumed to remain unchanged, so that,

$$\mu_1 = \cdots = \mu_m. \tag{3.12}$$

To satisfy this assumption, the modified off-line CP test previously presented is run in order to identify a suitable m. With m determined, the on-line procedure can be used to check whether (3.12) holds as new data become available. In the form of a statistical hypothesis test, the on-line problem becomes,

$$\begin{aligned} H_0 &: I = 0, \\ H_1 &: I \neq 0. \end{aligned} \tag{3.13}$$

The on-line sequential analysis belongs to the category of stopping time stochastic processes. In general, a chosen on-line test statistic $TS_{on}(m,l)$ and a given threshold $F(m,l)$ define the stopping time $\tau(m)$:

$$\tau(m) = \begin{cases} \min\{l \in \mathbb{N} : TS_{on}(m,l) \geqslant F(m,l)\}, \\ \infty, \text{ if } TS_{on}(m,l) < F(m,l) \ \forall l \in \mathbb{N}, \end{cases} \tag{3.14}$$

implying that $TS_{on}(m,l)$ is calculated on-line for every l in the monitoring period. The procedure stops if the test statistic exceeds the value of the threshold function $F(m,l)$. As soon as this happens, the null hypothesis is rejected and a CP is detected. The following properties should hold for $\tau(m)$,

$$\lim_{m\to\infty} Pr\{\tau(m) < \infty | H_0\} = \alpha,$$

ensuring that the probability of false alarm is asymptotically bounded by $\alpha \in (0,1)$, and,

$$\lim_{m\to\infty} Pr\{\tau(m) < \infty | H_1\} = 1,$$

ensuring that under H_1 the asymptotic power of the statistical test is unity. The threshold $F(m,l)$ is given by,

$$F(m,l) = cv_{on,a} g(m,l), \tag{3.15}$$

where: (i) the critical value $cv_{on,a}$ is determined from the asymptotic behavior of the stopping time procedure under H_0 by letting $m \to \infty$; and (ii) the weight function,

$$g(m,l) = \sqrt{m}\left(1 + \frac{l}{m}\right)\left(\frac{l}{l+m}\right)^{\gamma} \tag{3.16}$$

depends on the sensitivity parameter $\gamma \in [0, 1/2)$.

We use two different CUSUM approaches; the standard [11], with test statistic denoted by TS_{on}^{ct}, and, the ratio-type [12], with test statistic denoted by TS_{on}^{rt}. Their corresponding critical values are denoted by $cv_{on,a}^{ct}$ and $cv_{on,a}^{rt}$, respectively, and their stopping rules by $\tau_{ct}(m)$ and $\tau_{rt}(m)$, correspondingly. Both tests are based on the sequential CUSUM detector, $E(m,l)$,

$$E(m,l) = \left(\overline{X}_{m+1,m+l} - \overline{X}_{1,m}\right) \tag{3.17}$$

The standard CUSUM test is expressed as:

$$TS_{on}^{ct}(m,l) = l\widehat{\Omega}_m^{-\frac{1}{2}} E(m,l), \tag{3.18}$$

where $\widehat{\Omega}_m$ is the estimated long-run covariance, defined as in (4), that captures the dependence between observations. Then, the stopping rule $\tau_{ct}(m)$, is defined as:

$$\tau_{ct}(m) = \min\{l \in \mathbb{N} : \|TS_{on}^{ct}(m,l)\|_1 \geq cv_{on,a}^{ct} g(m,l)\}, \tag{3.19}$$

where the ℓ_1 norm is involved to modify TS_{on}^{ct} so that it can be compared to a one dimensional threshold function. The critical value, $cv_{on,a}^{ct}$, is derived from the asymptotic behavior of the stopping rule under H_0:

$$\lim_{m\to\infty} Pr\{\tau(m) < \infty\} = \lim_{m\to\infty} Pr\left\{ \sup_{1\leqslant l\leqslant\infty} \frac{\|TS_{on}^{ct}(m,l)\|_1}{g(m,l)} > cv_{on,\alpha}^{ct} \right\}$$
$$= Pr\left\{ \sup_{t\in[0,1]} \frac{\|W(t)\|_1}{t^\gamma} > cv_{on,\alpha}^{ct} \right\} = \alpha. \tag{3.20}$$

Unlike standard CUSUM tests, ratio type statistics do not require to estimate the long-run covariance and are also considered for this reason in this analysis. The precise form of the chosen statistic is given in the following quadratic form,

$$TS_{on}^{rt}(m,l) = \frac{l^2}{m} E^T(m,l) \left\{ \frac{1}{m^2} \sum_{j=1}^{m} j^2 \left(\overline{X}_{1,j} - \overline{X}_{1,m}\right)\left(\overline{X}_{1,j} - \overline{X}_{1,m}\right)^T \right\}^{-1} E(m,l), \tag{3.21}$$

with its equivalent stopping rule,

$$\tau_{rt}(m) = \min\{l \in \mathbb{N} : TS_{on}^{rt} \geq cv_{on,a}^{rt} g^2(m,l)\}. \tag{3.22}$$

Similarly to the standard CUSUM, the critical value, $cv_{on,a}^{rt}$, is estimated by,

$$\lim_{m\to\infty} Pr\{\tau(m) < \infty\} = Pr\left\{ \sup_{t\in[0,\infty)} \Delta_\gamma(t) > cv_{on,\alpha}^{rt} \right\} = \alpha, \tag{3.23}$$

where,

$$\Delta_\gamma(t) = \frac{1}{\eta_\gamma^2(t)} B^T(1+t) \left(\int_0^1 B(r) B^T(r) dr \right)^{-1} B(1+t), \eta_\gamma^2(t) = (1+t)\left(\frac{t}{1+t}\right)^\gamma,$$

and $B(t)$ is a standard Brownian bridge, $t \in [0,\infty)$.

Similarly to the off-line case, the on-line critical values for both test statistics can be computed using Monte Carlo simulations, considering that,

$$cv_{on,\alpha}^{ct} = \sup_{t\in[0,1]} \frac{W(t)}{t^\gamma}, \tag{3.24}$$

$$cv_{on,\alpha}^{rt} = \sup_{t\in[0,\infty)} \Delta_\gamma(t). \tag{3.25}$$

The estimated on-line CP, $\hat{k}_{on}^*$, is derived directly from the value of the stopping time $\tau(m)$, as,

$$\hat{k}_{on}^* = m + \{\tau(m) | \tau(m) < \infty\}. \tag{3.26}$$

3.5.2 Trend Indicator

Considering the on-line procedure, the hypothesis H_1 is two-tailed because the test statistics TS_{on}^{rt} and TS_{on}^{ct} are formulated in a quadratic form and a ℓ_1 norm, respectively. This means that the stopping time rule $\tau_{ct}(m)$ (or $\tau_{rt}(m)$) cannot be an indicator of the direction of a detected change. Thus, to estimate the direction of a change we introduce two indicators: i) based on the CUSUM detector in (3.17), denoted by TI_{ts}; and ii) based on the moving average convergence divergence (MACD) filter [44], denoted by TI_f.

Focusing on TI_{ts}, the indicator is directly derived from the form of the sequential CUSUM detector $E(m,l)$. The detector compares the mean value of the observations that are collected on-line for a chosen monitoring period l, with the mean value of a subsample of the historical data over the predetermined training sample. Hence, for a detected CP, we have that,

$$\begin{cases} E(m,l) > 0, \text{ denotes an upward change} \\ E(m,l) < 0, \text{ denotes a downward change} \end{cases} . \tag{3.27}$$

However, in certain cases, limiting the window over which the direction of a change is estimated to the immediate neighbourhood of a detected CP can be unreliable due to the continuous variability of the time-series. In such cases, we have to estimate the direction of a change by incorporating more elaborate filters; in this context, we estimate the direction of detected changes by applying the MACD indicator. The MACD is based on an exponential moving average (EMA) filter, of the form,

$$EMA_p(n) = \frac{2}{p+1} X_n + \frac{p-1}{p+1} EMA_p(n-1), \tag{3.28}$$

with p denoting the lag parameter. The MACD series can be derived from the subtraction from a short p_2 lag EMA (sensitive filter) of a longer p_3 lag EMA (blunt filter), as described below:

$$MACD(n) = EMA_{p_2} - EMA_{p_3}. \tag{3.29}$$

The trend indicator TI_f is then obtained by the subtraction of a short p_1 lag EMA filter of a MACD series from the raw MACD series, as described below

$$TI_f(n) = MACD(n) - EMA_{p_1}(MACD(n)),\ p_1 < p_2 < p_3. \tag{3.30}$$

In the evaluation of TI_f three exponential filters are involved. In essence, TI_f is an estimation of the second derivative over an interval around the change (considering that the subtraction of a filtered variable from the variable generates an estimate of its time derivative). In contrast to other works [44], we only adopt TI_f to characterize the direction from the specific value of TI_f at the estimated time of change. We announce an upward change if $TI_f(\hat{k}^*_{on}) > 0$, otherwise, if $TI_f(\hat{k}^*_{on}) < 0$, a downward change.

Finally, we propose a modification of the trend indicator TI_f, converting it from a point estimator to an interval estimator; instead of evaluating $TI_f(\hat{k}^*_{on})$, we propose to evaluate the trend indicator at a time interval $(\hat{k}^*_{on}, \hat{k}^*_{on} + h)$, where h is a threshold parameter:

$$TI_f(\hat{k}^*_{on}, h) = \sum_{l=\hat{k}^*_{on}}^{\hat{k}^*_{on}+h} TI_f(l). \tag{3.31}$$

The proposed $TI_f(\hat{k}^*_{on}, h)$ modification improves the estimator's accuracy; the calculation of the sum of a multitude of observations, after a CP, can smooth out a potential false one-point estimation, especially in the case of small changes.

3.5.3 Overall Algorithm

We outline in *Algorithm 1* the RCPD algorithm, as a combination of the off-line and the on-line phase, in the form of pseudo-code. Beginning from the initial value set for the monitoring starting period, denoted by m_s, the modified off-line algorithm is applied over the whole historical period; the training period m is then defined as the interval elapsed from the last detected off-line CP (if one exists) to m_s. As a second step, the on-line test statistic, $TS_{on}(m,l)$ in (14), is applied for a specified monitoring time frame l. If a content popularity change is detected at time instance $\hat{k}^*_{on}$, the trend indicator subroutine

Algorithm 1: The Real-time CP Detector (RCPD)

procedure RCPD(X_n,m_s,k)
; X_n: time-series of video views
; m_s: running end of training period
; m: training period
; l: monitoring time frame
; d: period assuming no change
; TS_{on}: on-line test statistic (eq. 3.18 or 3.21)
; cv_{on}: critical value (eq. 3.24 or 3.25)
; $\hat{k}^*_{on}$: the estimated on-line CP (eq. 3.26)
; TI: trend indicator (TI_{ts} or TI_f)
for n in X_n **do**
if $n = m_s$ **then**
s=MBS(1,m_s,1) ; calculate off-line CPs
if `array_length`(s) > 0 **then**
m=$\{\max(s), m_s\}$; $max(s)$ is the latest CP
else
m=$\{max(1, m_s - u), m_s\}$; u a large value
end if
else if $m_s < n < m_s + l$ **then**
calculate $\texttt{TS}_{on}$`(m,l)`
if $\texttt{TS}_{on}$`(m,l)`$>\text{cv}_{on}$ **then**
calculate `TI`
signal CP and estimated direction
$m_s = \widehat{cp}_{on} + d$; keep a distance from $\widehat{cp}_{on}$
end if
else if $n = m_s + l$ **then**
$m_s = m_s + l$; start a new training period
end if
end for
end procedure

is called to reveal the direction of change.[3] At this point the procedure stops and a new starting point for the monitoring window is defined as $m_s = \hat{k}^*_{on} + d$, where d is a constant value specifying a period assuming no change. Otherwise, if no change is detected after a maximum of l instances, the procedure restarts from the last time point, $m_s = m_s + l$.

3.6 Validation of the RCPD Using Synthetic Data

In this Section, we validate the performance of the overall algorithm by performing a series of four different experiments on synthetic data. The use of synthetic data allows us to regulate the parameters of the time-series in terms of mean changes and thus obtain quantitative metrics for the performance of the proposed algorithms.

The choice of the time-series model for the generation of the synthetic data is based on the fact that several studies have shown that ARMA models capture very well content popularity evolution. For example, in [9] it has been concluded that an ARMA model can efficiently describe the daily access patterns of YouTube content, based on an extensive analysis of 100,000 videos. Similarly, in [45] an

[3] In the load balancing scenario discussed in Section VII, in the case of an increase in the content popularity a new content cache is being deployed, while conversely a decrease leads to the removal of an existing cache.

Table 3.1: Percentage of the successful CP detections for the standard and modified BS algorithm

	Test 1: two CPs		Test 2: four CPs	
μ	BS	modified BS	BS	modified BS
	True (false) alarm rate		True (false) alarm rate	
μ_1=1	0.94 (0.06)	0.95 (0.05)	0.5 (0.258)	0.7 (0.05)
μ_2=1.5	0.95 (0.05)	0.95 (0.05)	0.5 (0.258)	0.9 (0.08)
μ_3=2	0.95 (0.05)	0.95 (0.05)	0.47 (0.53)	0.9 (0.1)

Table 3.2: Success rates of trend indicators

	Test 1: two CPs		Test 2: four CPs	
μ	TI_{ts}	TI_f	TI_{ts}	TI_f
	Success rate		Success rate	
μ_1=1	0.99	0.99	0.99	0.99
μ_2=1.5	1	1	1	1
μ_3=2	1	1	1	1

ARMA model has been proposed for the estimation of the popularity of video content. Motivated by these findings, for the validation of the proposed algorithm we use an ARMA$(1,1)$ time-series. We generate $1,000$ time-series of length $N = 600$ samples. Without loss of generality, we assume an initial mean value $\mu_0 = 0$, noting that the performance of the RCPD is independent of the initial mean value and only depends on the magnitude of the variation of the mean value before and after a CP.

In the first experiment, we begin with a comparison of the standard BS to the proposed modified BS algorithms described in Section 3.4. We perform two tests; in the first test we introduce two CPs at the instances $k_i^* = (iN)/3,\ i = 1, 2$, while in second test, we introduce four CPs at $k_i^* = (iN)/5,\ i = 1, \ldots, 4$. The two tests are repeated for three different values of the magnitude of a change $\mu_1 = 1,\ \mu_2 = 1.5, \mu_3 = 2$, i.e., we randomly increase or decrease the mean value by $\mu_j,\ j = 1, \ldots, 3$ at the time of change. Table 3.1 summarizes our findings regarding the true and false alarm rates of the two algorithms.

Both the standard and the modified BS algorithms provide similar true alarm rates, exceeding 94%, in the first test. On the contrary, in the more challenging second test, the superiority of the modified BS over the standard BS algorithm is clear. The modified BS algorithm achieves true alarm rates in excess of 70%, even in the demanding scenario of a relatively small change in the mean $\mu_1 = 1$. On the other hand, the standard BS algorithm has in all cases a true alarm rate of less than 50%, rendering any CP detection highly questionable. The second test confirms that the standard BS algorithm is prone to an overestimation of the number of CPs as shown by the high false alarm rates (in excess of 25% in all cases), an issue that can be effectively addressed by the modified BS algorithm which scores false alarm rates below 10%.

Next, in the second experiment, using the same test sets as above, we measure the success rates achieved by the proposed trend indicators TI_{ts} and TI_f for $h = 0$ (larger thresholds provided the same true identification rates). The results are summarized in Table 3.2. The two trend indicators successfully identify the direction of a change in more than 99% of the cases, which shows that they can be interchangeably employed. In the assessment of the performance using real datasets in Sections 3.6 and 3.7, we solely employ the TI_f trend indicator.

Table 3.3: Results of the RCPDs' algorithm CPs detection for one change in the mean value.

		ARMA(1,1)							
μ	l	standard CUSUM				ratio-type CUSUM			
		Number of detected CPs			$\hat{k}^*$	Number of detected CPs			$\hat{k}^*$
		0	1	> 1	med	0	1	> 1	med
	25	0.95	0.05	0	-	0.95	0.05	0	-
$\mu = 0$	50	0.95	0.05	0	-	0.95	0.05	0	-
	100	0.94	0.06	0	-	0.95	0.05	0	-
	25	0.7	0.29	0.01	-	0.8	0.19	0.01	-
$\mu = 0.5$	50	0.16	0.8	0.04	343	0.55	0.43	0.02	-
	100	0	0.93	0.07	341	0.2	0.76	0.04	348
	25	0.26	0.73	0.01	332	0.69	0.3	0.01	-
$\mu = 0.7$	50	0	0.96	0.04	326	0.3	0.65	0.05	328
	100	0.01	0.91	0.08	331	0.05	0.89	0.06	335
	25	0.01	0.97	0.02	327	0.52	0.46	0.02	-
$\mu = 1$	50	0	0.96	0.04	316	0.08	0.86	0.06	321
	100	0	0.92	0.08	321	0	0.95	0.05	323
	25	0.01	0.97	0.02	323	0.43	0.54	0.03	331
$\mu = 1.2$	50	0	0.95	0.05	316	0.02	0.93	0.05	317
	100	0	0.93	0.07	318	0	0.93	0.07	318
	25	0	0.97	0.03	320	0.36	0.6	0.04	329
$\mu = 1.5$	50	0	0.95	0.05	310	0	0.94	0.06	313
	100	0	0.93	0.07	314	0	0.94	0.06	318
	25	0	0.97	0.03	310	0.26	0.71	0.03	317
$\mu = 2$	50	0	0.95	0.05	307	0	0.93	0.07	310
	100	0	0.94	0.06	310	0	0.94	0.06	313

We proceed by assessing the proposed RCPD algorithm using both the standard and the ratio type CUSUM. In this third experiment, we measure the average number of CPs detected, averaged over $1,000$ simulations when a single CP is introduced in the ARMA time-series at the time instance $\frac{N}{2} = 300$. We consider different values for the magnitude of change $\mu \in \{0, 0.5, 0.7, 1, 1.2, 1.5, 2\}$ and the monitoring window length $l \in \{25, 50, 100\}$. We note that we included the case $\mu = 0$ – which corresponds to the absence of a change – to evaluate the false alarm rate of the overall algorithm. We omit results with true alarm rates lower than 50% as they are statistically unreliable. In terms of the remaining algorithmic parameters, we have set the minimum distance between two successive CPs to $d = 50$,[4] the sensitivity parameter to $\gamma = 0.25$ [46] (we choose a neutral value as the behaviour of γ is well studied), and, the significance level to $\alpha = 0.05$. In each test of the third experiment we measure the exact number of CPs detected, tabulated as one the following three values: i) 0 when (falsely[5]) no

[4]This choice is justified by our observations of the minimum distance between successive CPs in real data sets, presented in Section VI.

[5]Except for the $\mu = 0$ case.

Table 3.4: Results of the RCPDs' algorithm CPs detection for two mean changes.

		ARMA(1,1)									
μ	l	standard CUSUM					ratio-type CUSUM				
		Number of detected CPs			$\hat{k}_1^*$	$\hat{k}_2^*$	Number of detected CPs			$\hat{k}_1^*$	$\hat{k}_2^*$
		< 2	2	> 2	med		< 2	2	> 2	med	
$\mu_1 = 0.5$	25	0.88	0.12	0	-	-	0.95	0.05	0	-	-
	50	0.38	0.60	0.02	251	440	0.79	0.2	0.01	-	-
	100	0.1	0.87	0.03	242	443	0.54	0.44	0.02	-	-
$\mu_1 = 0.7$	25	0.41	0.58	0.01	230	427	0.9	0.1	0	-	-
	50	0.06	0.91	0.03	223	427	0.58	0.41	0.01	-	-
	100	0.01	0.93	0.06	227	428	0.25	0.72	0.03	231	439
$\mu_1 = 1$	25	0.04	0.93	0.03	219	420	0.74	0.25	0.01	-	-
	50	0.03	0.93	0.04	215	419	0.26	0.71	0.03	221	423
	100	0	0.94	0.06	217	420	0.05	0.9	0.05	220	424
$\mu_1 = 1.2$	25	0.01	0.96	0.03	214	414	0.56	0.42	0.02	-	-
	50	0	0.95	0.05	212	416	0.17	0.79	0.04	215	428
	100	0	0.94	0.06	217	420	0.02	0.93	0.05	216	421
$\mu_1 = 1.5$	25	0	0.98	0.02	211	411	0.33	0.63	0.04	213	417
	50	0	0.94	0.06	209	413	0.1	0.85	0.05	213	415
	100	0	0.94	0.06	211	415	0	0.96	0.04	216	419
$\mu_1 = 2$	25	0	0.98	0.02	208	407	0.12	0.85	0.03	210	412
	50	0	0.95	0.05	207	410	0.3	0.91	0.06	209	413
	100	0	0.94	0.06	209	411	0	0.96	0.04	211	414

CP is detected; ii) 1 when (correctly) a single CP is detected; and iii) > 1 when (falsely) multiple CPs are detected. Finally, we measure the median of the time instance of the single CP detection, denoted by $\hat{k}^*$.[6] The results of this experiment are presented in Table 3.3 and are discussed below.

Firstly, we observe that both the standard and the ratio type CUSUM achieve very small false alarm rates, inferior to 6% when no CP is inserted, irrespective of the choice of l. On the contrary, the choice of l readily affects the algorithm's success rate for $\mu > 0$; for small changes in the mean value, $\mu = 0.5, 0.7$, a larger monitoring window l increases the algorithm's true alarm rates in identifying correctly the existence of the CP. For medium and high changes in the magnitude of change $\mu = 1, 1.2, 1.5, 2$, it is observed that a high true alarm rate – in excess of 93% for the standard CUSUM – is achieved, while choosing a smaller l can slightly increase the true alarm rates. As a result, depending on the application, a choice of a larger l can be appropriate if the algorithm is to be employed as a universal CP detector. Alternatively, a smaller l can be chosen when the focus is on the identification of large changes in the mean value, i.e., we are interested primarily in detecting CPs of larger magnitude.

Secondly, we observe that overall, the ratio type CUSUM is outperformed by the standard CUSUM in all tests. Consequently, the standard CUSUM based detector can be considered as an efficient universal choice. Finally, we observe that the lag between $\hat{k}^*$ and the actual instance of change at

[6] We omit the results with true detection rate lower than 50%.

the point 300 decreases with increasing μ, ranging from 343 to 307, while it appears less sensitive to changes in l. This demonstrates that, intuitively, larger magnitude changes can be detected faster. This result is important for load balancing applications as it provides us with the means to quickly respond to significant changes in the network traffic.

Subsequently, in Table 3.4 in the previous page, we present the outputs of the fourth experiment in which we assess the performance, averaged over $1,000$ simulations, of the RCPD algorithm when two CPs are inserted in the ARMA time-series. We introduce a change at the time instance $k_1^* = \frac{N}{3} = 200$ and a second CP at the time instance $k_2^* = \frac{2N}{3} = 400$. We investigate the true and false alarm rates for $\mu \in \{0.5, 0.7, 1, 1.2, 1.5, 2\}$ and $l \in \{25, 50, 100\}$, while the rest of the parameters retain the values of the third experiment. In each test of the fourth experiment we measure the exact number of CPs detected, tabulated as one the following three values: i) < 2 when (falsely) less than two CPs are detected, ii) 2 when (correctly) two CPs are detected, and iii) > 2 when (falsely) more than two CPs are detected. Finally, we measure the median of the detection instances of the two CPs, denoted by $\hat{k}_1^*$ and $\hat{k}_2^*$, respectively (we omit the results with true detection rate lower than 50%).

Similarly to the third experiment, we observe that increasing l increases the true alarm rates for small magnitudes in the mean changes $\mu = 0.5,\ 0.7$, while this trend is reversed in high magnitudes $\mu = 1.5,\ 2$. For medium values $\mu = 1,\ 1.2$ the effect of l on the true alarm rates is less than 2%. Furthermore, in agreement with the outputs of the third experiment, with increasing μ the algorithms achieve increasingly high success rates, over 93% for the standard CUSUM when $\mu \geq 1$.

In addition, the superior performance of the standard CUSUM is re-confirmed in all the tests of the fourth experiment. Finally, with respect to the lag in the estimation of the time instances of the CPs, we observe that, as in experiment three, larger magnitude changes can be detected faster, e.g., for $\mu = 2$ a lag inferior to 11 instances is observed for both CPs with the standard CUSUM, irrespective of l.

Concluding this Section, we have presented an extensive set of experiments that provide strong evidence for the efficiency of the proposed algorithms. We have explicitly demonstrated the superiority of the modified BS over the standard BS algorithm and confirmed the validity of the proposed trend indicators. Subsequently, we evaluated the performance of the overall algorithm for various values of μ and l. We have shown that the RCPD algorithm achieves extremely high true alarm rates for larger values of μ, while increasing the length of the monitoring window l can significantly impact the performance for small values of μ. Finally, overall, the standard type CUSUM outperforms the ratio type CUSUM and should be preferred.

3.7 Performance Evaluation Using Real Data

In this Section we investigate the performance of the proposed algorithms using a real dataset provided within the framework of the CONGAS project [47]; the dataset consists of the number of views of 882 YouTube videos, observed over $N = 1,000$ instances.

3.7.1 Statistical Properties of the Real Dataset

First, we evaluate the validity of the most important underlying assumption of this analysis, that the content popularity can be modelled as the sum of a constant mean and a weak-dependent (t-dependent) stochastic process, as given in (3.34). A first intuitive method to test whether the time-series is short-range dependent (SRD) is through its autocorrelation function (ACF). The ACF for a weakly-stationary process $\{X_t : t \in \mathbb{N}$ with mean value μ is given by,

$$\rho(k) = \frac{(X_t - \mu)(X_{t+k} - \mu)}{\sigma^2}.$$

Note that if $\sum_{k=-\infty}^{\infty} \rho(k) \to \infty$ the process has long-range dependence (LRD), while if $\sum_{k=-\infty}^{\infty} |\rho(k)| < \infty$ it exhibits SRD. To distinguish between these two phenomena, we use the following functional form

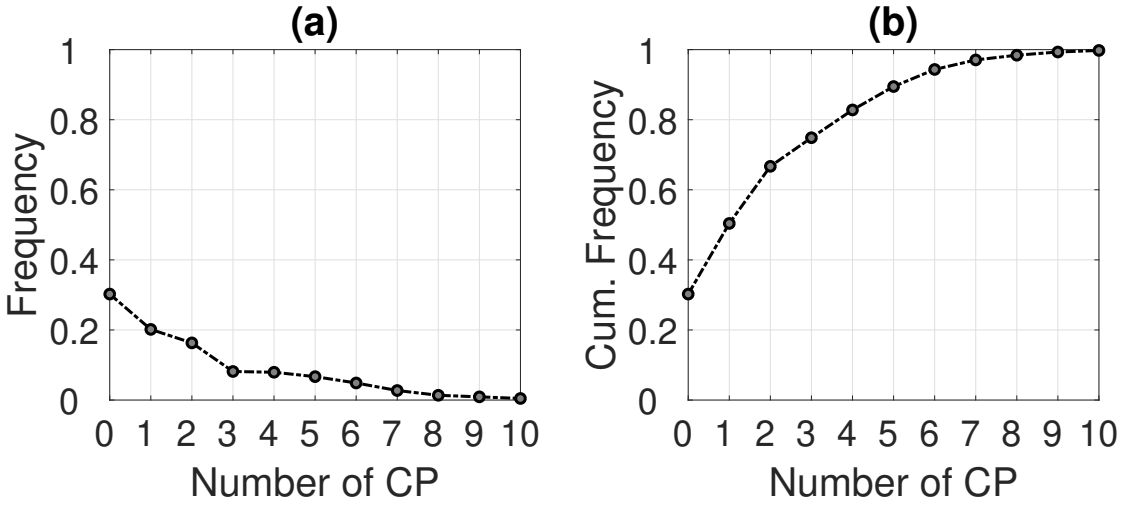


Figure 3.1: Estimated a) frequency and b) cumulative frequency of the number of CPs per time-series.

Table 3.5: Success rates of TI_f trend indicator

h	0	3	5	7	10
Video Set 1	0.69	0.91	0.95	0.97	0.98
Video Set 2	0.90	0.99	0.99	0.99	0.99

of the ACF,

$$\rho(k) \sim C_i^{2H-2}, \text{ as } i \to \infty,$$

where $C_i > 0$ and $H \in (0, 1)$ is the Hurst exponent characterizing the LRD,i.e., $H \in (1/2, 1)$ indicates the presence of LRD. It is challenging to accurately estimate the Hurst exponent out of real data [48] and several methods have been proposed in the literature [49]. In this work, we apply two semi-parametric tests, identified as accurate options among others presented in the survey paper [49]. The first method uses the discrete second order derivative in the time domain while the second uses the discrete second order derivative in the wavelet domain. Both methods estimate an $H \leq 0.5$ for 95% of the YouTube time-series, indicating the validity of our assumptions related to the equation (3.34).

3.7.2 Performance of the Off-line Training Phase

First, we test the hypothesis H_0 of no change in the mean structure on our dataset. H_0 is rejected in approximately 70% of the cases, for a significance level of $a = 0.05$. This outcome indicates that CP algorithms can identify changing content dynamics in real times series. Next, we estimate the number of CPs, by applying the extended off-line algorithm. The corresponding results are illustrated in Fig. 3.1 and indicate a sufficiently high number of content popularity anomalies (i.e., mean changes). Hence, a CP analysis is indeed a suitable tool for content popularity detection.

To evaluate the performance of the proposed trend indicator TI_f, we need a baseline independent assessment of the direction of change. We declare that a real increase in the mean value of content visit exists if

$$\mathsf{E}[X(\hat{k}^*_{i-1,off}) : X(\hat{k}^*_{i,off})] < \mathsf{E}[X(\hat{k}^*_{i,off}) : X(\hat{k}^*_{i+1,off})], \tag{3.32}$$

or, that a real decrease in the number of visits exists if

$$\mathsf{E}[X(\hat{k}^*_{i-1,off}) : X(\hat{k}^*_{i,off})] > \mathsf{E}[X(\hat{k}^*_{i,off}) : X(\hat{k}^*_{i+1,off})], \tag{3.33}$$

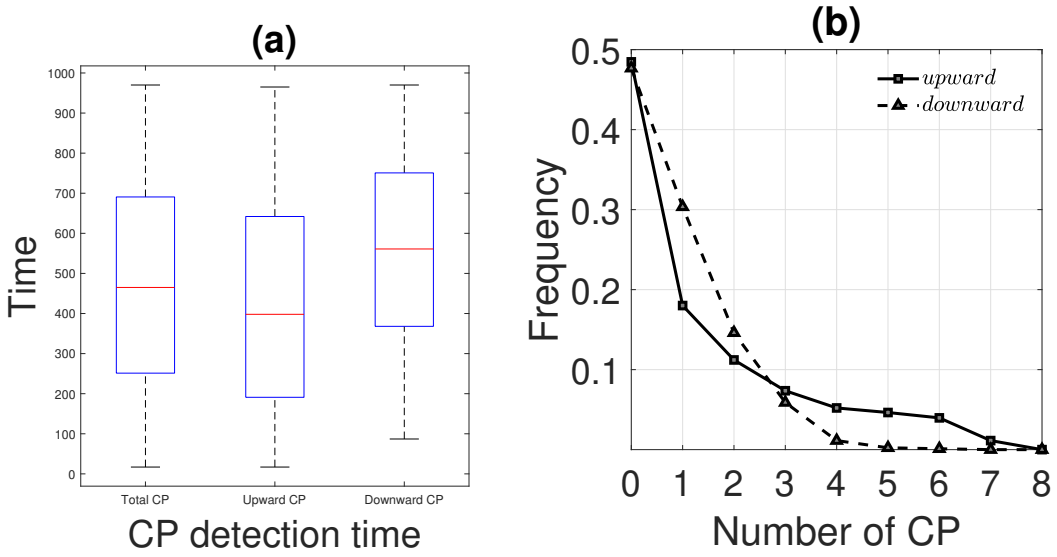


Figure 3.2: Frequency values of the number of upward and downward CPs, per time-series.

where $i = 2, \ldots, N-1$ and $\mathsf{E}[\cdot]$ denotes the numerical average. We test the modified MACD TI_f on two sets of videos. The first set, Video Set 1, comprises the whole dataset, while the second set, Video Set 2, comprises only the videos with a considerable average number of visits (> 10), i.e., for which, $\mathsf{E}[X(1) : X(1000)] > 10$.

The percentage of successful TI_f identifications are tabulated in Table 3.5 for five values of the parameter h, namely $h = 0$, 3, 5, 7 and 10, where h denotes the TI_f's calculation threshold. Commenting on the results for Video Set 1, the TI_f trend indicator works well, except for $h = 0$, providing at least 90% correct direction identifications. As expected, as h increases the procedure works better. More specifically, an $h \geq 5$ parameter choice yields a success rate of 95%, while if a more agile estimation is needed then an $h \geq 3$ still maintains a 91% accuracy. Considering the interim time between consecutive changes, we deduce that an $h \leq 7$ is preferable. Regarding Video Set 2, we see that the results are highly improved, indicating that the procedure works even better for the most popular videos. In practice, this represents the more interesting scenario as it will have a greater impact in terms of the applied load balancing mechanism.

Furthermore, in Fig. 3.2, the time instances of upward and downward changes are shown in the form of a boxplot. It is intuitive that upward changes occur earlier than downward changes. Moreover, Fig. 3.2 demonstrates that the multitude of upward changes is greater than the respective of downward changes, indicating that decreases in popularity are sharper than increases. In particular, we estimated that out of the total number of changes, 67% are upward.

Finally, we analyze the interim time between consecutive CPs. The results presented in Fig. 3.3 illustrate the existence of a sufficiently large gap between consecutive potential changes. 90% of the intervals corresponding to consecutive CPs exceed 70 time instances and only 5% of them are shorter than 50 time instances, ensuring that a sufficiently large training window can be applied. The results depicted in Fig. 3.3 allow adjusting parameters of the on-line phase, in particular the minimum time interval between consecutive changes, denoted by the parameter d.

3.7.3 Evaluation of the RCPD Algorithm

In the previous subsection we have evaluated the performance of the off-line algorithm and demonstrated its efficiency as well as how it is employed in determining parameters of the on-line phase, such as the interval assuming no change d and the threshold parameter of TI_f h.

We further employ the off-line algorithm as a benchmark against which the performance of the

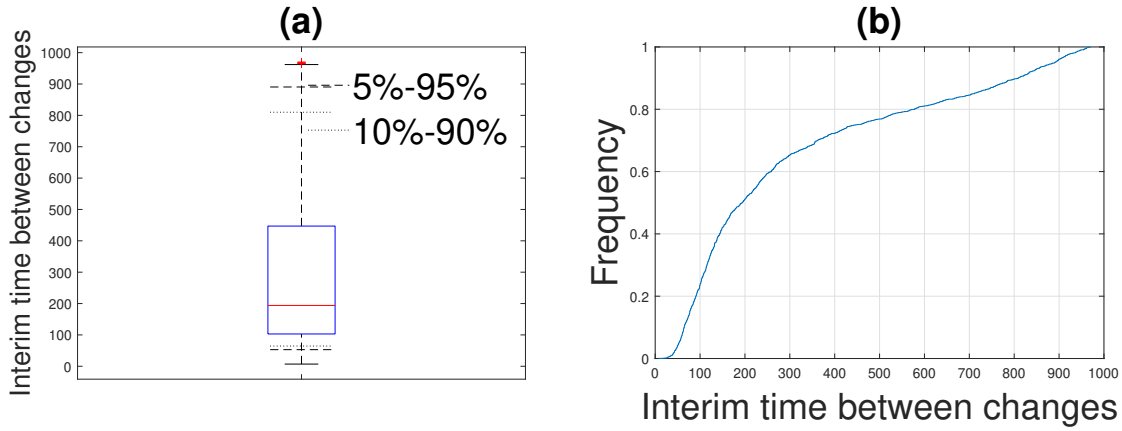


Figure 3.3: a) Boxplot including the interval $(5\% - 95\%)$ (dashed line) and $(10\% - 90\%)$ interval (dotted line), b) Cumulative frequency for the interim time of consecutive CPs.

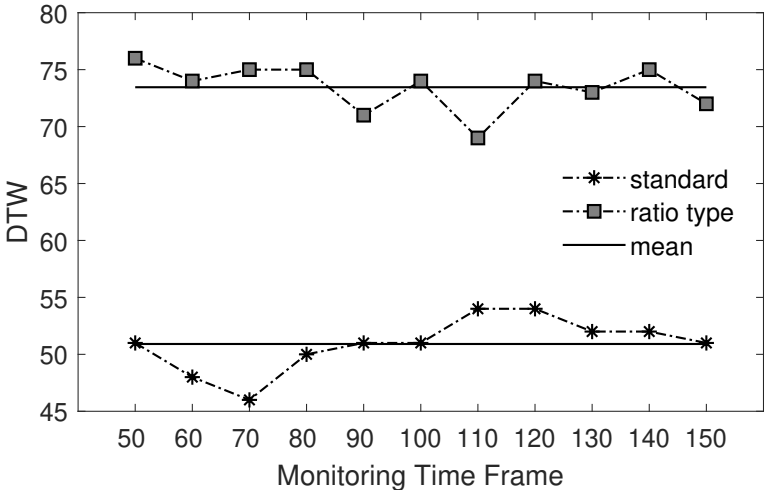


Figure 3.4: DTW distances for the two on-line detection schemes.

RCPD algorithm will be evaluated. We note that the off-line analysis provides the *best possible statistical detection* of the actual mean changes, as off-line algorithms operate retrospectively over the entirety of each of the time-series. Thus, in absence of a priori knowledge of the actual CPs in the real data (as opposed to the synthetic data in which the CPs were controlled), we evaluate the performance of the RCPD procedure by measuring the "similarity" of its outputs (detected CPs, instances of detection and trends) to the corresponding outputs of the off-line version.

As the number of detected CPs and / or their exact positions are likely to differ at the output of the retrospective (off-line) and of the RCPD algorithm, in order to obtain a measure of their similarity, we estimate their dynamic time warping (DTW) distance. The DTW is a dynamic programming tool that measures distances between asynchronous sequences and is widely used by the speech processing community [14].

The results are presented in Fig. 3.4, where the estimated DTW distances are depicted for several values of the monitoring window length $l \in [40, 150]$, to investigate the consistency of parameter l over different values. In the RCPD algorithm we use $d = 50$ (minimum distance between two changes) and have set the sensitivity parameter to $\gamma = 0.25$. The estimated mean DTW distance for the standard CUSUM is 52 and for the ratio-type CUSUM is 73. For comparison purposes, we note that the corresponding DTW distance over the synthetic data is 20 for medium / large changes, while the true CP detections are around 95%. As a result, we can infer, that the outputs of the on-line algorithm,

Table 3.6: Empirical percentiles of mean values change rate.

	Percentiles Threshold			
	10%	15%	25%	50%
Standard	9%	13.1%	20.8%	42.21%
Ratio type	9.5%	14.82%	28.22%	67.40%

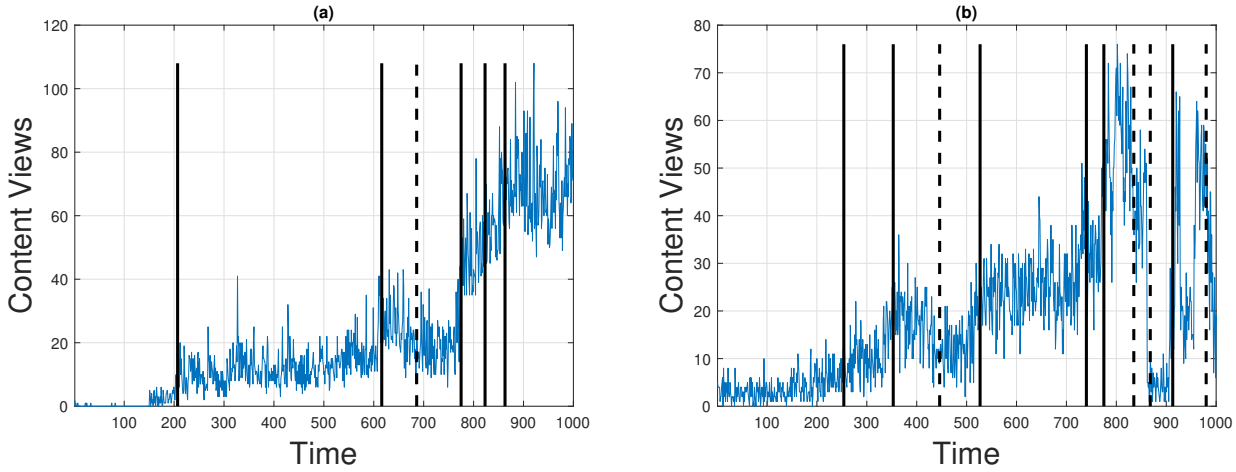


Figure 3.5: Outputs of the RCPD algorithm using standard CUSUM for different time-series. Solid and dashed lines depict an upward and a downward change, respectively.

using the standard CUSUM, are "very close" to the outputs of the benchmark off-line algorithm. In agreement with our observations over the synthetic data, the DTW distance using the ratio-type CUSUM is clearly larger.

We also study the magnitude of the detected CPs. We define as the CP magnitude the percentage-wise change in the mean values before and after the CP. We group the measured magnitudes for all change points using the four percentile threshold values 10%, 15%, 25% and 50%, i.e., reflecting the frequency of magnitudes exceeding the respective thresholds. The results are summarized in Table 3.6. According to our results, both the standard and ratio type CUSUM algorithms detect the most significant changes in the content popularity. Moreover, ratio-type CUSUM detects, in general, CPs with the largest magnitude of change, in agreement with synthetic data results.

Additionally, for illustration purposes, we depict the RCPD algorithm's outputs for four different time-series. We set the beginning of the monitoring period at $m_s = 200$ and monitoring horizon $l = 50$, the on-line parameter $g = 0.25$ and the significance level to $a = 0.05$. The corresponding results are depicted in Fig. 3.5 and 3.6, showing the estimated CPs by applying the standard CUSUM and the ratio type CUSUM procedures, respectively. In both cases, the estimated changes correspond to the real content popularity changes; visual inspection suggests that the performance of the standard CUSUM is more reasonable (e.g., Fig. 3.6d). The RCPD, as it is illustrated in Fig. 3.5b seems to be adaptable to "fast" changes; without getting "confused" by random peaks in the time-series, such as those in Fig. 3.5a or in Fig. 3.6c.

3.7.4 Time Dependencies of Piecewise time-series

We also measure the autocorrelation function of the piecewise - divided by the detected CPs - time-series. Results are tabulated in Table 3.7 and verify the short dependence structure of the dataset; significant lags in time dependencies higher than 30 instances can be found in less than 5% of the time-series. Furthermore, the fact that the ACF of the piecewise time-series drops to zero quickly indicates that

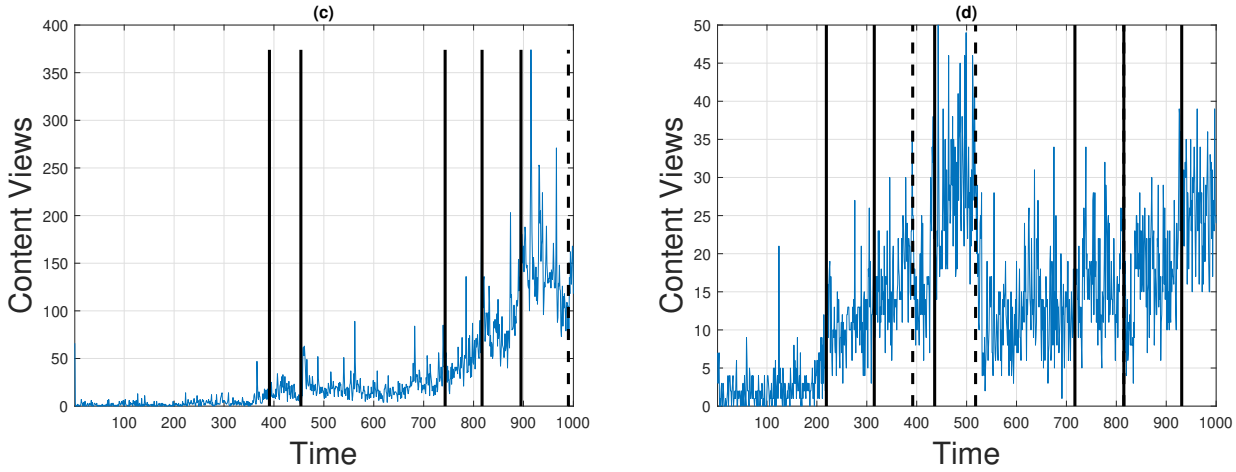


Figure 3.6: Outputs of the RCPD algorithm using standard type CUSUM for different time-series. Solid and dashed lines depict an upward and a downward change, respectively.

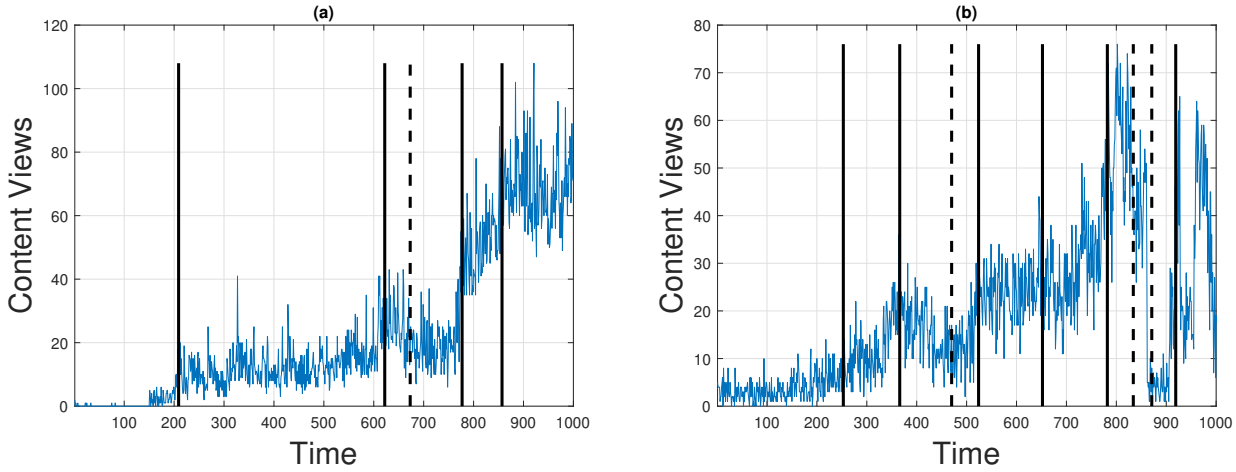


Figure 3.7: Outputs of the RCPD algorithm using ratio type CUSUM for different time-series. Solid and dashed lines depict an upward and a downward change, respectively.

the detected CPs split the time-series into stationary segments, which, additionally, confirms indirectly the accuracy of the off-line CP estimations over the changes in the real data.

3.7.5 Computational Complexity and Scalability

Finally, we present a MATLAB ® implementation of the overall algorithm with a large number of time-series (882 in this experiment) to quantify its performance in terms of processing cost. The computational time is measured on a Lenovo IdeaPad 510-15IKB laptop, with an Intel Core i7-7500U @ 2.70 GHz processor and 12 GB RAM. In Fig. 3.9, we show the aggregate processing cost per time instance for the two on-line methods and the total number of time-series. For the first 100 time instances, the algorithm collects the initial data, since it bootstraps. The peaks indicate the off-line part of the algorithm, which is more processing demanding mainly due to the segmentation algorithms running in parallel. The on-line part in the standard on-line algorithm indicates a linear complexity, since it is based on (3.18), while the equivalent quantity in (3.21) of the ratio-type is more CPU intensive, justifying the comparatively higher processing cost of the latter algorithm. In both cases, the aggregate processing cost is typically much less than a second, which demonstrates the lightweight nature of the proposed scheme. Such results could be further improved with a distributed deployment

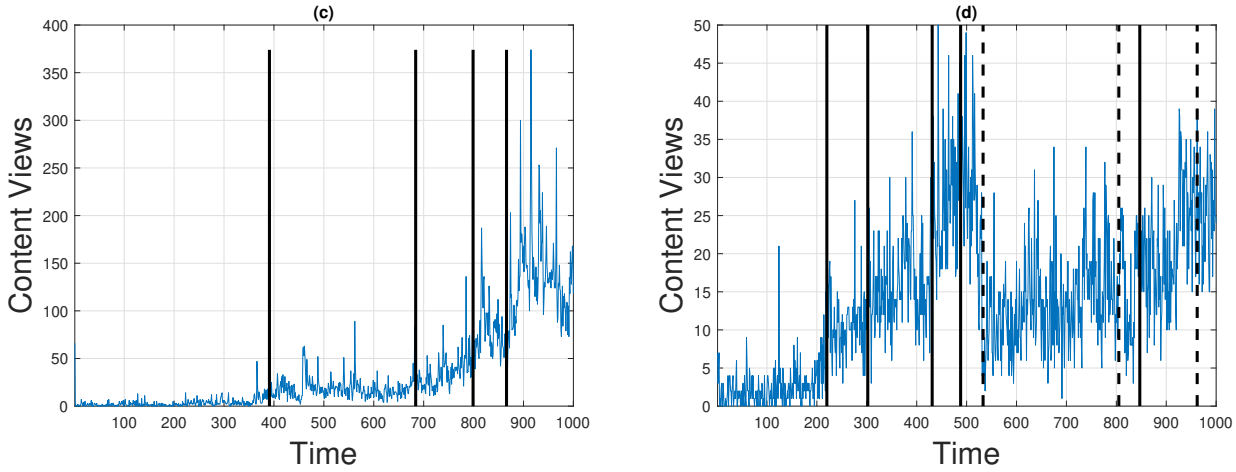


Figure 3.8: Outputs of the RCPD algorithm; using ratio type CUSUM for different time-series. Solid and dashed lines depict an upward and a downward change, respectively.

Table 3.7: Percentages of time-series with Time Dependencies Exceeding t Samples

t	≥ 1	≥ 5	≥ 15	≥ 30	≥ 50
piecewise	0.93	0.57	0.23	0.05	0.04

of scheme replicas since each of the time-series could be processed independently.

3.8 The RCPD Algorithm in a Load Balancing Scenario

In this Section, we demonstrate our proposal in a real content distribution scenario, balancing the traffic between web clients and content caches with a bespoke DNS-based load-balancer. We implement the RCPD algorithm as a client-server MATLAB ® application. The RCPD engine receives periodic content popularity measurements; if a CP is detected, the corresponding upward or downward changes are signalled to the load balancer. The load balancer: (i) distributes the load between the deployed content caches, in a round-robin fashion; (ii) tracks content visits and communicates them to the RCPD engine; and (iii) deploys or removes content caches based on the RCPD outputs.

We implement the web clients using with the httpperf tool (https://github.com/httperf/httperf). The number of clients at each time instance is based on a real time-series of YouTube content views, illustrated in Fig. 3.10a. In practice, an experimental run without the RCPD mechanisms uses three content caches constantly and a run with the RCPD mechanism enabled uses initially two and then three, four and five content caches, after each of the three detected change points, respectively. As we show in Fig. 3.10b, the web clients improve their connectivity times to download the content, while as demonstrated in Fig. 3.10c the CPU utilization in the servers hosting the content remains almost the same. A relevant experimental platform is presented in [6].

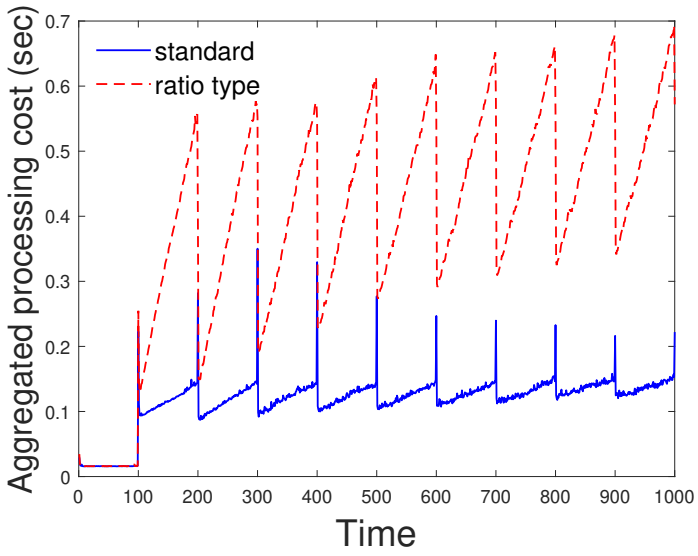


Figure 3.9: The aggregated overall processing cost, per time-instance, of the RCPD algorithm over 882 time-series.

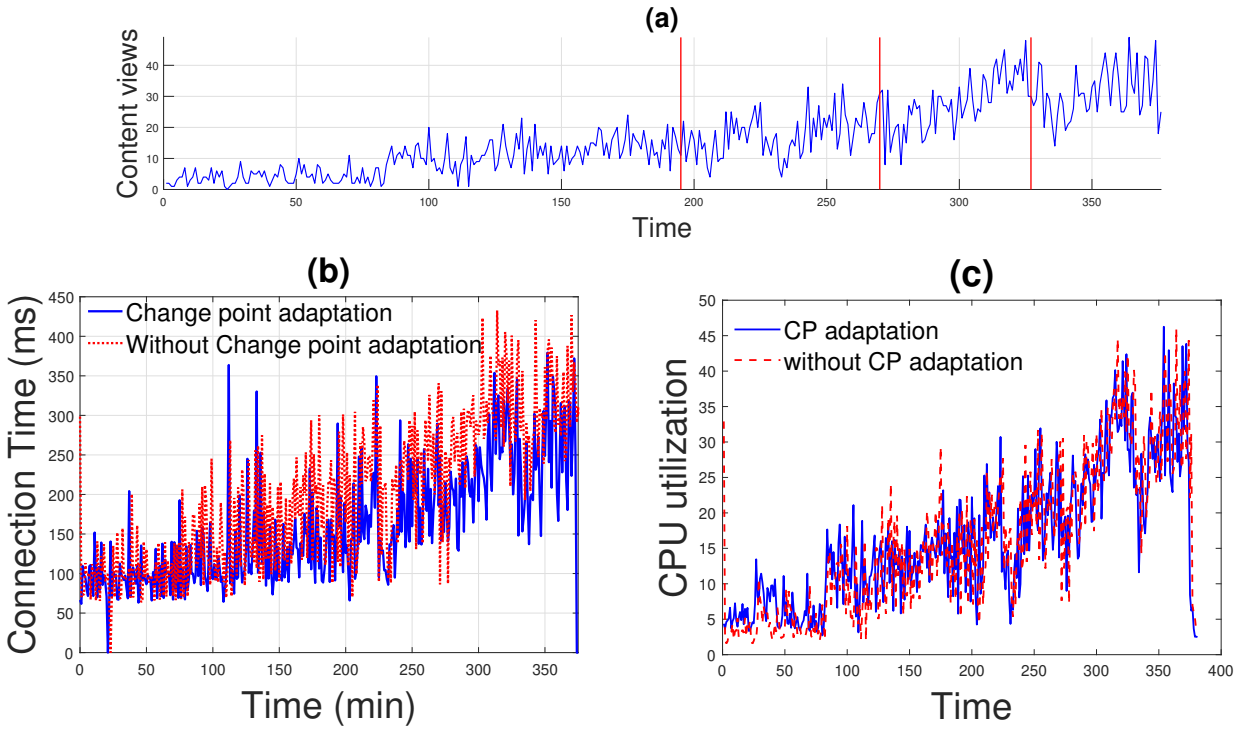


Figure 3.10: a) time-series of video content views, red lines depict the detected CPs, b) the connection time with and without RCPD adaptation and c) the equivalent servers' CPU utilization.

3.9 Application of the RCPD for Intrusion Detection in SDWSNs

Considering the limitations of previous works in SDWSN anomaly detection, outlined in Section 3.3, our main objective is to propose in the remainder of this Chapter a mechanism for DDoS detection with, i) a high detection rate, and, ii) low complexity, so that it would be suitable for "restricted" networks. To this end, we propose the employment of the RCPD. As will be explained in detail next, we study two different DDoS attacks: a false data flow forwarding (FDFF) attack, and a false neighbor

information (FNI) attack, chosen to illustrate the proposed algorithm's capabilities in the case of specific SDWSN vulnerabilities that exhibit largely different behavior. Both attacks are explained in Section 3.9.2, in the following.

3.9.1 SDWSN Security Analysis

The SDN networks security threats are grouped in three sets [50]: application plane attacks, control plane attacks, and data plane attacks. Among the three, the control plane attacks are pointed out as the most high impact and attractive [50] [51], as the control plane is responsible for the overall management of the network [52]. This characteristic turns the control plane prone to distributed denial of service (DDoS) attacks. For example, an intruder may flood the network with flow rule requests, which could lead to an exhaustion of the controller's resources. This attack can be intensified using multiple intruders.

The threats and vulnerabilities explained before also apply to SDWSNs. Moreover, there are specific attacks that can attain SDWSNs due to resources constraints, for example: in SDWSN the forwarding devices have low storage capacity, which limits the memory assigned for flow tables and buffers. These constraints make the forwarding devices prone to saturation attacks. Also, SDWSN networks are characterized for having a limited bandwidth and low processing power. This means that a saturation attack can also result in a DoS attack.

Another vulnerability concerns the gateway between the SDN controller and the WSN. The gateway has a radio module of limited bandwidth, rendering it a weak link even when the controller has enough resources to overcome an attack.

For the reasons outlined above, most of the security mechanisms designed for standard SDN networks have to be adapted or redesigned. This is one of the major challenges for SDWSN security.

3.9.2 Impact of DDoS Attacks on Network Performance

Based on SDWSN specific security vulnerabilities, in a previous work, we studied the impact of three DDoS attacks on SDWSN performance [53]. The attacks investigated were: false flow request (FFR), false data flow forwarding (FDFF), and false neighbor information (FNI).

The FFR attack aimed at increasing the SDWSN controller's processing overhead, as well as the packets' traffic, thus, increasing the number of collisions. Each attacker sent multiple flow rule requests to the controller, while the latter calculated the rule and replied to the request. The impact of the attack was observed to be negligible. The FDFF attack followed the FFR attack main idea of sending false flow rule requests to the controller, however, the execution was based on using each attacker's neighbors (benign nodes). Each attacker sent one data packet to its neighbors tagged with an unknown flow identifier; as the neighbors did not have a rule to apply to the packet, they sent a flow request to the controller asking a rule for the unknown flow identifier. Thus, compared to the FFR, the intensity of the attack was multiplied by the number of neighbors. The FDFF attack tripled the number of control packets in the whole network, but had a minor impact on the delivery rate. For both control and data packets, the delivery rate decreased only between 2% and 4%.

In the FNI attack, each attacker intercepted packets containing neighbor information, modified them with false neighbor information and forwarded them to the controller. The controller updated the network topology graph using the false information, and then reconfigured the network with wrong forwarding rules. Our main results [53] showed that the FNI attack could double the number of control packets in the whole network and had a significant impact on the delivery rate. In the case of the control packets, the delivery rate decreased between 35% and 50%. In the case of the data packets, the delivery rate decreased between 20% and 70%.

3.9.3 RCPD for Intrusion Detection

We employed the RCPD algorithm in SDWSNs under FDFF and FNI attacks. We simulated grid topologies with 36 and 100 nodes, varying the number of attackers in the network (5% and 20%). Each simulation run during 10 hours and each scenario was replicated 30 times. During the first 8 hours the network operated normally, then the attack was triggered. The choice of 8 hours was made because empirically it was seen that we needed at least 250 samples for the training period and we obtained one sample every 2 minutes. The simulations were performed using the COOJA simulator [54] and sky motes. The MAC layer was the IEEE 802.15.4, configured to work without radio duty cycle (`nullrdc_driver`). The data sink received the application data, while the management sink received performance metrics information. Notice that the SDN controller is a different node from the sink. Table 3.8 depicts the simulation parameters.

Table 3.8: Simulation Parameters

Simulation parameters	
Topology	Square grid
Number of nodes	36 and 100
Simulation duration	36000 s
Node boot interval	$[0, 1]$ s
Number of sinks	2
Sinks position	Middle of the grid edge
Data traffic rate	1 packet every 30 seconds
Management traffic rate	1 packet every two minutes
Data payload size	10 bytes
Management payload size	10 bytes
Data traffic start time	$[2, 3]$ min
Radio module power	0 dB
Distance between neighbors	50 m
Attacks begins after	28800 s

IT-SDN parameters	
Controller position	center
ND protocol	Collect-based
Link metric	ETX
CD protocol	none
Flow setup	source routed
Route calculation algorithm	Dijkstra
Route recalculation threshold	10%
Flow setup types	regular or source routed
Flow table size	10 entries

We analyzed the data packets delivery rate and the control packets overhead. The delivery rate

was calculated by dividing the total number of packets successfully received by the total number of packets sent. The control packets overhead was quantified as the total amount of control packets sent. Those metrics were updated every two minutes.

The metrics measuring the performance of the intrusion detection algorithm were the following: i) the detection rate (DR); ii) the false positive rate (FPR); iii) the false negative rate (FNR); iv) the detection time median (DTM), indicating the median of the time instances elapsed from the launch of the attack to the instance it was identified; and v) the median absolute deviation (MAD). The detection rate is defined as the ratio between the correctly detected attacks and the total number of attacks. The false positive rate is defined as the ratio between the number of attack events classified as attack and the total number of attack events. The false negative rate is defined as the ratio between attack events classified as non-attack event and the number of attack events. The detection time median is defined as the median of the number of samples required to detect the attack. The median absolute deviation measures the variability of the detection times and is calculated as shown in (3.34), where X_i is the detection time for replication i, and $\tilde{X}$ is the median of all the detection times,

$$\text{MAD} = \text{median}(|X_i - \tilde{X}|). \tag{3.34}$$

The delivery rate and control overhead time series were analyzed for three monitoring windows and three critical values. We used monitoring periods $K \in \{50, 100, 150\}$ samples. This means that the test statistic was run over K samples to extract changes in the mean value. As critical values we used $\alpha \in \{90\%, 95\%, 99\%\}$. Finally, in this analysis, we discarded the first 15 samples because during this time the network was bootstrapping.

3.10 Results and Analysis

In this Section we present and analyze the simulation results. In Section 3.10.1 we compare the FDFF attack detection performance when monitoring the data packets delivery rate and the control overhead. In Section 3.10.2 we repeat this analysis for the FNI attack.

3.10.1 FDFF Attack Detection

Tables 3.9 and 3.10 summarize the FDFF attack detection results when 5% of nodes are attackers. The results show that when monitoring the data packets delivery rate, the DR is between 57% and 73% for 36 nodes, and between 60% and 83% for 100 nodes. The results when monitoring the control packets overhead show two main points: (i) the algorithm has the same detection performance if configured with a monitoring period K of 50 or 150 samples, and (ii) when the monitoring period is configured as $K = 100$ samples we obtained a DR between 97% and 100%.

Comparing the FPR and the FNR metrics, we observed that the number of cases classified as false negative is higher than the number of cases classified as false positive. This means, it is more common for the algorithm not to detect a change in the metrics when the network is under attack than to detect a suspicious change in a network without attackers. For example, looking at the results when monitoring the control overhead in Table 3.9, only in one out of nine cases the FPR was different than zero. Conversely, the FNR was different than zero in six of nine cases.

The DTM (detection time median) results show that when monitoring the control packets overhead, the attack detection is faster than when monitoring the delivery rate in all the cases. When monitoring the data packets delivery rate, the DTM is between 31 and 37 samples for 36 nodes, and between 20 and 31 samples for 100 nodes. When monitoring the control packets overhead, the DTM is between 9 and 19 samples for 36 nodes, and between 10 and 19 samples for 100 nodes. The fastest detection is obtained monitoring the control packets overhead using a monitoring period of 100 samples, highlighted in red color.

Table 3.9: FDFF Attack Detection, 36 Nodes, 5% Attackers

Data packets delivery rate

K	**50**			**100**			**150**		
α	**90**	**95**	**99**	**90**	**95**	**99**	**90**	**95**	**99**
DTM	31	33	31	31	37	33	31	31	31
MAD	4	6	4	8	9	10	4	4	4
DR	63	67	67	57	70	63	67	73	70
FPR	7	10	7	0	0	0	0	0	0
FNR	30	23	27	43	30	37	33	27	30

Control overhead

K	**50**			**100**			**150**		
α	**90**	**95**	**99**	**90**	**95**	**99**	**90**	**95**	**99**
DTM	19	16	18	12	9	11	19	16	18
MAD	3	3	3	3	2	2	3	3	3
DR	67	73	67	100	97	100	67	73	67
FPR	0	0	0	0	3	0	0	0	0
FNR	33	27	33	0	0	0	33	27	33

Table 3.10: FDFF Attack Detection, 100 nodes, 5% Attackers

Data packets delivery rate

K	**50**			**100**			**150**		
α	**90**	**95**	**99**	**90**	**95**	**99**	**90**	**95**	**99**
DTM	24	26	27	22	20	21	29	31	31
MAD	7	6	13	9	10	11	13	9	15
DR	60	67	67	77	83	73	63	67	63
FPR	23	20	20	10	7	13	0	3	7
FNR	17	13	13	13	1	13	37	30	30

Control overhead

K	**50**			**100**			**150**		
α	**90**	**95**	**99**	**90**	**95**	**99**	**90**	**95**	**99**
DTM	19	17	19	13	10	12	19	17	19
MAD	3	3	3	3	2	3	3	3	3
DR	60	73	63	100	100	100	60	73	63
FPR	0	0	0	0	0	0	0	0	0
FNR	40	27	37	0	0	0	40	27	37

Tables 3.11 and 3.12 summarize the FDFF attack detection results when 20% of nodes are attackers. In the case of 36 nodes, the DR was between 73% and 83% when monitoring the data packets delivery

Table 3.11: FDFF Attack Detection, 36 nodes, 20% Attackers

Data packets delivery rate

K	**50**			**100**			**150**		
α	**90**	**95**	**99**	**90**	**95**	**99**	**90**	**95**	**99**
DTM	28	28	28	30	24	28	29	28	28
MAD	5	8	6	11	7	8	6	5	8
DR	77	80	73	73	83	73	77	80	77
FPR	3	07	7	0	3	0	0	3	0
FNR	20	13	20	27	13	27	23	17	23

Control overhead

K	**50**			**100**			**150**		
α	**90**	**95**	**99**	**90**	**95**	**99**	**90**	**95**	**99**
M	8	7	7	5	5	5	8	7	7
MAD	2	2	2	1	1	1	2	2	2
DR	100	100	100	97	87	97	100	100	100
FPR	0	0	0	3	13	3	0	0	0
FNR	0	0	0	0	0	0	0	0	0

Table 3.12: FDFF Attack Detection, 100 nodes, 20% Attackers

Data packets delivery rate

K	**50**			**100**			**150**		
α	**90**	**95**	**99**	**90**	**95**	**99**	**90**	**95**	**99**
DTM	15	13	14	8	7	7	15	14	14
MAD	5	6	5	6	5	5	5	5	5
DR	100	93	100	97	93	97	100	97	97
FPR	0	7	0	3	7	3	0	3	3
FNR	0	0	0	0	0	0	0	0	0

Control overhead

K	**50**			**100**			**150**		
α	**90**	**95**	**99**	**90**	**95**	**99**	**90**	**95**	**99**
DTM	4	4	4	3	3	3	4	4	4
MAD	0	0	0	0	0	0	0	0	0
DR	100	97	100	97	90	97	100	97	100
FPR	0	3	0	3	10	3	0	3	0
FNR	0	0	0	0	0	0	0	0	0

rate, and between 87% and 100% when monitoring the control packets overhead. In terms of detection time, the best DTM when monitoring the data packets delivery rate was 24 samples and the DTM

Table 3.13: FNI Attack Detection, 36 nodes, 5% Attackers

Data packets delivery rate

K	**50**			**100**			**150**		
α	**90**	**95**	**99**	**90**	**95**	**99**	**90**	**95**	**99**
DTM	7	6	7	8	7	6	7	6	6
MAD	3	4	3	4	3	3	2	4	4
DR	93	83	93	93	80	93	93	83	87
FPR	0	10	0	0	13	0	0	10	7
FNR	7	7	7	7	7	7	7	7	6

Control overhead

K	**50**			**100**			**150**		
α	**90**	**95**	**99**	**90**	**95**	**99**	**90**	**95**	**99**
DTM	28	25	27	35	26	33	28	25	27
MAD	6	7	9	4	3	5	6	7	9
DR	27	33	27	20	27	23	27	33	27
FPR	3	3	3	0	0	0	0	0	0
FNR	70	63	70	80	73	77	73	67	73

Table 3.14: FNI Attack Detection, 100 nodes, 5% Attackers

Data packets delivery rate

K	**50**			**100**			**150**		
α	**90**	**95**	**99**	**90**	**95**	**99**	**90**	**95**	**99**
DTM	6	6	6	6	6	6	6	6	6
MAD	4	4	3	3	3	2	4	4	4
DR	87	93	83	83	83	83	83	90	87
FPR	13	7	17	17	17	17	13	10	13
FNR	0	0	0	0	0	0	3	0	0

Control overhead

K	**50**			**100**			**150**		
α	**90**	**95**	**99**	**90**	**95**	**99**	**90**	**95**	**99**
DTM	34	29	33	35	37	37	34	29	33
MAD	7	7	7	10	7	8	7	8	8
DR	63	70	67	30	47	37	63	70	67
FPR	0	0	0	0	0	0	0	0	0
FNR	37	30	33	70	53	63	37	30	33

when monitoring the control packets overhead was 5 samples. Configuring the monitoring period in 100 we obtain the best DTM, but there was a drop in the DR if compared with the cases when using

Table 3.15: FNI Attack Detection, 36 nodes, 20% Attackers

Data packets delivery rate

K	**50**			**100**			**150**		
α	**90**	**95**	**99**	**90**	**95**	**99**	**90**	**95**	**99**
DTM	7	7	7	7	7	7	8	7	7
MAD	2	2	2	3	4	3	2	2	2
DR	100	100	100	100	100	100	100	100	100
FPR	0	0	0	0	0	0	0	0	0
FNR	0	0	0	0	0	0	0	0	0

Control overhead

K	**50**			**100**			**150**		
α	**90**	**95**	**99**	**90**	**95**	**99**	**90**	**95**	**99**
DTM	26	24	26	26	24	27	26	24	26
MAD	8	7	7	17	11	13	8	7	7
DR	57	70	60	43	63	57	57	70	60
FPR	0	0	0	0	0	0	0	0	0
FNR	43	30	40	57	37	43	43	30	40

Table 3.16: FNI Attack Detection, 100 nodes, 20% Attackers

Data packets delivery rate

K	**50**			**100**			**150**		
α	**90**	**95**	**99**	**90**	**95**	**99**	**90**	**95**	**99**
DTM	9	10	10	8	9	8	10	12	11
MAD	5	8	7	4	6	4	5	9	8
DR	100	100	100	100	100	100	100	100	97
FPR	0	0	0	0	0	0	0	0	3
FNR	0	0	0	0	0	0	0	0	0

Control overhead

K	**50**			**100**			**150**		
α	**90**	**95**	**99**	**90**	**95**	**99**	**90**	**95**	**99**
DTM	27	24	26	26	25	25	27	24	26
MAD	6	3	6	6	6	6	6	3	6
DR	93	97	97	93	97	93	93	97	97
FPR	0	0	0	0	0	0	0	0	0
FNR	7	3	3	7	3	7	7	3	3

monitoring periods of 50 and 150 samples.

The results for 100 nodes showed it is possible to obtain a DR of 100% monitoring any of the

metrics, but there were significant differences in the detection time. The DTM when monitoring the control overhead is between 3 and 4 samples, while when monitoring the data packets delivery rate the DTM was between 7 and 15 samples. Considering the earliest detection with the highest DR for both monitoring metrics, it occurred when using a monitoring period of 100 samples. For both cases the DR obtained was 97%. In terms of FPR and FNR, the best performance was obtained when monitoring the control overhead and using a monitoring period of 50 and 150 samples. Monitoring the control overhead using a monitoring window of 100 samples provided a FPR between 3% and 10%.

Summarizing, the algorithm was able to detect the FDFF attack using either the data packet packets delivery rate or the control packets overhead as inputs. Notably, the algorithm obtained a DR of 100% with both metrics when 20% of nodes behave as attackers. However, aiming for the quickest detection captured through the detection time median, the algorithm achieved far better results when monitoring the control packets overhead in all scenarios. This is a direct consequence of the type of the attack; the attacker creates multiple flow rule request packets to increase the packet traffic and the controller processing overhead. After some time, the flow table of the nodes around the attacker start to saturate, affecting the data packets delivery rate. This means that the change in the delivery will be detected only after the tables saturation; on the contrary, the number of control packets start to change immediately after the attack is triggered.

3.10.2 FNI Attack Detection

Tables 3.13 and 3.14 summarize the FNI attack detection results when 5% of nodes were attackers. Opposite to the FDFF attack results, the algorithm obtained a better performance detecting the FNI attack when monitoring the data packets delivery rate. In the case of 36 nodes, the DR when monitoring the data packets delivery rate was between 80% and 93%, and the DR when monitoring the control packets overhead was between 23% and 33%. In the case of 100 nodes, the DR when monitoring the data packets delivery rate was between 83% and 93%, and the DR when monitoring the control packets overhead was between 30% and 70%. This means, even the best DR when monitoring the control packets overhead was under the worse DR when monitoring the data packets delivery rate. Also, the results showed that using a critical value of 90%, we obtained a negligible FPR (in our simulation calculated zero). With respect to the DTM, the best result was obtained by monitoring the data packets delivery rate and the control packets overhead were 6 and 25 samples, respectively. This means the algorithm detected the attack four times faster when monitoring the data packets delivery rate. For 100 nodes, the best DTM when monitoring the data packets delivery rate remained in 6 samples, but when monitoring the control packets overhead it was 29 samples.

Lastly, Tables 3.15 and 3.16 summarize the FNI attack detection results when 20% of nodes were attackers. For 36 nodes, the results remained similar to the case of 5% of nodes are attackers. In the case of 100 nodes, the DR when monitoring the data packets delivery rate was between 97% and 100%, and the DR when monitoring the control packets delivery rate was between 93% and 97%. About the DTM, the results for the scenarios when monitoring the data packets delivery rate were between 4 and 9 samples. The results for this same metric when monitoring the control packets overhead were between 24 and 26 samples. This means, for grid topologies with 100 nodes where 20% of nodes were attackers, we obtained similar DRs regardless of the monitoring metric, but when monitoring the delivery rate the detection was at least 3 times faster.

Summarizing our findings, the algorithm was able to detect the FNI attack monitoring either the data packet packets delivery rate or the control packets overhead. Then, comparing the detection performance based on the detection rate and the detection time median, the algorithm obtained a far better performance when monitoring the data packets delivery rate in all scenarios. This effect was directly related to the type of the attack; in the FNI attack, the attackers intercept the control packets that contained neighbor information, modify them, and then forward them to the controller. This means this attack could lead to a network misconfiguration using few control packets.

3.11 Conclusion

In this Chapter, we proposed the RCPD, a novel algorithm for the real-time detection of changes in the mean value of content popularity. Approaching the problem statistically, we efficiently combined off-line and on-line non-parametric CUSUM procedures to avoid restrictive assumptions for content popularity behavior and to reduce the overall computational cost. We divided the algorithm in two phases. The first phase was an extended retrospective (off-line) procedure with a modified BS algorithm and was used to adjust on-line parameters, based on historical data of the particular video. The second phase integrated one of two alternative trend indicators to the sequential (on-line) procedure, to reveal the direction of a detected change. We provided extensive simulations, using synthetic and real data, that demonstrated the performance of the proposed algorithm for the successful identification of content popularity changes in real-time. We also demonstrated through experimental measurements that the RCPD's processing cost is almost imperceptible. Finally we provided proof-of-concept by applying the algorithm in a load balancing application, highlighting its efficiency in a realistic setting.

Furthermore, we have used the RCPD for intrusion detection in SDWSNs. We performed experiments for two SDWSN DDoS attacks, in topologies of 36 and 100 nodes, and with varying number of attackers. Our results showed that it is feasible to detect different types of attacks by monitoring either the data packets delivery rate or control packets metrics. As the detector's algorithmic complexity is linear to the size of the network and the number of metrics monitored, the proposed approach could scale to include other metrics.

References

[1] Tom Goethals, Merlijn Sebrechts, Ankita Atrey, Bruno Volckaert, and Filip De Turck. Unikernels vs containers: An in-depth benchmarking study in the context of microservice applications. In *IEEE Int. Symp. Cloud Service Comput. (SC2)*, Nov. 2018.

[2] Joao Martins, Ahmed Mohamed, Costin Raiciu, and Felipe Huici. Enabling fast, dynamic networking processing with ClickOS. In *Second ACM SIGCOMM Workshop on Hot Topics in Software Defined Networking*, pages 67–72, 2013.

[3] CISCO Visual Networking. Cisco global cloud index: forecast and methodology, 2015-2020. San Jose, CA, USA, CISCO, Tech. Rep., 2017.

[4] Nick McKeown, Tom Anderson, Hari Balakrishnan, Guru Parulkar, Larry Peterson, Jennifer Rexford, et al. Openflow: enabling innovation in campus networks. *ACM SIGCOMM Comput. Commun. Rev.*, 38(2):69–74, Mar. 2008.

[5] Necos project: Towards lightweight slicing of cloud federated infrastructures. https://intrig.dca.fee.unicamp.br/2017/09/05/necos-2-year-eu-brazil-collaborative-project-starting-in-nov2017/.

[6] Polychronis Valsamas, Sotiris Skaperas, and Lefteris Mamatas. Elastic content distribution based on unikernels and change-point analysis. In *Proc. 24th Eur. Wireless Conf. (EW)*, pages 1–7, Catania, Italy, May 2018.

[7] Alexandru Tatar, Marcelo Dias De Amorim, Serge Fdida, and Panayotis Antoniadis. A survey on predicting the popularity of web content. *J. Internet Services Appl.*, 5(1):8, Dec. 2014.

[8] Gabor Szabo and Bernardo A Huberman. Predicting the popularity of online content. *Commun. ACM*, 53(8):80–88, Aug. 2010.

[9] Gonca Gürsun, Mark Crovella, and Ibrahim Matta. Describing and forecasting video access patterns. In *Proc. IEEE Int. Conf. Comput. Commun. (IEEE INFOCOM)*, pages 16–20, Shanghai, China, Apr. 2011.

[10] Justin Cheng, Lada Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. Can cascades be predicted? In *Proc. 23rd Int. Conf. World Wide Web (WWW)*, pages 925–936, Seoul, Republic of Korea, Apr. 2014.

[11] Stefan Fremdt. Asymptotic distribution of the delay time in page's sequential procedure. *J. Statist. Planning Inference*, 145:74–91, Feb. 2014.

[12] Yannick Hoga. Monitoring multivariate time series. *J. Multivariate Anal.*, 155:105–121, Mar. 2017.

[13] E Brodsky and Boris S Darkhovsky. *Nonparametric methods in change point problems*. Dordrecht, The Netherlands: Kluwer, 2013.

[14] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *Proc. AAAI Workshop Knowl. Disc. Databases (KDD)*, volume 10, pages 359–370, Seattle, USA, Aug. 1994.

[15] Rohit J Kate. Using dynamic time warping distances as features for improved time series classification. *Data Mining Knowledge Discovery*, 30:283–312, Mar. 2016.

[16] D. Kreutz, F. M. V. Ramos, P. E. Veríssimo, C. E. Rothenberg, S. Azodolmolky, and S. Uhlig. Software-Defined Networking: A Comprehensive Survey. *Proc. IEEE Proc.*, 103(1):14–76, Jan 2015.

[17] H. I. Kobo, A. M. Abu-Mahfouz, and G. P. Hancke. A Survey on Software-Defined Wireless Sensor Networks: Challenges and Design Requirements. *IEEE Access*, 5:1872–1899, 2017.

[18] Nick McKeown, Tom Anderson, Hari Balakrishnan, Guru Parulkar, Larry Peterson, Jennifer Rexford, Scott Shenker, and Jonathan Turner. OpenFlow: Enabling Innovation in Campus Networks. *SIGCOMM Comput. Commun. Rev.*, 38(2):69–74, March 2008.

[19] R. C. A. Alves, D. A. G. Oliveira, G. A. Nunez Segura, and C. B. Margi. The Cost of Software-Defining Things: A Scalability Study of Software-Defined Sensor Networks. *IEEE Access*, 7:115093–115108, Aug 2019.

[20] Sotiris Skaperas, Lefteris Mamatas, and Arsenia Chorti. Early Video Content Popularity Detection with Change Point Analysis. In *IEEE Global Commun. Conf. (GLOBECOM)*, Abhu-Dhabi, United Arab Emirates, December 2018.

[21] S. Skaperas, L. Mamatas, and A. Chorti. Real-Time Video Content Popularity Detection Based on Mean Change Point Analysis. *IEEE Access*, 7:142246–142260, 2019.

[22] Henrique Pinto, Jussara M Almeida, and Marcos A Gonçalves. Using early view patterns to predict the popularity of youtube videos. In *Proc. 6th ACM Int. Conf. Web Search and Data Mining (WSDM)*, pages 365–374, Rome, Italy, Feb. 2013.

[23] Sotiris Skaperas, Lefteris Mamatas, and Arsenia Chorti. Early video content popularity detection with change point analysis. In *Proc. IEEE Global Commun. Conf. (IEEE GLOBECOM)*, pages 1–7, Abu Dhabi, UAE, Dec. 2018.

[24] Michèle Basseville, Igor V Nikiforov, et al. *Detection of abrupt changes: theory and application*, volume 104. Prentice Hall Englewood Cliffs, 1993.

[25] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surveys (CSUR)*, 41(3):1–58, Sept. 2009.

[26] Alexander G Tartakovsky, Aleksey S Polunchenko, and Grigory Sokolov. Efficient computer network anomaly detection by changepoint detection methods. *IEEE J. Sel. Topics Signal Process.*, 7(1):4–11, Feb. 2013.

[27] Alexander G Tartakovsky, Boris L Rozovskii, Rudolf B Blazek, and Hongjoong Kim. A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods. *IEEE Trans. Signal Process.*, 54(9):3372–3382, Sept. 2006.

[28] Haining Wang, Danlu Zhang, and Kang G Shin. Change-point monitoring for the detection of dos attacks. *IEEE Trans. Depend. Sec. Comput.*, 1(4):193–208, Oct.-Dec. 2004.

[29] Yanxiang Jiang, Miaoli Ma, Mehdi Bennis, Fuchun Zheng, and Xiaohu You. A novel caching policy with content popularity prediction and user preference learning in fog-ran. In *Proc. IEEE Global Commun. Conf. (IEEE GLOBECOM) Workshops*, pages 1–6, 2017.

[30] Gautam Thatte, Urbashi Mitra, and John Heidemann. Parametric methods for anomaly detection in aggregate traffic. *IEEE/ACM Trans. Netw. (TON)*, 19(2):512–525, Apr. 2011.

[31] Augustin Soule, Kavé Salamatian, and Nina Taft. Combining filtering and statistical methods for anomaly detection. In *Proc. 5th ACM SIGCOMM Conf. Internet Measurement*, pages 1–14, New York, NY, USA, Oct. 2005.

[32] Ido Nevat, Dinil Mon Divakaran, Sai Ganesh Nagarajan, Pengfei Zhang, Le Su, Li Ling Ko, and Vrizlynn LL Thing. Anomaly detection and attribution in networks with temporally correlated traffic. *IEEE/ACM Trans. Netw. (TON)*, 26(1):131–144, Feb. 2018.

[33] S. S. Bhunia and M. Gurusamy. Dynamic attack detection and mitigation in IoT using SDN. In *27th Int. Telecommun. Netw. and Appl. Conf. (ITNAC)*, pages 1–6, Nov 2017.

[34] D. Yin, L. Zhang, and K. Yang. A DDoS Attack Detection and Mitigation With Software-Defined Internet of Things Framework. *IEEE Access*, 6:24694–24705, 2018.

[35] Rui Wang, Zhiyong Zhang, Zhiwei Zhang, and Zhiping Jia. ETMRM: An Energy-efficient Trust Management and Routing Mechanism for SDWSNs. *Computer Networks*, 139:119 – 135, 2018.

[36] Xiaobo Zhou and Cheng-Zhong Xu. Optimal video replication and placement on a cluster of video-on-demand servers. In *in Proc. Int. Conf. Parallel Process. (ICPP)*, pages 547–555, Vancouver, Canada, Aug. 2002.

[37] Wenting Tang, Yun Fu, Ludmila Cherkasova, and Amin Vahdat. Modeling and generating realistic streaming media server workloads. *Comput. Netw.*, 51(1):336–356, Jan. 2007.

[38] Alexander Aue and Lajos Horváth. Structural breaks in time series. *J. Time Series Anal.*, 34(1):1–16, Jan. 2013.

[39] Donald WK Andrews. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica: J. Econometric Soc.*, 59:817–858, May 1991.

[40] Dominik Wied. A nonparametric test for a constant correlation matrix. *Econometric Rev.*, 36(10):1157–1172, Apr. 2017.

[41] Marc Lavielle and Gilles Teyssiere. Adaptive detection of multiple change-points in asset price volatility. In *Long Memory in Economics*, pages 129–156. Springer, G. Teyssiere and A. Kirkman, Eds. Berlin, Germany: Springer–Verlag, 2007.

[42] Daniele Angelosante and Georgios B Giannakis. Sparse graphical modeling of piecewise-stationary time series. In *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process (IEEE ICASSP)*, pages 1960–1963, Prague, Czech Republic, May 2011.

[43] Carla Inclan and George C Tiao. Use of cumulative sums of squares for retrospective detection of changes of variance. *J. Amer. Statist. Assoc.*, 89(427):913–923, Sept. 1994.

[44] Huang Kai, Qi Zhengwei, and Liu Bo. Network anomaly detection based on statistical approach and time series analysis. In *Proc. Int. Conf. Advanced Inform. Netw. Appl. (WAINA) Workshops*, pages 205–211, Bradford, UK, May 2009.

[45] Nesrine Ben Hassine, Ruben Milocco, and Pascale Minet. Arma based popularity prediction for caching in content delivery networks. In *Proc. Wireless Days*, pages 113–120, Porto, Portugal, Mar. 2017.

[46] Dominik Wied and Pedro Galeano. Monitoring correlation change in a sequence of random variables. *J Statist. Planning Inference*, 143(1):186–196, Jan. 2013.

[47] Mattia Zeni, Daniele Miorandi, and Francesco De Pellegrini. Youstatanalyzer: a tool for analysing the dynamics of youtube content popularity. In *Proc. 7th Int. Conf. Perform. Eval. Methodol. Tools*, pages 286–289, Torino, Italy, Dec. 2013.

[48] Richard G Clegg. A practical guide to measuring the hurst parameter. *Int. J. Simul. Syst. Sci. Technol.*, 7(2):3–14, Nov. 2006.

[49] Jean-Marc Bardet, Gabriel Lang, Georges Oppenheim, Anne Philippe, Stilian Stoev, and Murad S Taqqu. Semi-parametric estimation of the long-range dependence parameter: a survey. *Theory and applications of long-range dependence*, pages 557–577, 2003.

[50] I. Ahmad, S. Namal, M. Ylianttila, and A. Gurtov. Security in Software Defined Networks: A Survey. *IEEE Commun. Surveys Tuts.*, 17(4):2317–2346, Fourthquarter 2015.

[51] Zhaogang Shu, Jiafu Wan, Di Li, Jiaxiang Lin, Athanasios V. Vasilakos, and Muhammad Imran. Security in Software-Defined Networking: Threats and Countermeasures. *Mobile Netw. and Appl.*, 21(5):764–776, Oct 2016.

[52] A. Akhunzada, E. Ahmed, A. Gani, M. K. Khan, M. Imran, and S. Guizani. Securing software defined networks: taxonomy, requirements, and open issues. *IEEE Commun. Mag.*, 53(4):36–44, April 2015.

[53] Gustavo A. Nunez Segura, Cintia B. Margi, and Arsenia Chorti. Understanding the Performance of Software Defined Wireless Sensor Networks Under Denial of Service Attack. *Open Journal of Internet Of Things (OJIOT)*, 2019. Special Issue: Proc. Int. Workshop Very Large Internet of Things (VLIoT 2019) in conjunction with the VLDB 2019 Conf. Los Angeles, United States.

[54] F. Osterlind, A. Dunkels, J. Eriksson, N. Finne, and T. Voigt. Cross-Level Sensor Network Simulation with COOJA. In *Proc. IEEE Conf. Local Comput. Netw. (LCN)*, pages 641–648, Nov 2006.

Chapter 4

Uplink Non-Orthogonal Multiple Access (NOMA) Under Statistical QoS Delay Constraints

4.1 Introduction

Various verticals in 5G and beyond (B5G) networks require very stringent latency guarantees, while at the same time envisioning massive connectivity. As a result, choosing the optimal multiple access (MA) technique to achieve low latency is a key enabler of B5G. In particular, this issue is more acute in uplink transmissions due to the potentially high number of collisions. On this premise, in the present contribution we discuss the issue of delay-sensitive uplink connectivity; to this end, we perform a comparative analysis of various MA approaches with respect to the achievable effective capacity (EC). As opposed to standard rate (PHY) or throughput (MAC) analyses, we propose the concept of the effective capacity as a suitable metric for characterizing jointly PHY-MAC layer delays. The palette of investigated MA approaches includes standard orthogonal MA (OMA) and power domain non-orthogonal MA (NOMA) in uplink scenarios.

For two-user networks, we propose novel closed-form expressions for the EC of the NOMA users and show that in the high signal to noise ratio (SNR) region, the "strong" NOMA user has a limited EC, assuming the same delay constraint as the "weak" user. We demonstrate that for the weak user, OMA achieves higher EC than NOMA at small values of the transmit SNR, while NOMA outperforms OMA in terms of EC at high SNRs. On the other hand, for the strong user the opposite is true, i.e., NOMA achieves higher EC than OMA at small SNRs, while OMA becomes more beneficial at high SNRs. This result raises the question of introducing "adaptive" OMA / NOMA policies, based jointly on the users' delay constraints as well as on the available transmit power.

4.2 Contributions and Chapter Organization

Non-orthogonal multiple access (NOMA) schemes have attracted a lot of attention recently, allowing multiple users to be served simultaneously with enhanced spectral efficiency; it is known that the boundary of achievable rate pairs (in the case of two users) using NOMA is outside the capacity region achievable with orthogonal multiple access (OMA) techniques [1] or other schemes [2]. Superior achievable rates are attainable through the use of superposition coding at the transmitter and of successive interference cancellation (SIC) at the receiver [3, 4]. The SIC receiver decodes multi-user signals with descending received signal power and subtracts the decoded signal(s) from the received superimposed signal, so as to improve the signal-to-interference ratio. The process is repeated until the signal of interest is decoded. In uplink NOMA networks, the strongest user's signal is decoded first (as opposed to downlink NOMA networks in which the inverse order is applied).

Besides, in a number of emerging applications, delay QoS requirements become increasingly important, e.g., for URLLC systems. Furthermore, in future wireless networks, users are expected to necessitate flexible delay guarantees for achieving different service requirements. In order to satisfy diverse delay requirements, a simple and flexible delay QoS model is imperative to be applied and investigated. In this respect, the EC theory can be employed [5], [6] [7], with EC denoting the maximum constant arrival rate which can be served by a given service process, while guaranteeing the required statistical delay provisioning. We studied delay-constrained downlink NOMA networks in [4] and with secrecy constraints [8] in [9]. The present analysis complements [4], focusing on uplink transmissions. In the following Sections, we derive novel closed-form expressions for the ECs of a two user network; we then provide four Lemmas for the asymptotic performance of the network with NOMA and OMA. The conclusions drawn are supported by an extensive set of simulations.

The rest of the Chapter is organized as follows. In Section 4.3 we investigate the EC of a two user uplink NOMA system under statistical delay QoS constraints. Simulation results are given in Section 4.4, followed by conclusions in Section 4.5.

4.3 Effective Capacity of Two-user NOMA Uplink Network

Assume a two-user NOMA uplink network with users U_1 and U_2 in a Rayleigh block fading propagation channel, with respective channel gains during a transmission block denoted by $|h_1|^2 < |h_2|^2$. The users transmit corresponding symbols S_1, S_2 respectively, with power $\mathbb{E}[|S_i|^2] = P_i, i = 1, 2$ and total power $P_T = \sum_{i=1}^{2} P_i = 1$. Here, P_i is the power coefficient for the user i [10]. The received superimposed signal can be expressed as [11]

$$Z = \sum_{i=1}^{2} \sqrt{P_i} h_i S_i + w, \tag{4.1}$$

where w denotes a zero mean circularly symmetric complex Gaussian random variable with variance σ^2. The receiver will first decode the symbol of the strong user treating the transmission of the weak as interference. After decoding it, the receiver will suppress it from Z and decode the signal of the weak user. Following the SIC principle and denoting by $\rho = \frac{1}{\sigma^2}$ the transmit SNR, the achievable rates, in b/s/Hz, for user $\mathrm{U}_i, i = 1, 2$, is expressed as: [12]

$$R_i = \log_2 \left(1 + \frac{\rho P_i |h_i|^2}{1 + \rho \sum_{l=1}^{i-1} P_l |h_l|^2} \right). \tag{4.2}$$

Introducing statistical delay QoS constraints, let θ_i be the statistical delay QoS exponent of the i-th user, and assume that the service process satisfies the Gärtner-Ellis theorem [6]. The delay exponent θ_i captures how strict the delay constraint is [6]. A slower decay rate can be represented by a smaller θ_i, which indicates that the system is more delay tolerant, while a larger θ_i corresponds to a system with more stringent QoS requirements. Applying the EC theory in a uplink NOMA with two users, the i-th user's EC over a block-fading channel, is defined as:

$$E_c^i = -\frac{1}{\theta_i T_f B} \ln \left(\mathbb{E} \left[e^{-\theta_i T_\mathrm{f} B R_i} \right] \right) \quad \text{(in b/s/Hz)}, \tag{4.3}$$

where T_f is the fading-block length, B is the bandwidth and $\mathbb{E}[\cdot]$ denotes expectation over the channel gains. By inserting R_i into (4.3), we obtain the following expression for the EC of the i-th user

$$E_c^i = \frac{1}{\beta_i} \log_2 \left(\mathbb{E} \left[(1 + \frac{\rho P_i |h_i|^2}{1 + \rho \sum_{l=1}^{i-1} P_l |h_l|^2})^{\beta_i} \right] \right), \tag{4.4}$$

where $\beta_i = -\frac{\theta_i T_f B}{\ln 2}$, $i = 1, 2$, is the normalized (negative) QoS exponent.

4.3.1 ECs in a Two-user NOMA Uplink Network

For the ordering of the channel gains we make use of the theory of order statistics in the following analysis [13]. Assuming a Rayleigh wireless environment, the channel gains, denoted by $x_i = |h_i|^2, i = 1, 2$, are exponentially distributed with probability density function (PDF) and cumulative density function (CDF) respectively given by $f(x_i) = e^{-x_i}$, $F(x_i) = 1 - e^{-x_i}$. Then, according to order statistics [13], the ordered channel gains have respective PDFs $f_{i:2}(x_i), i = 1, 2$, and joint PDF $f(x_1, x_2)$ that are expressed as

$$f_{1:2}(x_1) = 2e^{-2x_1}, \tag{4.5}$$
$$f_{2:2}(x_2) = 2e^{-x_2}\left(1 - e^{-x_2}\right), \tag{4.6}$$
$$f(x_1, x_2) = 2e^{-x_1}e^{-x_2}. \tag{4.7}$$

As a result, the EC of User 1, denoted by E_c^1 is expressed as

$$\begin{aligned} E_c^1 &= \frac{1}{\beta_1}\log_2(\mathbb{E}[(1+\rho P_1 x_1)^{\beta_1}]) = \frac{1}{\beta_1}\log_2\left(\int_0^\infty (1+\rho P_1 x_1)^{\beta_1} f_{1:2}(x_1)dx_1\right) \\ &= \frac{1}{\beta_1}\log_2\left(\frac{2}{P_1\rho}\times U\left(1, 2+\beta_1, \frac{2}{\rho P_1}\right)\right). \end{aligned} \tag{4.8}$$

where $U(\cdot,\cdot,\cdot)$ denotes the confluent hypergeometric function [4]. On the other hand, the EC of the User 2 is evaluated as

$$\begin{aligned} E_c^2 &= \frac{1}{\beta_2}\log_2\left(\mathbb{E}\left[\left(1+\frac{\rho P_2 x_2}{1+\rho P_1 x_1}\right)^{\beta_2}\right]\right) = \frac{1}{\beta_2}\log_2\left(\int_0^\infty\int_{x_1}^\infty\left(1+\frac{\rho P_2 x_2}{1+\rho P_1 x_1}\right)^{\beta_2} f(x_1,x_2)dx_2dx_1\right) \\ &= \frac{1}{\beta_2}\log_2\left(2P_2^{1-\beta_2}(\rho P_2)^{\beta_2}e^{\frac{1}{\rho P_2}}e^{-\frac{(P_1-P_2)}{\rho P_2}}\right) + \frac{1}{\beta_2}\log_2\left(\sum_{j=0}^{-\beta_2}\binom{-\beta_2}{j}(\rho P_1)^j \times \sum_{k=0}^{\infty}\frac{(-1)^k(P_2-P_1)^k}{k!(1+j+k)}\right. \\ &\left.\times\left[\Gamma[2+\beta_2+j+k, \frac{1}{\rho P_2}] - (\rho P_2)^{-1-j-k}\Gamma[1+\beta_2, \frac{1}{\rho P_2}]\right]\right), \end{aligned} \tag{4.9}$$

with $\Gamma(\cdot,\cdot)$ denoting the incomplete Gamma function [4]. The proof for deriving E_c^1 is omitted as it can be verified with standard software (MAPLE or Mathematica) while for E_c^2 is provided in Appendix I.

In order to perform a comparative performance analysis, here we provide the achievable data rates for a two-user OMA network, denoted by $\widetilde{R}_i, i = 1, 2$, given as

$$\widetilde{R}_i = \frac{1}{2}\log_2\left(1+\rho P_T|h_i|^2\right), i = 1, 2 \tag{4.10}$$

Note that the coefficient $\frac{1}{2}$ is due to the equal allocation of resources to both users. The corresponding expressions are obtained for the ECs of both users in a OMA network, denoted by $\widetilde{E}_c^i$, given as

$$\tilde{E}_c^i = \frac{1}{\beta_i}\log_2\left(\mathbb{E}\left[(1+\rho P_T|h_i|^2)^{\frac{\beta_i}{2}}\right]\right)$$

so that,

$$\begin{aligned} \tilde{E}_c^1 &= \frac{1}{\beta_1}\log_2\left(\frac{2}{\rho}\times U\left(1, 2+\frac{\beta_1}{2}, \frac{2}{\rho}\right)\right) \\ \widetilde{E}_c^2 &= \frac{1}{\beta_2}\log_2\left(\frac{2}{\rho}\sum_{k=0}^{1}\binom{1}{k}(-1)^k \times U\left(1, 2+\frac{\beta_2}{2}, \frac{1+k}{\rho}\right)\right) \end{aligned} \tag{4.11}$$

The proof is omitted as it can be verified with software (MAPLE or Mathematica).

4.3.2 Asymptotic Analysis

We first perform an asymptotic analysis with respect to the SNR. Our results are summarized in Lemma 1.

Lemma 1: In the low and high SNR regimes, respectively, the following conclusions hold:

1. *When* $\rho \to 0$, *then,* $E_c^1 \to 0$, $E_c^2 \to 0$, $\widetilde{E}_c^1 \to 0$, $\widetilde{E}_c^2 \to 0$, $E_c^1 - \widetilde{E}_c^1 \to 0$, $E_c^2 - \widetilde{E}_c^2 \to 0$;
2. *When* $\rho \to +\infty$, *then* $E_c^1 \to +\infty$, $E_c^2 \to \frac{1}{\beta_2} \log_2 \left(\mathbb{E} \left[\left(1 + \frac{P_2 |h_2|^2}{P_1 |h_1|^2} \right)^{\beta_2} \right] \right)$, $\widetilde{E}_c^1 \to +\infty$, $\widetilde{E}_c^2 \to +\infty$, $E_c^1 - \widetilde{E}_c^1 \to +\infty$, $E_c^2 - \widetilde{E}_c^2 \to -\infty$.

Proof: The proof is provided in Appendix II.

Lemma 1 indicates that the ECs of both users are vanishingly small at low values of ρ, irrespective of employing NOMA or OMA. On the other hand, at high SNRs, we notice that the EC of the strong user with NOMA is limited to a finite value. On the contrary, for the weaker user, when $\rho >> 1$, its achievable EC in the NOMA uplink increases without bound. This is the exact opposite of the downlink scenario, where it is the weaker user which is limited in terms of EC, when $\rho >> 1$ [4].

Now, the question is how the ECs evolve with ρ between the two asymptotic regimes. To answer this question and to further analyze the impact of ρ on the individual EC, we look at the derivatives with respect to ρ [4] in Lemma 2.

Lemma 2: For the EC of User 1, in a two-user uplink network the following hold:

1. $\frac{\partial E_c^1}{\partial \rho} \geq 0$ *and* $\frac{\partial \widetilde{E}_c^1}{\partial \rho} \geq 0$, $\forall \rho$;
2. *When* $\rho \to 0$, *then* $\lim_{\rho \to 0} \left(\frac{\partial (E_c^1 - \widetilde{E}_c^1)}{\partial \rho} \right) = \frac{P_1 - \frac{1}{2}}{\ln 2} \mathbb{E}[|h_1|^2]$;
3. *When* $\rho >> 1$, *then* $\frac{\partial (E_c^1 - \widetilde{E}_c^1)}{\partial \rho} \approx \frac{1}{2\rho \ln 2} \geq 0$ *and it approaches* 0 *when* $\rho \to \infty$.

Proof: The proof is provided in Appendix III.

Lemma 2 indicates that for User 1, when the transmit SNR ρ is very small, the EC with OMA increases faster than the EC with NOMA. On the other hand, Lemma 2 shows that when the transmit SNR is very large, the EC with NOMA increases faster than with OMA.

Combining Lemma 2 and Lemma 1, we can conclude that, $E_c^1 - \widetilde{E}_c^1$ starts at vanishingly small value, first decreases, and subsequently increases to ∞ at a gradually reducing speed. This means that for the weaker user, OMA achieves higher EC than NOMA at small values of the transmit SNR ρ. At high values of ρ, NOMA becomes more beneficial for the weak user. Finally, when $\rho \to \infty$ the performance gain of NOMA over OMA reaches a constant value in the case of User 1.

Lemma 3: For the EC of User 2, in a two-user uplink network the following hold:

1. $\frac{\partial E_c^2}{\partial \rho} \geq 0$ *and* $\frac{\partial \widetilde{E}_c^2}{\partial \rho} \geq 0$, $\forall \rho$;
2. *When* $\rho \to 0$, *then* $\lim_{\rho \to 0} \left(\frac{\partial (E_c^2 - \widetilde{E}_c^2)}{\partial \rho} \right) = \frac{P_2}{2 \ln 2} \mathbb{E}[|h_2|^2]$
3. *When* $\rho >> 1$, *then* $\frac{\partial (E_c^2 - \widetilde{E}_c^2)}{\partial \rho} \approx -\frac{1}{2 \ln 2} \frac{1}{\rho} < 0$ *and it approaches 0 when* $\rho \to \infty$.

Proof: The proof is provided in Appendix IV.

Lemma 3 indicates that, for User 2, when the transmit SNR ρ is very small, the uplink EC with NOMA increases faster than that with OMA. On the other hand, when the transmit SNR is very large, the uplink EC with OMA increases faster than that with NOMA. Combining Lemma 3 and Lemma 1, we can conclude that, $E_c^2 - \widetilde{E}_c^2$ starts at an initial vanishingly small value, first increases, and subsequently decreases to $-\infty$ with a gradually diminishing rate. This means that for the stronger user, NOMA achieves higher EC than OMA at small values of the transmit SNR ρ. At high values of

ρ, OMA becomes more beneficial for the strong user. Finally, when $\rho \to \infty$ the performance gain of OMA over NOMA reaches a constant value, for the stronger user.

Finally, we investigate the sum ECs when using OMA and NOMA, denoted by V_N and V_O,

$$V_N = E_c^1 + E_c^2, \tag{4.12}$$
$$V_O = \tilde{E}_c^1 + \tilde{E}_c^2. \tag{4.13}$$

Our conclusions are drawn in Lemma 4.

Lemma 4: For the sum EC with NOMA, denoted by V_N, and with OMA, denoted by V_O, in a two-user uplink network, the following hold:

1. $\frac{\partial V_N}{\partial \rho} \geq 0$ *and* $\frac{\partial V_O}{\partial \rho} \geq 0$, $\forall \rho$;
2. *When* $\rho \to 0$, $V_N \to 0$, $\lim_{\rho \to 0}(\frac{\partial V_N}{\partial \rho}) = \frac{P_1}{\ln 2}\mathbb{E}[|h_1|^2] + \frac{P_2}{\ln 2}\mathbb{E}[|h_2|^2] \geq 0$, *and* $V_O \to 0$, $\lim_{\rho \to 0}(\frac{\partial V_O}{\partial \rho}) = \frac{P_1}{2\ln 2}\mathbb{E}[|h_1|^2] + \frac{P_2}{2\ln 2}\mathbb{E}[|h_2|^2] \geq 0$;
3. *When* $\rho >> 1$, $V_N \to \infty$, $\lim_{\rho \to \infty}(\frac{\partial V_N}{\partial \rho}) = 0$, *and* $V_O \to \infty$, $\lim_{\rho \to \infty}(\frac{\partial V_O}{\partial \rho}) = 0$.

The proof is provided in Appendix V.

Lemma 4 indicates that when NOMA is applied, the sum EC has a constant increasing rate at small value of the transmit SNR ρ that depends on the average of the channel power gains and the allocated power coefficients. A similar conclusion is reached when using OMA. On the other hand, when $\rho >> 1$, Lemma 4 indicates that the rate at which the sum ECs increase reaches a plateau, both in the case of NOMA and OMA.

4.4 Numerical Results

In this Section, the Lemmas presented in Section 4.3 are validated through Monte Carlo simulations. We consider a two user uplink NOMA system, with the following settings: normalized transmission power levels for both users, $P_1 = 0.2$, $P_2 = 0.8$, normalized delay exponent $\beta_1 = \beta_2 = -1$ for both users, unless otherwise stated.

In Fig. 4.1 the ECs of the two-user uplink NOMA and OMA networks are depicted versus the transmit SNR. We note that for the weak user, OMA is more advantageous than NOMA for low transmit SNRs, and NOMA is more advantageous than OMA at high transmit SNRs. Reverse conclusions can be drawn for the strong user. We notice also that the EC of the strong user converges at high SNRs. This provides numerical validation for Lemma 1.

Figs. 4.2 and 4.3, show respectively the EC of User 1 and User 2, versus the transmit SNR, for different values of $\beta_1 = \beta_2 = \beta$. When the delay constraints become more stringent, i.e., β decreases (equivalently, θ increases), the individual link-layer rates in NOMA decrease, for both users.

In Fig. 4.4, the ECs of the strong and weak users are depicted across different SNR values, $\rho \in \{1, 10, 30, 40, 50\}$ dB, as functions of the (negative) normalized delay exponent, for NOMA and OMA scenarios. We notice that the EC of each user is identical for NOMA and OMA, for small and large values of the normalized delay exponent. And with incresing transmit SNR ρ, the EC increases for both users.

Fig. 4.5 shows the difference of the EC in NOMA and the EC in OMA of the weak user. This curve starts initially at zero, then decreases to a certain minimum and starts increasing at the high values of transmit SNR. This confirms Lemma 2. When the delay is equal to -1, we see that for $\rho \in [0, 30]$ dB, the difference values are negative, indicating that OMA outperforms NOMA in this range. But when $\rho > 30$ dB, the values are positive, i.e., NOMA offers better link-layer rates. However, the particular ranges depend not only on the delay exponents but also on the power allocation coefficients. By increasing the transmission power of the weak user and reducing the transmission power of the

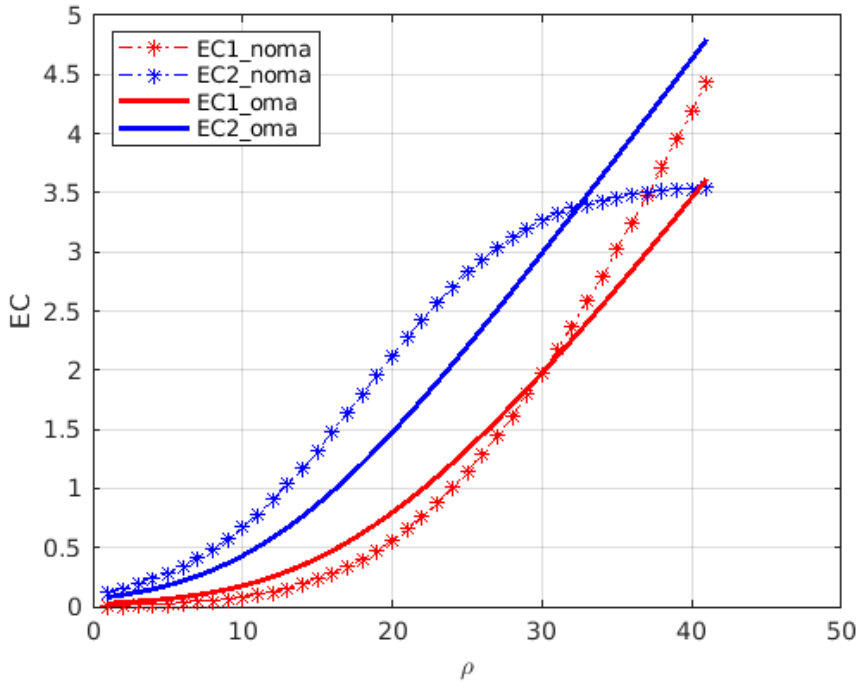


Figure 4.1: E_c^1, E_c^2in a two-user NOMA uplink network compared to Ecs of two users OMA, versus ρ

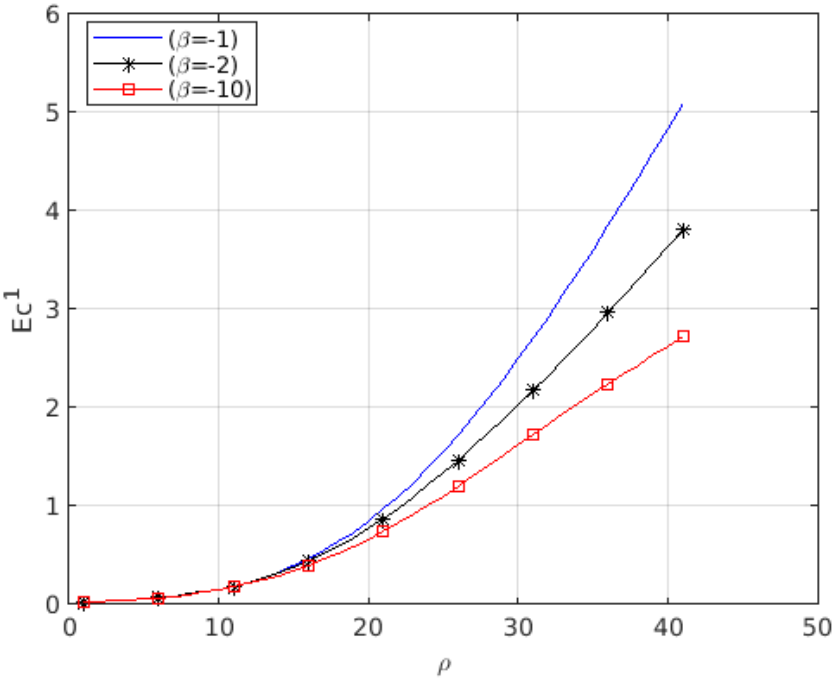


Figure 4.2: E_c^1 versus the transmit SNR, for several delays.

strong user, we notice that the range is reduced. That range expands when we do the inverse. Also, when the delay becomes more stringent, e.g., β_1=β_2=-2, the zero crossing moves from 30 to 36 dB.

Figure 4.6 shows the difference of the EC in NOMA and the EC in OMA for the strong user. This

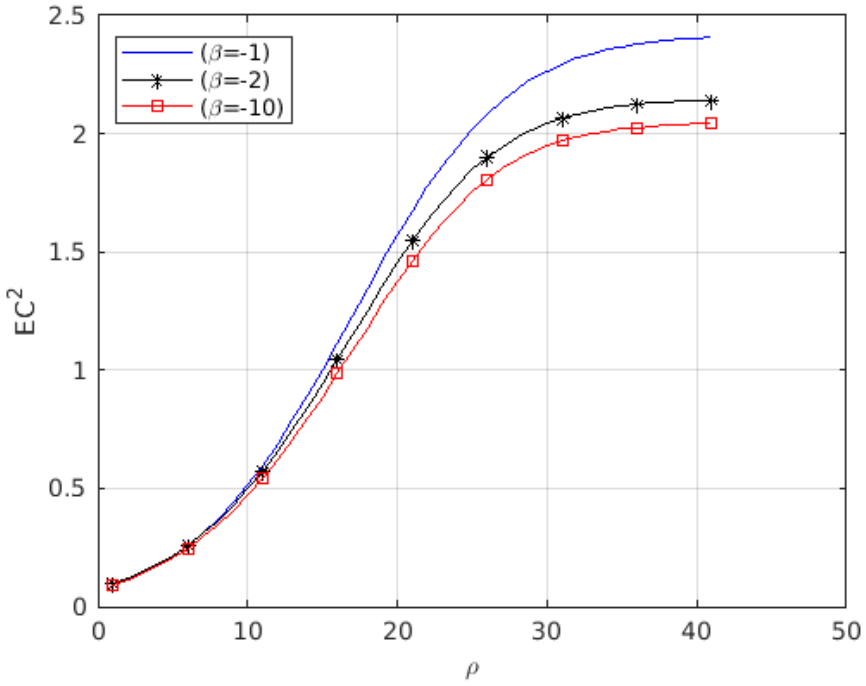


Figure 4.3: E_c^2 versus the transmit SNR ρ for several delays.

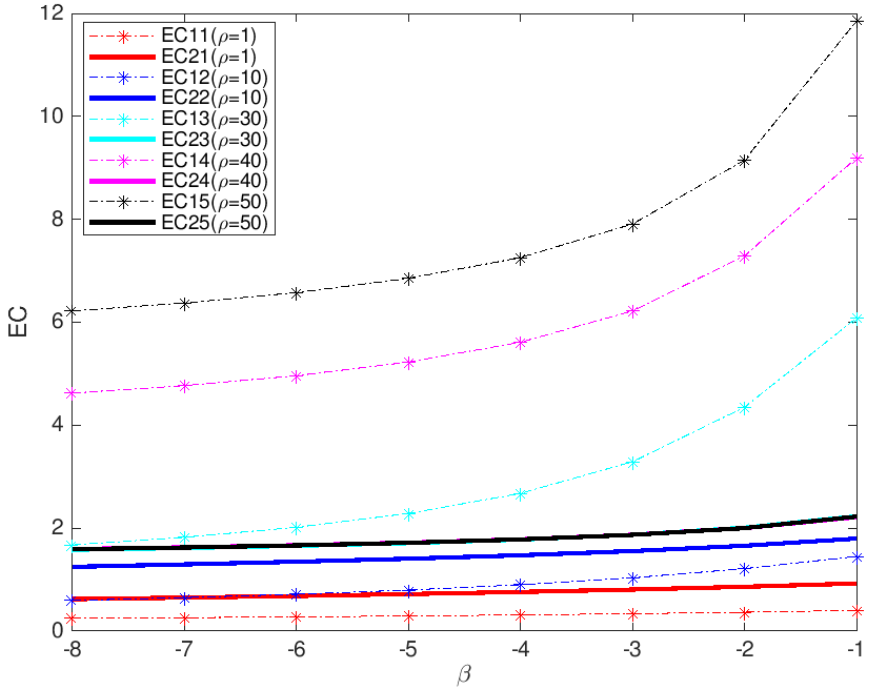


Figure 4.4: E_c^1 and E_c^2 in a two-user NOMA compared to ECs of two users OMA, versus normalized delay β, for different values of ρ.

curve starts initially at zero, then increases to a certain maximum and starts decreasing without bound at high values of the transmit SNR. This confirms Lemma 3. We note that the maximum of these curves decreases when the delay becomes more stringent.

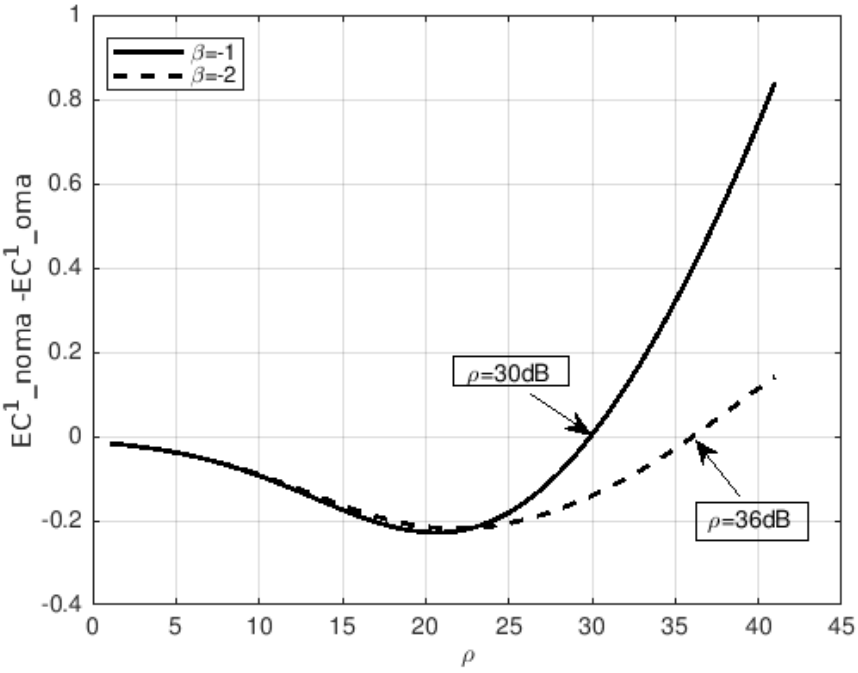


Figure 4.5: $E_c^1 - \tilde{E}_c^1$ versus ρ, for several values of the normalized delay exponent.

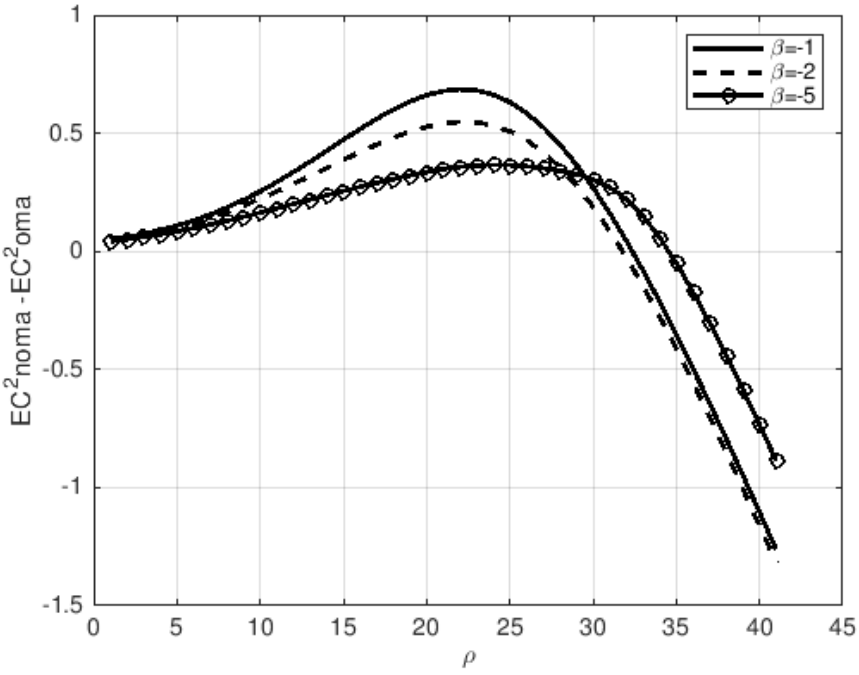


Figure 4.6: $E_c^2 - \tilde{E}_c^2$ versus ρ, for various normalized delay exponent.

To investigate the impact of ρ on the performance of the total link-layer rate for the two-user system, in Fig. 4.7 the plots for V_N in NOMA and V_O in OMA, versus the transmit SNR are depicted for various delay exponents. The curves demonstrate that for both NOMA and OMA, the total EC for

the two users starts at the initial value of 0 and then increases with the transmit SNR, as outlined in Lemma 4. When ρ is very small, the total link-layer rate for the two user in NOMA, V_N, increases faster than V_O in OMA. On the contrary, with the increase of the transmit SNR, V_O becomes gradually higher than V_N. At very high values of the transmit SNR, the gap between the sum EC with NOMA and OMA increases further. Finally, when the delay becomes more stringent, the sum EC of both NOMA and OMA decreases.

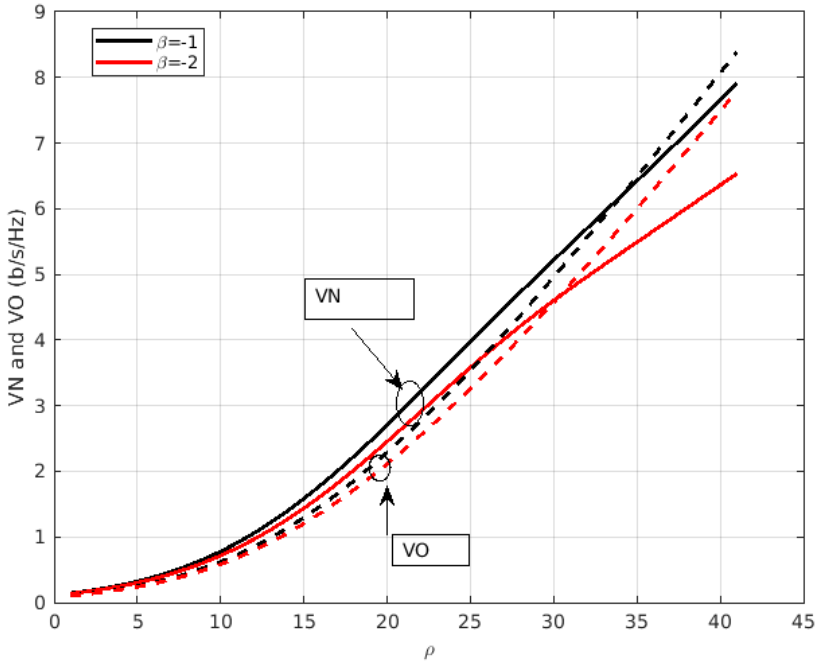


Figure 4.7: V_N and V_O versus ρ, for several values of normalized delay exponent.

Finally, Figs 4.8 and 4.9 depict the sum ECs versus ρ, for several values of the (negative) normalized delay exponent. In Fig. 4.8, the delay of the strong user is fixed, while the delay exponent of the weak user varies. It is shown that in this case, the highest delay QoS (i.e., the smallest negative normalized delay exponent) of the weak user corresponds to the highest gap between the sum ECs $V_N - V_O$. On the other hand, when the delay of the weak user is fixed, Fig. 4.9 shows that the smallest delay Qos (i.e., the highest negative normalized delay exponent) for the strong user corresponds to the largest gap in $V_N - V_O$.

The curve of $V_N - V_O$ starts at zero, increases to a maximum, and returns to negative values. The transition to zero is at $\rho = 31$, and ρ =36 respectively for the figures 4.8 and 4.9. That means from 0 to 31dB (36dB in the Figure 4.9), the total link-layer rate of NOMA is higher than the OMA one. And when ρ becomes larger than this transition point, the total link-layer rate of OMA outperforms the NOMA one.

4.5 Conclusions

The concept of the EC enabled us to study the achievable data-link layer rates when statistical delay QoS guarantees are in place, expressed in the form of delay exponents. We investigated the EC for the uplink of a two-user NOMA network, assuming a Rayleigh block fading channel. We derived novel closed-form expressions for the ECs of the two users and provided a comparison between NOMA and OMA. In NOMA networks, we showed that the ECs of both users decrease as the delay constraints

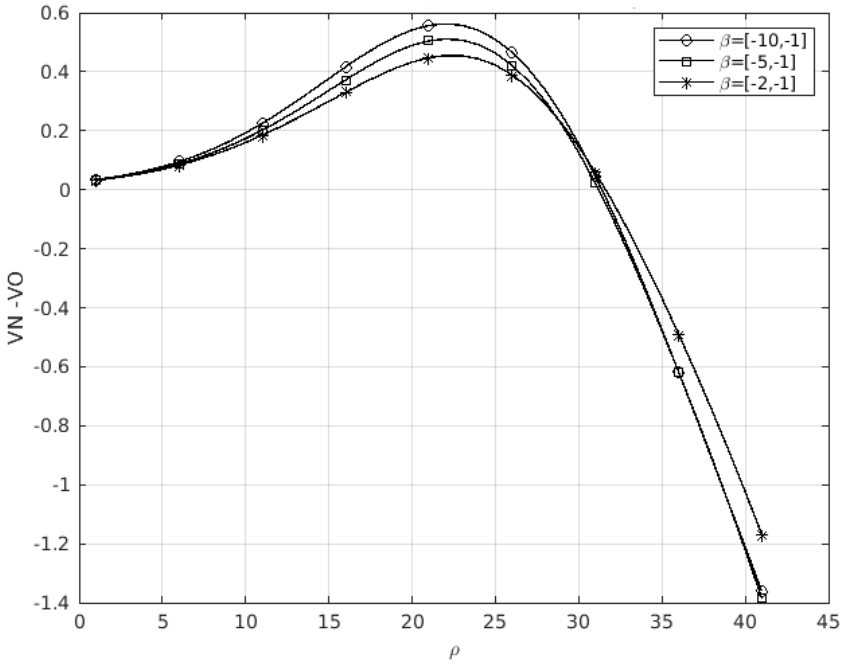


Figure 4.8: V_N - V_O versus ρ for various normalized delay.

become stringent. On the other hand, at high transmit SNRs, the EC of the weak user can surpass the EC of the strong user, as the latter is limited due to interference. This provides the possibility of switching between NOMA and OMA according to the individual users' delay constraints and transmit power levels.

Appendix I

For the second user, we have:

$$E_c^2 = \frac{1}{\beta_2}\log_2\left(2\int_0^\infty(\frac{\rho P_2}{1+\rho P_1 x_1})^{\beta_2}e^{-x_1}\int_{x_1}^\infty\Big(\frac{1+\rho P_1 x_1}{\rho P_2}+x_2\Big)^{\beta_2}e^{-x_2}dx_2dx_1\right).$$

Set $z=\frac{1+\rho P_1 x_1}{\rho P_2}+x_2$, which means we have $x_2=z-\frac{1+\rho P_1 x_1}{\rho P_2}$ and $dx_2=dz$. Then,

$$\begin{aligned}
E_c^2 &= \frac{1}{\beta_2}\log_2\left(2e^{\frac{1}{\rho P_2}}\int_0^\infty(\frac{\rho P_2}{1+\rho P_1 x_1})^{\beta_2}e^{-x_1}e^{\frac{P_1 x_1}{P_2}}\int_{\frac{1+\rho x_1}{\rho P_2}}^\infty z^{\beta_2}e^{-z}dzdx_1\right)\\
&= \frac{1}{\beta_2}\log_2\left(2(\rho P_2)^{\frac{\beta_2}{2}}e^{\frac{1}{2\rho P_2}}\int_0^\infty(1+\rho P_1 x_1)^{-\beta_2}(1+\rho x_1)^{\frac{\beta_2}{2}}e^{\frac{(2P_1-2P_2-1)x_1}{2P_2}}\Big[\mathbf{W}_{\frac{\beta_2}{2},\frac{1+\beta_2}{2}}(\frac{1+\rho x_1}{\rho P_2})\Big]dx_1\right)\\
&= \frac{1}{\beta_2}\log_2\left(2P_2(\rho P_2)^{\beta_2}e^{\frac{1}{\rho P_2}}e^{-\frac{(P_1-P_2)}{\rho P_2}}\int_{\frac{1}{\rho P_2}}^\infty P_2^{-\beta_2}(1+\rho P_1 y)^{-\beta_2}e^{(P_1-P_2)y}\Gamma(1+\beta_2,y)dy\right),
\end{aligned}$$

where $\mathbf{W}$ is the Whittaker W function.

Using the binomial expansion, we have $(1+\rho P_1 y)^{-\beta_2}=\sum_{j=0}^{-\beta_2}\binom{-\beta_2}{j}(\rho P_1 y)^j$ and we get the expression given in (4.9).

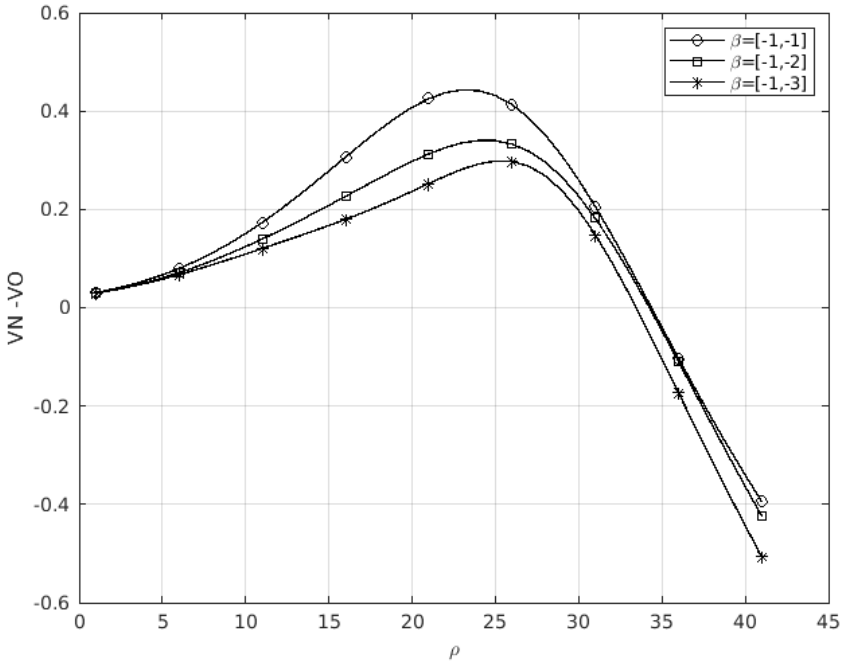


Figure 4.9: V_N - V_O versus ρ for various normalized delay.

The closed-form expression for the EC OMA, of the m-th user with M total users, is determined in [4] as follow.

$$\begin{aligned}\tilde{E}_c^m &= \frac{1}{\beta_m}\log_2\left(E\left[(1+\rho|h_m|^2)^{\frac{\beta_m}{2}}\right]\right) \\ &= \frac{1}{\beta_m}\log_2\left(\frac{\psi_m}{\rho}\int_0^\infty (1+\gamma_m)^{\beta_m} e^{-\frac{(M-m+1)\gamma_m}{\rho}}(1-e^{\frac{-\gamma_m}{\rho}})^{m-1}d\gamma_m\right) \\ &= \frac{1}{\beta_m}\log_2\left(\frac{\psi_m}{\rho}\sum_{k=0}^{m-1}\binom{m-1}{k}(-1)^k\int_0^\infty (1+\gamma_m)^{\beta_m} e^{-\frac{(M-m+1+k)\gamma_m}{\rho}}d\gamma_m\right).\end{aligned} \tag{4.14}$$

$$\tilde{E}_c^m = \frac{1}{\beta_m}\log_2\left(\frac{\psi_m}{\rho}\sum_{k=0}^{m-1}\binom{m-1}{k}(-1)^k U\left(1,2+\frac{2}{M}\beta_m,\frac{M-m+1+k}{\rho}\right)\right) \tag{4.15}$$

For the two users case, $M=2$, we have:

$$\widetilde{E}_c^1 = \frac{1}{\beta_1}\log_2\left(\frac{2}{\rho}\times U\left(1,2+\beta_1,\frac{2}{\rho}\right)\right) \tag{4.16}$$

and,

$$\widetilde{E}_c^2 = \frac{1}{\beta_2}\log_2\left(\frac{2}{\rho}\sum_{k=0}^{1}\binom{1}{k}(-1)^k\times U\left(1,2+\beta_2,\frac{1+k}{\rho}\right)\right) \tag{4.17}$$

where $U(\cdot,\cdot,\cdot)$ is the confluent hypergeometric function of the second kind, defined as

$$U(a,b,z) = \frac{1}{\Gamma(a)}\int_0^\infty e^{-zt}t^{a-1}(1+t)^{b-a-1}dt \tag{4.18}$$

Appendix II

By inserting $\rho \to 0$ into (4.8) and (4.9), we get 1) of Lemma 1, i.e.,

$$\lim_{\rho\to 0}(E_c^1 - \widetilde{E}_c^1) = \frac{1}{\beta_1}\log_2\left(\frac{\mathbb{E}[(1+\rho P_1|h_1|^2)^{\beta_2}]}{\mathbb{E}[(1+\rho|h_1|^2)^{\frac{\beta_2}{2}}]}\right) = 0,$$

$$\lim_{\rho\to 0}(E_c^2 - \widetilde{E}_c^2) = \frac{1}{\beta_2}\log_2\left(\frac{\mathbb{E}[(1+\frac{\rho P_2|h_2|^2}{1+\rho P_1|h_1|^2})^{\beta_2}]}{\mathbb{E}[(1+\rho|h_2|^2)^{\frac{\beta_2}{2}}]}\right) = 0.$$

In the same way, by inserting $\rho \to \infty$ into (4.8) and (4.9), we get 2) in Lemma 1, given below.

$$\lim_{\rho\to\infty} E_c^2 \to \frac{1}{\beta_2}\log_2\left(\mathbb{E}[(1+\frac{P_2|h_2|^2}{P_1|h_1|^2})^{\beta_2}]\right),$$

$$\lim_{\rho\to\infty}(E_c^1 - \widetilde{E}_c^1) = \frac{1}{\beta_1}\log_2\left(\rho^{\frac{\beta_1}{2}}\frac{\mathbb{E}[(\frac{1}{\rho}+P_1|h_1|^2)^{\beta_2}]}{\mathbb{E}[(\frac{1}{\rho}+|h_1|^2)^{\frac{\beta_2}{2}}]}\right) = \infty,$$

$$\lim_{\rho\to\infty}(E_c^2 - \widetilde{E}_c^2) = \frac{1}{\beta_2}\log_2\left(\frac{\mathbb{E}[(\frac{\frac{1}{\rho}+P_1|h_1|^2+P_2|h_2|^2}{\frac{1}{\rho}+P_1|h_1|^2})^{\beta_2}]}{\rho^{\frac{\beta_2}{2}}\mathbb{E}[(\frac{1}{\rho}+|h_2|^2)^{\frac{\beta_2}{2}}]}\right) = -\infty.$$

Appendix III

To analyze the trends of E_c^1 and $\widetilde{E}_c^1$ with respect to ρ, we start with

$$\frac{\partial E_c^1}{\partial\rho} = \frac{1}{\beta_1\ln 2}\frac{\left(\mathbb{E}[(1+\rho P_1|h_1|^2)^{\beta_1}]\right)'}{\mathbb{E}[(1+\rho P_1|h_1|^2)^{\beta_1}]} = \frac{P_1}{\ln 2}\frac{\mathbb{E}[|h_1|^2(1+\rho P_1|h_1|^2)^{\beta_1-1}]}{\mathbb{E}[(1+\rho P_1|h_1|^2)^{\beta_1}]} \geq 0.$$

Similarly, for user 1 in OMA we have

$$\frac{\partial\widetilde{E}_c^1}{\partial\rho} = \frac{1}{\beta_1\ln 2}\frac{\left(\mathbb{E}[(1+\rho|h_1|^2)^{\frac{\beta_1}{2}}]\right)'}{\mathbb{E}[(1+\rho|h_1|^2)^{\frac{\beta_1}{2}}]} = \frac{1}{2\ln 2}\frac{\mathbb{E}[|h_1|^2(1+\rho|h_1|^2)^{\frac{\beta_1}{2}-1}]}{\mathbb{E}[(1+\rho|h_1|^2)^{\frac{\beta_1}{2}}]} \geq 0.$$

Then, we get that

$$\frac{\partial(E_c^1 - \widetilde{E}_c^1)}{\partial\rho} = \frac{P_1}{\ln 2}\frac{\mathbb{E}[|h_1|^2(1+\rho P_1|h_1|^2)^{\beta_1-1}]}{\mathbb{E}[(1+\rho P_1|h_1|^2)^{\beta_1}]} - \frac{1}{2\ln 2}\frac{\mathbb{E}[|h_1|^2(1+\rho|h_1|^2)^{\frac{\beta_1}{2}-1}]}{\mathbb{E}[(1+\rho|h_1|^2)^{\frac{\beta_1}{2}}]}, \tag{4.19}$$

and $\lim_{\rho\to 0}(\frac{\partial(E_c^1-\widetilde{E}_c^1)}{\partial\rho}) = \frac{(P_1-\frac{1}{2})}{\ln 2}\mathbb{E}[|h_1|^2] \leq 0$. When $\rho >> 1$, we have

$$\frac{\partial(E_c^1 - \widetilde{E}_c^1)}{\partial\rho}) = \frac{P_1}{\ln 2}\frac{\mathbb{E}[|h_1|^2(\rho P_1|h_1|^2)^{\beta_1-1}]}{\mathbb{E}[(\rho P_1|h_1|^2)^{\beta_1}]} - \frac{1}{2\ln 2}\frac{\mathbb{E}[|h_1|^2(\rho|h_1|^2)^{\frac{\beta_1}{2}-1}]}{\mathbb{E}[(\rho|h_1|^2)^{\frac{\beta_1}{2}}]} = \frac{1}{2\rho\ln 2} \geq 0. \tag{4.20}$$

When $\rho \to \infty$, this term approaches 0.

Appendix IV

$E_c^2 = \frac{1}{\beta_2}\log_2(\mathbb{E}[(1+\frac{\rho P_2|h_2|^2}{1+\rho P_1|h_1|^2})^{\beta_2}])$, and

$$\frac{\partial E_c^2}{\partial \rho} = \frac{1}{\beta_2 \ln 2}\frac{\left(\mathbb{E}[(1+\frac{\rho P_2|h_2|^2}{1+\rho P_1|h_1|^2})^{\beta_2}]\right)'}{\mathbb{E}[(1+\frac{\rho P_2|h_2|^2}{1+\rho P_1|h_1|^2})^{\beta_2}]} = \frac{1}{\ln 2}\frac{\mathbb{E}[\frac{P_2|h_2|^2}{(1+\rho P_1|h_1|^2)^2}(1+\frac{\rho P_2|h_2|^2}{1+\rho P_1|h_1|^2})^{\beta_2-1}]}{\mathbb{E}[(1+\frac{\rho P_2|h_2|^2}{1+\rho P_1|h_1|^2})^{\beta_2}]} \geq 0. \tag{4.21}$$

In the same way, for the user 2 in OMA, we have

$$\frac{\partial \widetilde{E}_c^2}{\partial \rho} = \frac{1}{\beta_2 \ln 2}\frac{\left(\mathbb{E}[(1+\rho|h_2|^2)^{\frac{\beta_2}{2}}]\right)'}{\mathbb{E}[(1+\rho|h_2|^2)^{\frac{\beta_2}{2}}]} = \frac{1}{2\ln 2}\frac{\mathbb{E}[|h_2|^2(1+\rho|h_2|^2)^{\frac{\beta_2}{2}-1}]}{\mathbb{E}[(1+\rho|h_2|^2)^{\frac{\beta_2}{2}}]} \geq 0, \tag{4.22}$$

and

$$\frac{\partial (E_c^2-\widetilde{E}_c^2)}{\partial \rho} = \frac{1}{\ln 2}\frac{\mathbb{E}[\frac{P_2|h_2|^2}{(1+\rho P_1|h_1|^2)^2}(1+\frac{\rho P_2|h_2|^2}{1+\rho P_1|h_1|^2})^{\beta_2-1}]}{\mathbb{E}[(1+\frac{\rho P_2|h_2|^2}{1+\rho P_1|h_1|^2})^{\beta_2}]} - \frac{1}{2\ln 2}\frac{\mathbb{E}[|h_2|^2(1+\rho|h_2|^2)^{\frac{\beta_2}{2}-1}]}{\mathbb{E}[(1+\rho|h_2|^2)^{\frac{\beta_2}{2}}]}. \tag{4.23}$$

When $\rho \to 0$, we have that $\lim_{\rho\to 0}(\frac{\partial(E_c^2-\widetilde{E}_c^2)}{\partial\rho}) = \frac{(P_2-\frac{1}{2})}{\ln 2}\mathbb{E}[|h_2|^2]$. When ρ is very large,

$$\begin{aligned}
\frac{\partial (E_c^2-\widetilde{E}_c^2)}{\partial \rho} &= \frac{\mathbb{E}[\frac{P_2|h_2|^2}{\rho^2(\frac{1}{\rho}+P_1|h_1|^2)^2}(1+\frac{\rho}{\rho}\frac{(P_2|h_2|^2)}{(\frac{1}{\rho}+P_1|h_1|^2)})^{\beta_2-1}]}{\ln 2\mathbb{E}[(1+\frac{\rho}{\rho}\frac{P_2|h_2|^2}{(\frac{1}{\rho}+P_1|h_1|^2)})^{\beta_2}]} - \frac{1}{2\ln 2}\frac{1}{\rho}\frac{\mathbb{E}[|h_2|^2(\frac{1}{\rho}+|h_2|^2)^{\frac{\beta_2}{2}-1}]}{\mathbb{E}[(\frac{1}{\rho}+|h_2|^2)^{\frac{\beta_2}{2}}]} \\
&= \frac{\mathbb{E}[\frac{P_2|h_2|^2}{\rho^2(P_1|h_1|^2)^2}(1+\frac{P_2|h_2|^2}{P_1|h_1|^2})^{\beta_2-1}]}{\ln 2\mathbb{E}[(1+\frac{P_2|h_2|^2}{P_1|h_1|^2})^{\beta_2}]} - \frac{1}{2\ln 2}\frac{1}{\rho}\frac{\mathbb{E}[(|h_2|^2)^{\frac{\beta_2}{2}}]}{\mathbb{E}[(|h_2|^2)^{\frac{\beta_2}{2}}]} \\
&= \frac{P_2}{\rho^2 P_1^2}\frac{\mathbb{E}[\frac{|h_2|^2}{(|h_1|^2)^2}(1+\frac{P_2|h_2|^2}{P_1|h_1|^2})^{\beta_2-1}]}{\ln 2\mathbb{E}[(1+\frac{P_2|h_2|^2}{P_1|h_1|^2})^{\beta_2}]} - \frac{1}{2\ln 2}\frac{1}{\rho} = \frac{\frac{P_2}{P_1^2\ln 2}A - \frac{1}{2\ln 2}\rho}{\rho^2},
\end{aligned} \tag{4.24}$$

where $A = \frac{\mathbb{E}[\frac{|h_2|^2}{(|h_1|^2)^2}(1+\frac{P_2|h_2|^2}{P_1|h_1|^2})^{\beta_2-1}]}{\mathbb{E}[(1+\frac{P_2|h_2|^2}{P_1|h_1|^2})^{\beta_2}]}$, unrelated to ρ. Hence, when ρ is very large, $\frac{\partial(E_c^2-\widetilde{E}_c^2)}{\partial\rho}$ can be approximated by $-\frac{1}{2\ln 2}\frac{1}{\rho}$, and it gradually approaches 0 when $\rho \to \infty$.

Appendix V

Note that $V_N = E_c^1 + E_c^2$. By using Lemma 1, we have $\lim_{\rho\to 0}(V_N) = 0$ and $\lim_{\rho\to\infty}(V_N) = \infty$. Then, we get that

$$\frac{\partial V_N}{\partial \rho} = \frac{\partial(E_c^1+E_c^2)}{\partial\rho} = \frac{P_1}{\ln 2}\frac{\mathbb{E}[|h_1|^2(1+\rho P_1|h_1|^2)^{\beta_1-1}]}{\mathbb{E}[(1+\rho P_1|h_1|^2)^{\beta_1}]} + \frac{1}{\ln 2}\frac{\mathbb{E}[\frac{P_2|h_2|^2}{(1+\rho P_1|h_1|^2)^2}(1+\frac{\rho P_2|h_2|^2}{1+\rho P_1|h_1|^2})^{\beta_2-1}]}{\mathbb{E}[(1+\frac{\rho P_2|h_2|^2}{1+\rho P_1|h_1|^2})^{\beta_2}]} \geq 0. \tag{4.25}$$

When $\rho \to 0$, we have $\lim_{\rho\to 0}(\frac{\partial V_N}{\partial \rho}) = \frac{P_1}{\ln 2}\mathbb{E}[|h_1|^2] + \frac{P_2}{\ln 2}\mathbb{E}[|h_2|^2]$. When $\rho \to \infty$, we get that

$$\lim_{\rho\to\infty} \frac{\partial V_N}{\partial \rho} = \frac{1}{\rho \ln 2} + \frac{\mathbb{E}[\frac{P_2|h_2|^2}{(P_1|h_1|^2)^2}(1+\frac{P_2|h_2|^2}{P_1|h_1|^2})^{\beta_2-1}]}{\rho^2 \ln 2 \mathbb{E}[(1+\frac{P_2|h_2|^2}{P_1|h_1|^2})^{\beta_2}]} = 0.$$

For V_O in the case of OMA, we note that $V_O = \widetilde{E}_c^1 + \widetilde{E}_c^2$. By using Lemma 1, we have $\lim_{\rho\to 0}(V_0) = 0$ and $\lim_{\rho\to\infty}(V_0) = \infty$. Then,

$$\frac{\partial V_0}{\partial \rho} = \frac{\partial(\widetilde{E}_c^1 + \widetilde{E}_c^2)}{\partial \rho} = \frac{1}{2\ln 2}\frac{\mathbb{E}[|h_1|^2(1+\rho|h_1|^2)^{\frac{\beta_1}{2}-1}]}{\mathbb{E}[(1+\rho|h_1|^2)^{\frac{\beta_1}{2}}]} + \frac{1}{2\ln 2}\frac{\mathbb{E}[|h_2|^2(1+\rho|h_2|^2)^{\frac{\beta_2}{2}-1}]}{\mathbb{E}[(1+\rho|h_2|^2)^{\frac{\beta_2}{2}}]} \geq 0.$$

When $\rho \to 0$, we have $\lim_{\rho\to 0}(\frac{\partial V_O}{\partial \rho}) = \frac{1}{2\ln 2}\mathbb{E}[|h_1|^2] + \frac{1}{2\ln 2}\mathbb{E}[|h_2|^2]$. When $\rho \to \infty$, we have that $\lim_{\rho\to\infty}(\frac{\partial V_O}{\partial \rho}) = \lim_{\rho\to\infty}(\frac{1}{2\rho\ln 2} + \frac{1}{2\rho\ln 2}) = \lim_{\rho\to\infty}(\frac{1}{\rho\ln 2})$, which equals to 0.

References

[1] SM Riazul Islam, , Nurilla Avazov, Octavia A Dobre, and Kyung-Sup Kwak. Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges. *IEEE Commun. Surv, & Tut.*, 19(2):721–742, 2016.

[2] Y. Kanaras and I. Darwazeh A. Chorti, M. Rodrigues. An optimum detection for a spectrally efficient non orthogonal FDM system. In *Proc. 13th Int. OFDM WS*, pages 65–68, Aug 2008.

[3] Yuya Saito, Yoshihisa Kishiyama, Anass Benjebbour, Takehiro Nakamura, Anxin Li, and Kenichi Higuchi. Non-orthogonal multiple access (NOMA) for cellular future radio access. In *Proc. IEEE VTC Spring*, pages 1–5, 2013.

[4] Wenjuan Yu, Leila Musavian, and Qiang Ni. Link-layer capacity of NOMA under statistical delay QoS guarantees. *IEEE Trans. Commun.*, 66(10):4907–4922, 2018.

[5] Wenjuan Yu, Leila Musavian, and Qiang Ni. Tradeoff analysis and joint optimization of link-layer energy efficiency and effective capacity toward green communications. *IEEE Trans. Wireless Commun.*, 15(5):3339–3353, 2016.

[6] Dapeng Wu and Rohit Negi. Effective capacity: a wireless link model for support of quality of service. *IEEE Trans. Wireless Commun.*, 2(4):630–643, 2003.

[7] Jia Tang and Xi Zhang. Cross-layer modeling for quality of service guarantees over wireless links. *IEEE Trans. Wireless Commun.*, 6(12):4504–4512, 2007.

[8] A. Chorti and H. V. Poor. Achievable secrecy rates in physical layer secure systems with a helping interferer. In *2012 Int Conf. Computing, Networking Commun. (ICNC)*, pages 18–22, Maui, HI, 2012.

[9] W. Yu, A. Chorti, L. Musavian, H. Vincent Poor, and Q. Ni. Effective secrecy rate for a downlink NOMA network. *IEEE Trans. Wireless Commun.*, pages 5673–5690, Dec. 2019.

[10] Zheng Yang, Zhiguo Ding, Pingzhi Fan, and Naofal Al-Dhahir. A general power allocation scheme to guarantee quality of service in downlink and uplink noma systems. *IEEE Transactions on Wireless Communications*, 15(11):7244–7257, 2016.

[11] Ningbo Zhang, Jing Wang, Guixia Kang, and Yang Liu. Uplink nonorthogonal multiple access in 5G systems. *IEEE Commun. L.*, 20(3):458–461, 2016.

[12] Li Fan, Shi Jin, Chao-Kai Wen, and Haixia Zhang. Uplink achievable rate for massive MIMO systems with low-resolution ADC. *IEEE Commun. L.*, 19(12):2186–2189, 2015.

[13] Hong-Chuan Yang and Mohamed-Slim Alouini. *Order statistics in wireless communications: diversity, adaptation, and scheduling in MIMO and OFDM systems*. Cambridge University Press, 2011.

Chapter 5

Perspectives

5.1 Introduction

We are currently witnessing a tremendous increase in the volume of Internet of Things (IoT) generated data on one hand, and, the emergence of delay critical use cases on the other. These factors have played a major role towards the processing of the respective data flows close to their sources as opposed to in core Clouds, giving rise to Edge/Fog computing. However, in this context the question of what is the "optimal" point in the end to end (E2E) path for processing to occur, has not yet been systematically addressed. As a rule of thumb, it is conjectured that strict real-time applications require computation to be executed in the vicinity of the data source while tasks such as big data analytics are considered more appropriate to run on the Cloud. A further question down the path, concerns securing connections to Edge nodes while offloading, particularly when heterogeneous systems are to be interconnected.

In parallel, proposals for the flexible allocation of the infrastructure resources under the umbrella of network slicing, build on two key technologies, namely software defined networking (SDN) and network function virtualization (NFV) [1]. Due to the complexity of the problem at hand, the new network *management layer* is typically thought of as machine learning (ML) unit enabling multi-domain orchestration [2], [3]. As we move gradually away from the standard client-server networking paradigm and enter a new era of truly E2E quality of service (QoS), service level agreements (SLAs) in the near future will be expected to include guarantees about the *quality of security* (QoSec) [4].

Meanwhile, in the radio community a different discussion is currently taking place around the 1949 paper of Warren Weaver [5] on the premise that in sixth generation (6G) systems it will be possible to move towards semantic communications, i.e., conveying reliably the "meaning" of the messages, rather than simply conveying reliably the data that carry the messages. Early approaches in the direction of defining a formal framework for semantic communications have appeared [6], in essence extending the standard Shannon model of reliable communication of binary sequences to reliable communication of semantic messages. Through the use of propositional logic, the concept of semantic entropy was introduced and employed in semantic source and channel coding statements. In other communities, notably in natural language processing, moving from a semantic to a computational representation of meaning was undertaken with artificial intelligence (AI) tools with great success. It therefore only seems natural to invest on AI to interpret semantics in 6G, while the interplay between information theory and machine learning has recently attracted attention [7].

On the other hand, the intelligence of any actual system in the physical world sits on, and depends upon, its physical infrastructure.[1] The author of this manuscripts posits that PHY layer properties, e.g., the direction of arrival of a beam, the location of a node, the time of communication, the ambient temperature, etc., carry important *contextual* information, directly related to semantics. A consequential question would be then how to incorporate in the definition of context "deeper" properties of what constitutes the "PHY signal / system", for example:

[1] An ancient Greek proverb states that "a healthy mind (lives) in a healthy body".

- Deterministic / stochastic aspects of source processes and evolutionary dynamics;
- Nonlinear behavior of sources, sinks, transceivers and media;
- Memory depth of the individual segments of the network and overall memory of the system;
- Scale (size) of the network and interconnectivity properties, e.g., well connected / self similar / scale-free networks;

The list can grow...

The implications of the previous discussion are multi-fold and relate to important open research problems, for example: i) the ML orchestration unit emerges as a preferential point of attack in the network slicing era; ii) the standard "on or off" security paradigm in which security protocols provide hard guarantees of authentication, confidentiality, integrity, non-repudiation, etc., for the client-server exchanges might not be well aligned with QoSec; iii) accounting for the context of communication carries important semantic information; iv) to truly enable E2E QoS the E2E latency, comprising both communication and processing delays need to be mutually optimized.

In the following Sections my short and long term research directions, motivated by the previous points, will be presented, starting from wireless security and moving on to 6G wireless.

5.2 The Role of of PLS in 6G

In the context of security, unarguably, 5G security enhancements present a big improvement with respect to long term evolution (LTE). However, as the complexity of the application scenarios increases with the introduction of novel use cases, notably ultra-reliable low latency (URLLC) and massive machine type communications (mMTC), novel security challenges arise that might be difficult to address using the standard paradigm of complexity based classical crypto solutions, particularly because of low latency constraints. At the same time, in the longer 10-year horizon novel security concepts based on "trust models" and *risk-based, adaptive identity management and access control* will come to life, enabled to a large extend by AI. Early elements of these directions are seen in standards used for digital (monetary) transactions, e.g., the Internet engineering task force (IETF) request for comments (RFC) *Vectors of Trust* [8]. Furthermore, to allow for flexible QoSec in the differentiated services framework, the development and integration of security controls at *all* layers of the communications system is envisioned [9].

Currently, physical layer security (PLS) is being considered as a possible way to emancipate networks from classical, complexity based, security approaches. Since the wireless channel is reciprocal, time-varying and random in nature, it offers a valid, inherently secure source for key agreement (KA) protocols between two communicating parties. This is pertinent to many forthcoming B5G applications that will require strong, but nevertheless, lightweight KA mechanisms; in this direction, PLS may offer such solutions, or complement existing algorithms, with minimal changes in the control plane. With respect to authentication, physical unclonable functions (PUFs), wireless fingerprinting / localization, combined with more classical approaches, could also enhance authentication and key agreement (AKA) in demanding scenarios, including (but not limited to) device to device (D2D) and Industry 4.0. In parallel, mmWave in the Terahertz range will rely upon setting up invisible radio "wires"; although on their own they cannot ensure confidentiality, they will provide a concrete scenario for the wiretap channel. It is therefore pertinent to discuss advancements in wiretap secrecy encoders.

5.2.1 How Many Secret Bits Are Needed

A core question that needs to be addressed first in all of the above scenarios is "how many secret bits" we need to reach security levels that could be considered acceptable. If we envision hybrid PLS-crypto systems that leverage symmetric block ciphers and the employment of PLS serves primarily

for generating and communicating authentication and encryption keys, then the secret bit rates can be made very low. As an example, 256 authentication or encryption key bits suffice for the encryption of at least 1 gigabyte (GB) of data; if the keys are exchanged using secrecy encoders, this would correspond to a required secrecy rate of $\tau = 3.2 \cdot 10^{-8}$ bits/sec/Hz. For a "fair" comparison we can set the probability of secrecy outage to $P_{out} = 2^{-80}$, which is the practically tolerable value for the "advantage" gained by an adversary in semantic security proofs (chosen plaintext and chosen ciphertext proofs – CPA and CCA security). A quick calculation using these numbers in a basic single input single output (SISO) block fading channel with a single eavesdropper [10] reveals that we would need 44 transmission symbols if causal CSI was available along with a fixed power budget over a time horizon; in 5G this translates to less than 4 consecutive transmission time intervals (TTIs) of a total time duration of less than 2 msec. Further investigations assuming semi-blind scenarios in which only the legitimate users' CSI is available, extension to MIMO, looking at larger values of the probability of secrecy outage, etc., would be needed for a systematic study of the scenarios in which secrecy encoders could be applicable. Our initial conclusion is that hybrid systems show promise and could fit excellently in a framework of flexible QoSec with differing levels of secrecy guarantees.

A further motivation in exploring PLS solutions stems from the fact that a number of vulnerabilities, e.g., in the false base station attack [11] or due to jamming during the beam allocation in mmWave [12] arise during the establishment of the radio link; in this aspect, standard security protocols that *build on the premise that the communication link has already been established*, cannot offer solutions when this is not the case, whereas, PLS schemes can be seamlessly incorporated (e.g., can be interwoven with channel estimation).

In a nutshell, several advantages can be envisioned by rethinking the security design bottom-up, and in particular: 1) PLS can provide information-theoretic security guarantees with lightweight mechanisms (e.g., using LDPC encoders); 2) hybrid crypto-PLS protocols can provide alternatives in scenarios where classical mechanisms fall short and naturally lend themselves for the design of adaptive QoSec protocols, and, 3) PLS can act as an extra security layer, complementing other approaches.

My proposed research on PLS encompasses research on authentication, data confidentiality and anomaly detection, described below.

5.2.2 Authentication

With respect to authentication, my current research includes multi-factor AKA with PUFs, wireless fingerprinting and localization, while I also work on Slepian Wolf and Wyner Ziv PUF and SKG decoders in the short block-length.

My future research directions include:

- Characterization of the wireless channel from a security (as opposed to reliability) point of view. The baseline idea boils down to the fact that the *predictable* element of the wireless coefficient would be useful for authentication (e.g., through localization), while the purely random for extracting secret keys. Developing the mathematical and ML tools to resolve the two is an unchartered area of research so far;

- The proposal of hybrid crypto-PLS systems and the resilience of related systems to active attacks. Unlike in the network security paradigm, active attacks at PHY can be alleviated by PHY remedies, e.g., narrow beamforming, pilot randomization, energy harvesting and band hopping leveraging OFDM transmitters.

5.2.3 Data Confidentiality

Additionally, with respect to data confidentiality achieved in wiretap channels with the use of secrecy encoders, practical designs have so far been presented only for the wiretap-II channel (i.e., noiseless main channel) and the erasure channel. Building secrecy encoders for the standard wiretap-I channel is timely,

especially as degradedness of the eavesdropping channel can be substantiated in mmWave technologies enabled by narrow beamforming using multiple antennas. In my proposed research direction in this domain, I intend to seek input from the design of core building blocks of symmetric block ciphers, in particular of reversible S-boxes. While linear encoders purposed for reliability have proved instrumental to reach the Shannon limit for reliable communication in (linear) channel wireless settings, they might not be the optimal choice for secrecy. The study of bent functions [13] can constitute the starting point of the design of non-linear secrecy encoders, purposed to guarantee reliability in the communication and secrecy with respect to an eavesdropper.

Finally, the evaluation of the security level achieved with the proposed PLS methods will be sought, scrutinizing the hypothesis that security level 5 (post-quantum) is attainable with PLS in the finite block-length regime. In particular through a systematic analysis of the probability of secrecy outage, I will seek to provide quantitative measures regarding the security level achievable.

5.2.4 Anomaly Detection

In terms of anomaly detection, my focus is on

- Identifying attacks on the ML orchestrator;
- Proposing novel approaches on identifying hacking at the device level in IoT networks.

I will study whether such intrusions can be "inferred" by observing side channels. Side channels have customarily been used for negative security proofs, e.g., showcasing it is possible to compromise symmetric block ciphers with smart power monitoring. The idea of using them to identify intrusions or anomalous events breaks completely new ground. A second important aspect is the development of distributed anomaly detection algorithms. Understanding the trade-off between cluster size and speed of detection would be the primary initial goal in this setting.

5.3 Low Latency, Interference-free, Contextual 6G Communications

Three further topics of primary focus for me are articulated around:

- Interference management;
- E2E delay constraints;
- Context awareness.

With respect to the former two, I am interested in studying the interaction of two key technologies: the combination of random access (RA), e.g., slotted Aloha, with uplink NOMA on one hand and the limits of interference cancellation using the baseline ideas of generalised frequency division multiplexing (GFDM) [14]. With respect to the latter, the employment of advanced ML algorithms will be investigated.

5.3.1 NOMA for Collision Avoidance in mMTC Uplink

With respect to the first aspect, as an alternative to grant free access, it is possible to envision the combination of slotted Aloha with NOMA technologies. Let's us take the narrow-band IoT (NB-IoT) as an example. A contention-based RA procedure is performed for initial uplink grant which includes a four-message handshake between the IoT devices and the gNodeB (gNB). First, the IoT device transmits a preamble to the gNB on the Narrowband Physical RA CHannel (NPRACH). The preamble is composed of four symbol groups, each transmitted on a different subcarrier. If two or more IoT devices randomly choose the same initial subcarrier, the preamble sequence will collide, albeit the gNB will not notice the collision and the corresponding devices will receive the same RA response

(RAR) message containing the uplink resource grant and synchronization information. This causes the message transmission collision to take place only in step 3, causing further delays.

It is expected that preamble collision will occur more frequently in mMTC scenarios, causing severe network congestion and long access delay, because of the subsequent backoff. NOMA can be applied to enhance the conventional RA procedure, with early proposals in this direction recently appearing [15]. In these works, a set of pre-determined power levels is set and the NOMA-based RA is employed by allowing the IoT devices to randomly access one channel with a randomly chosen power level. In my future studies I am interested in applying this concept in Rayleigh environments and using the theory of order statistics to allow each device "self-evaluate" it's relative ranking in terms of channel quality and make an informed guess on the power level that should be used. Interesting further extensions would account for the impact of short packet transmissions [16] and MAC sub-layer QoS delay constraints [17].

5.3.2 Interference Cancellation Using Machine Learning

I will build on my previous research on a particular instance of GFDM in the frequency domain, referred to as faster than Nyquist (FTN) signalling [18]. In FTN, the orthogonality of the OFDM subcarriers is intentionally violated through the use of inverse fast fractional Fourier transform at the transmitter and fast fractional Fourier transform at the receiver. Our results so far demonstrate that it is possible to employ ML to detect particular sets of FTN signals. The trade-off between the increase in the number of subcarriers and the compression of the intercarrier spacing is captured in the fact that an increasing number of the system Gram matrice's singular values become vanishingly small. I am interested in transferring the lessons learned from this exercise to building similar detection schemes in the presence of high interference levels. The goal will be to investigate the limits of using ML for interference cancellation in partially overlapped transmissions from different users, e.g., due to collisions in mMTC using standard Aloha.

5.3.3 Towards Context Aware Communications in 6G

Finally, in the general context of 6G, understanding of the semantics is expected to become increasingly important. In this aspect, capturing different aspects of PHY signals' properties related to semantic content, e.g., as discussed in Section 5.1, could be envisioned through the use of ML. As a first approach to incorporate "logarithmic" type of memory (e.g., as exists in human languages), I will explore the use of *dilated* convolutional neural networks (CNNs) that apply recursively fractional weights to deeper layers. Furthermore, the use of generative adversarial networks (GANs) will be explored in the general context of learning the appropriate ML architecture.

Moreover, the opening up of the THz spectrum will provide new "sensing" capabilities to 6G devices, such as high definition imaging and frequency spectroscopy. In combination with decimeter precision localization, as showcased recently for mmWave systems [19], these enhanced sensing capabilities can prove instrumental in understanding context and could naturally be incorporating in trust building and predicting reliability. Enhancements by aggregating inertial movement unit (IMU) calibration, accelerometer readings with simultaneous localization and mapping (SLAM) as already proposed [20] can enhance substantially the positioning accuracy and reliability.

The combination of the above mentioned enhanced sensing and positioning traits of 6G devices and different types of system memory can pave the way towards semantic communications, not accounting exclusively for the literal meaning of the exchanged information (typically at the application layer), but also, for the context in which it was produced, transported and received.

References

[1] J. Gil Herrera and J. F. Botero. Resource allocation in nfv: A comprehensive survey. *IEEE Transactions on Network and Service Management*, 13(3):518–532, 2016.

[2] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella. On multi-access edge computing: A survey of the emerging 5g network edge cloud architecture and orchestration. *IEEE Communications Surveys Tutorials*, 19(3):1657–1681, 2017.

[3] V. K. Rathi, V. Chaudhary, N. K. Rajput, B. Ahuja, A. K. Jaiswal, D. Gupta, M. Elhoseny, and M. Hammoudeh. A blockchain-enabled multi domain edge computing orchestrator. *IEEE Internet of Things Magazine*, 3(2):30–36, 2020.

[4] Z. M. Fadlullah, C. Wei, Z. Shi, and N. Kato. Gt-qosec: A game-theoretic joint optimization of qos and security for differentiated services in next generation heterogeneous networks. *IEEE Transactions on Wireless Communications*, 16(2):1037–1050, 2017.

[5] W. Weaver. The mathematical theory of communication. *Scientific American*, 1949.

[6] J. Bao, P. Basu, M. Dean, C. Partridge, A. Swami, W. Leland, and J. A. Hendler. Towards a theory of semantic communication. In *2011 IEEE Network Science Workshop*, pages 110–117, 2011.

[7] A. R. Heravi and G. A. Hodtani. A new information theoretic relation between minimum error entropy and maximum correntropy. *IEEE Signal Processing Letters*, 25(7):921–925, 2018.

[8] J. Richer. IETF RFC 8485, Vectors of Trust, Oct. 2018.

[9] INGR. Security, 1st ed., 2019.

[10] A. Chorti, K. Papadaki, and H.V. Poor. Optimal power allocation in block fading channels with confidential messages. *IEEE Trans. Wireless Commun.*, 14(9):4708–4719, Sep. 2015.

[11] 3GPP. TR33.809, "Study on 5G security enhancements against false base stations (Rel 16)", Sep. 2018.

[12] Y. Arjoune and S. Faruque. smart jamming attacks in 5g new radio: A review. In *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*.

[13] O. S. Rothaus. On "bent" functions. *Journal of Combinatorial Theory*, Series A. 20(3):300–305, May 1976.

[14] N. Michailow, M. Matthé, I. S. Gaspar, A. N. Caldevilla, L. L. Mendes, A. Festag, and G. Fettweis. Generalized frequency division multiplexing for 5th generation cellular networks. *IEEE Transactions on Communications*, 62(9):3045–3061, 2014.

[15] Wenjuan Yu, Chuan Heng Foh, Atta ul Quddus, Yuanwei Liu, and Rahim Tafazolli. Throughput analysis and user barring design for uplink noma-enabled random access.

[16] Y. Polyanskiy, H. V. Poor, and S. Verdu. Channel coding rate in the finite blocklength regime. *IEEE Transactions on Information Theory*, 56(5):2307–2359, 2010.

[17] M. Pischella, A. Chorti, and I. Fijalkow. Performance analysis of uplink noma-relevant strategy under statistical delay qos constraints. *IEEE Wireless Communications Letters*, pages 1–1, 2020.

[18] I. Kanaras, A. Chorti, M. R. D. Rodrigues, and I. Darwazeh. Spectrally efficient fdm signals: Bandwidth gain at the expense of receiver complexity. In *2009 IEEE International Conference on Communications*, pages 1–6, 2009.

[19] O. Kanhere, S. Ju, Y. Xing, and T. S. Rappaport. Map-assisted millimeter wave localization for accurate position location. In *2019 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6, 2019.

[20] T. Rosinol Vidal, H.Rebecq, T. Horstschaefer, and D. Scaramuzza. Ultimate SLAM? Combining Events, Images, and IMU for Robust Visual SLAM in HDR and High Speed Scenarios. *IEEE Robotics and Automation Letters (RA-L)*, 2018.

Appendix A

Selected Recent Publications

This Appendix contains five recent publications, that have not been discussed in this thesis, on the following topics:

- Active attacks on SKG [J15];
- The use of energy harvesting to overcome jamming attacks [J16];
- Effective secrecy capacity in NOMA downlink networks [J19];
- Detection of changes in the variance of time series with application to resource allocation [J20];
- NOMA-relevant in uplink networks [J22].

Protecting Secret Key Generation Systems Against Jamming: Energy Harvesting and Channel Hopping Approaches

E. Veronica Belmega, *Member, IEEE*, and Arsenia Chorti, *Member, IEEE*

***Abstract*— Jamming attacks represent a critical vulnerability for wireless secret key generation (SKG) systems. In this paper, two counter-jamming approaches are investigated for SKG systems: first, the employment of energy harvesting (EH) at the legitimate nodes to turn part of the jamming power into useful communication power, and, second, the use of channel hopping or power spreading in block fading channels to reduce the impact of jamming. In both cases, the adversarial interaction between the pair of legitimate nodes and the jammer is formulated as a two-player zero-sum game and the Nash and Stackelberg equilibria are characterized analytically and in closed form. In particular, in the case of EH receivers, the existence of a critical transmission power for the legitimate nodes allows the full characterization of the game's equilibria and also enables the complete neutralization of the jammer. In the case of channel hopping versus power spreading techniques, it is shown that the jammer's optimal strategy is always power spreading while the legitimate nodes should only use power spreading in the high signal-to-interference ratio (SIR) regime. In the low SIR regime, when avoiding the jammer's interference becomes critical, channel hopping is optimal for the legitimate nodes. Numerical results demonstrate the efficiency of both counter-jamming measures.**

***Index Terms*— Secret key generation, jamming, energy harvesting, channel hopping, zero-sum game.**

I. Introduction

SECRET key generation (SKG) from shared randomness at two remote locations has been extensively studied [3]–[10] and has recently been extended to unauthenticated channels [11], [12]. SKG techniques have also been incorporated in protocols that are resilient to spoofing, tampering and man-in-the-middle active attacks [13]. Still, such key generation techniques are not entirely robust against active adversaries, particularly during the advantage distillation phase. Denial of service attacks in the form of jamming are a known vulnerability of SKG systems; in [14], it was demonstrated that when increasing the jamming power, the reconciliation rate normalized to the rate of the SKG increases sharply and the SKG process can in essence be brought to a halt. As SKG techniques are currently being considered for applications such as the Internet of things (IoT) [15], the study of appropriate counter-jamming approaches is timely.

Typically, jamming in wireless communication systems has been investigated using game theoretic tools [16]–[23]. Contrary to our work, these earlier studies focus on performance metrics that are either based on the legitimate nodes' signal-to-interference-plus-noise ratio (SINR) [16]–[21] and do not incorporate physical-layer security constraints at all, or are based on the secrecy capacity [22], [23]. The secrecy capacity is inherently different than the SKG capacity considered in this work; the former measures the maximum rate at which both confidential and reliable communication is possible, while the latter represents the maximum rate at which a common secret key can be extracted from the observation of correlated sequences at two remote locations [24].

In the past, two main counter-jamming approaches have been commonly considered: direct sequence spread spectrum (DSSS) and frequency hopping spread spectrum (FHSS) [25], [26]. In either approach, the impact of power constrained jammers can be limited because their optimal strategy has been proved to be the spreading of their available power over the entire bandwidth (and thus jam with potentially low power). However, DSSS and FHSS systems require a pre-shared secret to establish the spreading sequence or the hopping pattern at Alice and Bob; as such, they are not directly applicable to SKG systems that on the contrary *seek to establish* a secret key. Attempting to resolve this contradiction and reconcile DSSS and FHSS with SKG, uncoordinated frequency hopping and spreading techniques have recently been investigated in [27] and [28]. The main idea behind the proposed approaches was the randomization of the selection of the hopping/spreading sequences, at the cost of reducing the achievable rates for secret key establishment.

However, in uncoordinated hopping/spreading techniques there are minimum requirements regarding the length of the pseudorandom sequences employed. As a result, accounting for the strict bandwidth specifications of fourth and fifth generation networks, the use of long pseudorandom sequences can be a limiting factor. Thus, investigating different counter-jamming approaches based on the use of channel hopping or power spreading over multiple orthogonal subcarriers, e.g., orthogonal frequency division multiplexing (OFDM)

Manuscript received November 16, 2016; revised March 17, 2017 and April 24, 2017; accepted May 19, 2017. Date of publication June 7, 2017; date of current version July 26, 2017. This work was supported in part by ENSEA, Cergy-Pontoise, France, and in part by LABEX MME-DII. This paper was presented at the Proceedings IEEE International Conference on Communications 2017 [1], [2]. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Qian Wang. *(Corresponding author: E. Veronica Belmega.)*

E. V. Belmega is with ETIS, UMR 8051, Université Paris Seine, Université Cergy-Pointoise, ENSEA, CNRS, France, and also with Inria (e-mail: belmega@ensea.fr).

A. Chorti is with the School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, U.K. (e-mail: achorti@essex.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2017.2713342

systems [16], [18], is timely and offers an interesting alternative to [27] and [28] as in OFDM systems there is no need for coordination of the remote nodes. Furthermore, although in [27] and [28] the numerical investigations focused on the throughput, a Media Access Control (MAC) layer quantity, when analyzing physical layer security SKG systems the standard approach is to utilize the SKG capacity (a physical layer quantity).

On a different note, next generation terminals are likely to be enhanced with many new features that could prove pivotal in protecting against jamming. For example, greater energy autonomy exploiting energy harvesting (EH) approaches [29], [30] is being researched for systems such as wireless sensor networks for IoT applications. Thus, it is interesting to investigate whether EH could be utilized as a counter-jamming technique by exploiting the harvested jamming power to enhance the quality of the legitimate communication.

Motivated by the above, in the present work we propose two novel approaches for alleviating the impact of jamming in SKG systems. In both approaches, we model the interaction between the legitimate nodes and the adversarial jammer as a two-player zero-sum game in which the SKG capacity plays the role of the utility function. We investigate two non-cooperative solutions: the Nash equilibria (NE), when both players make their decision simultaneously and the Stackelberg equilibria (SE), when the legitimate nodes have an advantage and choose their strategy first while anticipating the jammer's response.

In the first part of this contribution, we study systems in which the legitimate nodes are equipped with EH capabilities and examine whether this added functionality is useful in preempting jamming attacks. We focus on time switching EH protocols [30]: for a fraction of time the legitimate nodes operate in EH mode and switch to the SKG procedure for the rest. To the best of our knowledge, this is among the first works to investigate EH as a counter-jamming approach with the exception of [21].[1]

Our analysis reveals the existence of a critical power thres-hold p_{th} for the legitimate nodes and of an associated threshold harvesting duration τ_{th}. When the legitimate nodes employ EH for longer than τ_{th}, the attacker's optimal strategy is not to jam at all, i.e., the jammer is effectively neutralized. However, neutralizing the jammer is not a stable solution to unilateral deviations (if the strategic decisions are taken simultaneously) and is therefore not a Nash equilibrium (NE) of the game. At the NE, it is found that both the legitimate nodes and the jammer transmit with full power and that the EH duration does not correspond always to the above threshold. At low signal to interference ratio (SIR) (e.g., relatively low transmit power or high jamming power), the EH optimal duration equals τ_{th}. Although the attacker jams with full power, the power collected from EH cancels out the impact of the attack and the SKG capacity is equivalent to the case of using EH for the same duration in absence of a jammer. At medium to high SIR, the EH optimal duration becomes lower than τ_{th} and decreases until the legitimate nodes do not harvest energy at all.

Furthermore, when moving to a hierarchical game formulation, the SE analysis reveals that the legitimate nodes should play the NE strategy. Whenever the legitimate nodes' harvest energy for a duration τ_{th} (at the NE), the jammer neutralization strategy is also a SE solution. This means that, in a hierarchical game, the jammer can potentially be deterred from launching the attack.

In the second part of this investigation, extending the studies in [19] and [21] to SKG systems, counter-jamming policies are investigated for N block fading additive white Gaussian noise (BF AWGN) channels, e.g., systems with N orthogonal subcarriers. At the NE, the jammer always spreads its power over all subcarriers, while for the legitimate nodes the optimality of channel hopping or power spreading depends on the channel parameters. In the high SIR regime, the legitimate nodes should use power spreading to exploit the entire available spectrum given the relatively low jamming interference. On the other hand, at low SIR, the legitimate nodes should use channel hopping and transmit over a single subcarrier to avoid most of the jammer's interference. Furthermore, in characterizing the game's SE we find that the optimal SE strategies reduce to the NE ones, demonstrating that there is no extra payoff to be earned from the advantage of playing first.

Preliminary results of this work have been presented in [1] and [2]. The major contributions and improvements of this journal paper as compared with [1] and [2] consist in: providing complete proofs of all the results regarding the NE analysis and the jammer neutralization state; relaxing the action set of the jammer, in the energy harvesting case, from the discrete choice between remaining silent and transmitting at full power into the continuous interval of all possible power values, which has brought to light the existence of additional NEs; providing the additional analysis of the Stackelberg equilibrium; providing a comparative discussion between the two counter-jamming methods in Sec. V-C.

The paper is organized as follows. In Sec. II, the SKG baseline system model is introduced. In Sec. III, the adversarial interaction between the EH legitimate nodes and the jammer is formulated and analyzed using a zero-sum non-cooperative game framework, while in Sec. IV this setting is used to study channel hopping vs. power spreading in BF AWGN systems. Numerical illustrations and a detailed discussion of these counter-jamming strategies are provided in Sec. V, while the conclusions are given in Sec. VI.

[1]The recent work [21] proposes to harvest energy from the jamming interference in a multi-user interference channel in which the jammer is not a strategic decision maker. In terms of formulation, a global optimization problem is investigated (as opposed to an adversarial game). Furthermore, the global performance metric in [21] does not incorporate security constraints and the harvested energy is not directly exploited in the communication phase, appearing only as an additional term in the utility function.

II. SKG System Model in the Presence of a Jammer

The baseline SKG system model with two legitimate nodes, denoted by Alice and Bob and a single adversary, denoted by Eve, is depicted in Fig. 1. Typically, the SKG process consists of three phases [4], [6]. In the first phase, referred

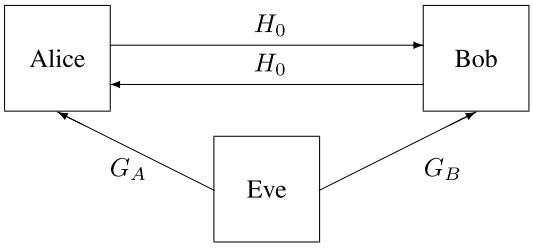


Fig. 1. SKG system model with two legitimate nodes and a single adversary.

to as *shared randomness distillation*, Alice and Bob observe dependent random variables denoted by Y_A, Y_B while an eavesdropper, referred to as Eve, observes Y_E. In wireless channels, a readily available source of shared randomness is the multipath fading due to the reciprocity of the wireless medium during the channel's coherence time [10]. Here, we focus exclusively on shared randomness extraction from Rayleigh fading coefficients.

In the next two phases, known as *information reconciliation* and *privacy amplification*, side information V is exchanged between Alice and Bob, generated by corresponding encoders f_A, f_B. At the end of the SKG process, a common key $K \in \mathcal{K}$ is extracted at Alice and Bob such that, for any $\epsilon > 0$, the following statements hold [8]:

$$Pr\left(K = f_A\left(Y_A, V\right) = f_B\left(Y_B, V\right)\right) \geq 1 - \epsilon, \quad (1)$$

$$I(K; V) \leq \epsilon, \quad (2)$$

$$H(K) \geq \log|\mathcal{K}| - \epsilon, \quad (3)$$

where $H(K)$ denotes the entropy of the key K and $I(K; V)$ denotes the mutual information between K and V.

The first inequality demonstrates that the SKG process can be made error free; (2) ensures that the exchange of side information through public discussion does not leak any information to eavesdroppers; while (3) establishes that the generated keys attain maximum entropy (i.e., are uniform). Under the three conditions, an upper bound on the rate for the generation of secret keys is given by $\min\{I(Y_A; Y_B), I(Y_A; Y_B|Y_E)\}$ [3], [4]. Assuming rich multipath environments, the decorrelation properties of the wireless channel over short distances can be exploited to ensure that Eve's observation Y_E is uncorrelated with Y_A and Y_B [7], [8], [10]; in this case, the SKG capacity is given by [3, Sec. II]

$$C = I(Y_A; Y_B). \quad (4)$$

We assume that this holds true in the rest of this study and consider the SKG capacity above to be the focal performance metric.

SKG in Rayleigh fading channels has been extensively analyzed, e.g., [7], [8]. In these works, it was assumed that Alice and Bob exchange unit probe signals to excite the fading channel and obtain respective observations Y_A and Y_B with

$$Y_A = H_0 + Z_A, \; Y_B = H_0 + Z_B,$$

where H_0 denotes the fading coefficient in the link between the legitimate nodes, modeled as a zero mean Gaussian random variable $H_0 \sim \mathcal{N}(0, \sigma_H^2)$, and, Z_A and Z_B model the effect of AWGN and denote independent and identically distributed (i.i.d.) Gaussian random variables $Z_A \sim \mathcal{N}(0, N_A)$, $Z_B \sim \mathcal{N}(0, N_B)$. Using this notation, the SKG capacity has been expressed as [8]:

$$C = I(Y_A; Y_B) = \frac{1}{2}\log_2\left(1 + \frac{\sigma_H^2}{N_A + N_B + \frac{N_A N_B}{\sigma_H^2}}\right). \quad (5)$$

In this work, we assume that Eve is no longer a passive eavesdropper but a malicious jammer. To include jamming attacks in the above model, we consider the following extension:

$$Y_A = \sqrt{p}H_0 + \sqrt{\gamma}G_A + Z_A, \quad (6)$$

$$Y_B = \sqrt{p}H_0 + \sqrt{\gamma}G_B + Z_B, \quad (7)$$

assuming that Alice and Bob exchange constant probe signals [8] with power $p \leq P$ and that Eve transmits constant jamming signals [14] with power $\gamma \leq \Gamma$. The fading coefficient in the link between Eve and Alice is denoted by $G_A \sim \mathcal{N}\left(0, \sigma_A^2\right)$ and in the link between Eve and Bob by $G_B \sim \mathcal{N}\left(0, \sigma_B^2\right)$. For simplicity and without loss of generality, the noise variables Z_A and Z_B are assumed to have unit variance, i.e., are modeled as i.i.d. Gaussian random variables $Z_A, \; Z_B \sim \mathcal{N}(0, 1)$.

Under these assumptions, a simple calculation reveals that the SKG capacity can be expressed as a function of p and γ:

$$C(p, \gamma) = \frac{1}{2}\log_2\left(1 + \frac{\sigma_H^2 p}{2 + (\sigma_A^2 + \sigma_B^2)\gamma + \frac{(1+\sigma_A^2\gamma)(1+\sigma_B^2\gamma)}{\sigma_H^2 p}}\right). \quad (8)$$

By inspecting the first-order derivatives of (8), we conclude that $C(p, \gamma)$ is a strictly increasing function of p for any fixed γ, and a strictly decreasing function of γ for any fixed p. This implies that the legitimate nodes will transmit at full power P to maximize the SKG capacity, whereas the jammer will also transmit with full power Γ to minimize the SKG capacity. Also, it is a strictly convex function with respect to (w.r.t.) γ for any fixed $p > 0$ as its second derivative w.r.t. γ is strictly positive.

III. Energy Harvesting Against Jamming

In order to study EH as a counter-jamming measure, we focus on a time-switching EH scheme [30], i.e., we assume that each transmission interval of duration T is divided in two parts. In the first period of duration τT ($0 < \tau \leq 1$ being the fraction of T dedicated to EH), both Alice and Bob operate in EH mode with efficiency $0 < \zeta \leq 1$; in the second period of duration $(1 - \tau)T$, the legitimate nodes operate in SKG mode using the overall available power (including harvested power). For simplicity, we assume that the energy harvested can be stored in a battery without any overflowing issues (unlimited storage) [31].

Furthermore, to ease the mathematical derivation and by ensuring symmetry in the energy harvested at Alice and Bob we assume that $\sigma_A^2 = \sigma_B^2 = \sigma^2$ (the Eve-Alice and Eve-Bob links have equal variance). Given the above considerations and

assuming that the energy harvested by Alice and Bob is linear in the received RF power [30]:

$$E = \zeta \tau T \gamma \sigma^2, \tag{9}$$

the harvested power for each legitimate node per transmission interval can be expressed as

$$p^{EH} = \frac{E}{(1-\tau)T} = \kappa \gamma, \tag{10}$$

where $\kappa = \frac{\zeta \tau \sigma^2}{1-\tau}$ is a convex increasing function of τ. Thus, the SKG capacity is given by:

$$\tilde{u}(p, \tau, \gamma) = \frac{1-\tau}{2} \log_2 \left(1 + \frac{\left(\frac{p}{1-\tau} + \kappa\gamma\right)\sigma_H^2}{2(1+\sigma^2\gamma) + \frac{(1+\sigma^2\gamma)^2}{\left(\frac{p}{1-\tau}+\kappa\gamma\right)\sigma_H^2}} \right), \tag{11}$$

with power constraints $p \leq P$, $\gamma \leq \Gamma$.

A simple inspection of (11) reveals that this scenario is a generalization of the standard SKG setting. Indeed, if the legitimate nodes decide not to harvest energy, i.e., $\tau = 0$, (8) is obtained for $\sigma_A^2 = \sigma_B^2 = \sigma^2$, $N_A = N_B = 1$. In the model with EH, the legitimate nodes can maximize $\tilde{u}$ by tuning the additional variable τ. However, it is no longer straightforward that the jammer should transmit with the maximum available power as $\tilde{u}(p, \gamma, \tau)$ is no longer monotonically decreasing in γ.

Non-cooperative game theory provides the natural framework to study the adversarial interaction between the legitimate nodes and the jammer. Although game theory has already been exploited in physical layer security problems, e.g. [22], [23], to the best of our knowledge, this work is among the first to investigate EH as an effective means to counteract on jamming attacks.

A. Jammer Neutralization

Before introducing the game framework, we make two important observations regarding the SKG utility in (11) and discuss their implications.

Remark 1: For any fixed τ and γ, $\tilde{u}(p, \tau, \gamma)$ is monotonically increasing in p and

$$\arg \max_{p \in [0,P]} \tilde{u}(p, \tau, \gamma) = P. \tag{12}$$

Remark 2: For any fixed p and τ, $\tilde{u}(p, \tau, \gamma)$ is monotone in γ. In particular, it is monotonically decreasing in γ if $p > p_{th}(\tau) \triangleq \zeta\tau$, a constant if $p = p_{th}(\tau)$, and monotonically increasing if $p < p_{th}(\tau)$. This implies that:

$$\arg \min_{\gamma \in [0,\Gamma]} \tilde{u}(p, \tau, \gamma) = 0, \quad \text{if } p < p_{th}(\tau) \tag{13}$$

$$\arg \min_{\gamma \in [0,\Gamma]} \tilde{u}(p, \tau, \gamma) \in [0, \Gamma], \quad \text{if } p = p_{th}(\tau) \tag{14}$$

$$\arg \min_{\gamma \in [0,\Gamma]} \tilde{u}(p, \tau, \gamma) = \Gamma, \quad \text{if } p > p_{th}(\tau). \tag{15}$$

Remark 1 shows that, to maximize the utility, the legitimate nodes should transmit at maximum power P. On the contrary, Remark 2 shows that the jammer should practically switch in between staying silent, i.e., $\gamma = 0$, and jamming at full power, i.e., $\gamma = \Gamma$, depending on the choice (p, τ) of the legitimate nodes.

Remark 2 reveals that the legitimate nodes can neutralize the jammer by transmitting at a relatively low power $p < p_{th}(\tau)$. Although this result may seem counter-intuitive at first, this condition is equivalent to $\tau > \tau_{th}(p) \triangleq \frac{p}{\zeta}$, which means that the legitimate nodes spend a relatively large proportion of time harvesting the jamming interference before actually transmitting. In other words, the jammer is forced to stay silent since the harm it can cause by interfering in the SKG phase is overcome by the harvested energy in the EH phase. This novel result shows that the jamming interference, which is commonly thought as being harmful to the legitimate communication, can be exploited and transformed into useful power via EH. If Alice and Bob transmit with exactly $p_{th}(\tau)$, the jammer becomes indifferent between all its choices $\gamma \in [0, \Gamma]$ and has no interest in actively jamming the transmission.

The necessary conditions for the jammer neutralization are formalized below.

Proposition 1: The optimal strategy for the legitimate nodes that maximizes the SKG utility while ensuring that the jammer has no interest in actively jamming the transmission is given by:

$$p^{NJ} = \min\{P, p_{th}(\tau^*)\} \quad \textit{and} \quad \tau^{NJ} = \min\{\tau_{th}(P), \tau^*\}, \tag{16}$$

where $\tau^* \in (0, 1)$ is the unique maximizer of $\tilde{u}(p_{th}(\tau), \tau, 0)$ w.r.t. τ.

For the detailed proof the reader is referred to Appendix VI-A. Notice that, if the jammer stays silent $\gamma = 0$, there is no actual energy harvested during the EH phase of duration τ^{NJ}. Rather, the legitimate nodes' choice to use EH for a fraction of time τ^{NJ} acts as an effective threat to ensure the jammer has no interest in actively jamming the transmission. However, neutralizing the jammer may not be the overall optimal strategy for the legitimate nodes. A hint for this is that whenever $\tau^{NJ} = \tau^* < \tau_{th}(P)$, the transmit power is $p^{NJ} < P$, which we know is not optimal from Remark 1.

B. Game Formulation and Nash Equilibria

The interaction between the legitimate nodes and the jammer is formalized as a two-player zero-sum game, defined as the tuple $\tilde{\mathcal{G}} = \{\tilde{\mathcal{A}}_L, \tilde{\mathcal{A}}_J, \tilde{u}(p, \tau, \gamma)\}$ in which the players are: player L representing the legitimate nodes (Alice and Bob act as a single player) on one side, and player J, the jammer, on the other. The action (p, τ) of player L lies in the set $\tilde{\mathcal{A}}_L = [0, P] \times [0, 1]$, and the action γ of player J lies in the set $\tilde{\mathcal{A}}_J = [0, \Gamma]$. The objective of player L is to maximize the SKG utility $\tilde{u}(p, \tau, \gamma)$ given in (11), whereas player J aims at minimizing it.

The two players are adversaries and the optimal strategy of one player depends on the choice of their opponent and cannot be determined unilaterally. In such interactive situations, the NE [32] is the natural solution concept. Intuitively, a profile $(p^{NE}, \tau^{NE}, \gamma^{NE}) \in \tilde{\mathcal{A}}_L \times \tilde{\mathcal{A}}_J$ is a NE if none of the players can benefit by deviating from this profile knowing

that their opponent plays accordingly. Hence, NEs are system states that are stable to unilateral deviations.

We can easily check that the state $(p^{NJ}, \tau^{NJ}, 0)$ is not a NE since the legitimate nodes gain by deviating from it. Knowing that the jammer stays silent, player L can increase the SKG utility by deviating to $\tau = 0$. Using the whole duration T in SKG mode increases the utility when no energy is harvested in the EH phase. This, in turn, will cause also the jammer to deviate from $\gamma = 0$ and actively jam the transmission.

Theorem 1 shows that the game $\tilde{\mathcal{G}}$ has at least one NE at which both players transmit with maximum power. This NE may be unique or not, depending on the system parameters.

Theorem 1: The game $\tilde{\mathcal{G}}$ has at least one NE. Moreover, the profile (P, τ^{NE}, Γ) is a NE solution such that the EH strategy is either $\tau^{NE} = 0$ or $\tau^{NE} = \min\{\tau_{th}(P), \tau_{max}\}$ with $\tau_{th}(P) = \frac{P}{\zeta}$ and $\tau_{max} \in (0, 1)$ representing the critical maximum point of $\tilde{u}(P, \tau, \Gamma)$ w.r.t. τ, depending on the system parameters. If $\tau^{NE} < \tau_{th}(P)$, then the profile (P, τ^{NE}, Γ) is the unique NE of the game almost surely.

The proof is detailed in Appendix VI-B. We observe that, at the NE above (P, τ^{NE}, Γ) and depending on the system parameters, player L may harvest energy for a fraction of time $\tau^{NE} \leq \tau^{NJ}$ or not at all $\tau^{NE} = 0$. Intuitively, not using the SKG mode for the entire transmission symbol (for example to neutralize the jammer) becomes too costly at high SIR when the jamming interference is relatively low or negligible.

Concerning the uniqueness of the NE, the only case in which the states $(P, 0, \Gamma)$ and $(P, \min\{\tau_{max}, \tau_{th}\}, \Gamma)$ can both be NEs is when the provided utilities are identical, i.e., $\tilde{u}(P, 0, \Gamma) = \tilde{u}(P, \min\{\tau_{max}, \tau_{th}(P)\}, \Gamma)$ in addition to the constraint on the system parameters $1 + \sigma^2\Gamma \geq \sqrt{2}\sigma_H^2 P$ (see Appendix VI-B). However, we argue that such an equality condition on the system parameters can only happen in very special cases, otherwise stated, with zero probability (on a continuous sample space).

Furthermore, whenever player L chooses a strategy of the form $(P, \tau_{th}(P))$ at the NE, the jammer becomes indifferent between all their possible transmit powers in $[0, \Gamma]$ (as per Remark 2). Hence, in such cases, the strategy profile $(P, \tau_{th}(P), \Gamma)$ may not be the unique NE.

Theorem 2: If the legitimate nodes' NE strategy in Theorem 1 is such that $\tau^{NE} = \tau_{th}(P)$, the game $\tilde{\mathcal{G}}$ may have other solutions of the form $(P, \tau_{th}(P), \gamma^{NE})$ with $\gamma^{NE} \in (0, \Gamma)$. More precisely, any strategy of the form $(P, \tau_{th}(P), \gamma^{NE})$ with $\gamma^{NE} \in (0, \Gamma)$ meeting the additional condition $\arg\max_{\tau \in [0,1]} \tilde{u}(P, \tau, \gamma^{NE}) = \tau_{th}(P)$ is also a NE of the game. All such NEs provide identical utility to $\tilde{u}(P, \tau_{th}(P), \Gamma)$.

The proof and the detailed system conditions under which the game may have other NEs of the type $(P, \tau_{th}(P), \gamma^{NE})$ with $\gamma^{NE} \in (0, \Gamma)$ aside from $(P, \tau_{th}(P), \Gamma)$ is provided in Appendix VI-B. These NEs may exist with non-zero proba-bility since the additional condition depends on the variable $\gamma^{NE} \in (0, \Gamma)$ and not only on the system para-meters, as opposed to the condition entailing that $(P, 0, \Gamma)$ and $(P, \min\{\tau_{max}, \tau_{th}\}, \Gamma)$ are both NEs. It suffices that $\arg\max_{\tau \in [0,1]} \tilde{u}(P, \tau, \gamma^{NE}) = \tau_{th}(P)$ holds for a single value of $\gamma^{NE} \in (0, \Gamma)$ to entail the existence of such NEs.

Apart from providing a complete NE analysis, the existence of the NEs in Theorem 2 is not very relevant in practice. First, whenever they exist, the utility at such NEs is identical to the utility of the NE profile: $(P, \tau_{th}(P), \Gamma)$ in Theorem 1. Second, given Remark 2, the jammer can be assumed to restrict their strategy space from $[0, \Gamma]$ to the discrete choices $\{0, \Gamma\}$ with no loss of optimality. Assuming $\tilde{\mathcal{A}}_J = \{0, \Gamma\}$, the resulting game $\tilde{\mathcal{G}}$ has a unique pure-strategy NE (almost surely) which is given in Theorem 1.

As a last result, it turns out that neutralizing the jammer (NJ) in Proposition 1 incurs a non-trivial cost and the obtained utility is lower or equal to the NE utility.

Proposition 2: The SKG utility obtained when neutralizing the jammer (NJ) can never be greater that the utility at the NE. Both utilities are equal, if and only if $\tau^{NE} = \tau_{th}(P)$.

Proof: Since $(P, \tau^{NE}) = \arg\max_{p,\tau} \tilde{u}(p, \tau, \Gamma)$, from the NE's best-response property, we have that $\tilde{u}(p^{NJ}, \tau^{NJ}, \Gamma) \leq \tilde{u}(P, \tau^{NE}, \Gamma)$. From Remark 2, we have that $\tilde{u}(p^{NJ}, \tau^{NJ}, \Gamma) = \tilde{u}(p^{NJ}, \tau^{NJ}, 0)$ (the jammer is indifferent between all its choices) and we obtain that $\tilde{u}(P, \tau^{NE}, \Gamma) \geq \tilde{u}(p^{NJ}, \tau^{NJ}, 0)$. Intuitively, when searching for the NJ state in Proposition 1 the additional condition that the jammer has to be neutralized (i.e., $p = p_{th}(\tau)$) restricts the feasible set of all pairs (p, τ) which results in an optimality loss compared to the NE. Notice that $\max_\tau \tilde{u}(p_{th}(\tau), \tau, 0) \equiv \max_\tau \tilde{u}(p_{th}(\tau), \tau, \Gamma)$. This further implies that, if $\tau^{NE} = \tau_{th}(P)$, the aforementioned restriction is optimal and $(p^{NJ}, \tau^{NJ}) = (P, \tau^{NE})$ which proves the direct implication of the second claim. The hypothesis of the reverse implication: $\tilde{u}(p^{NJ}, \tau^{NJ}, 0) = \tilde{u}(P, \tau^{NE}, \Gamma)$ implies that $\tau^{NE} = \tau_{th}(P)$ and, thus, $\tilde{u}(p^{NJ}, \tau^{NJ}, 0) = \tilde{u}(P, \tau^{NE}, 0)$. From Appendix VI-A, the function $\tilde{u}(p_{th}(\tau), \tau, 0)$ has a unique maximizer w.r.t $\tau \in [0, \tau_{th}(P)]$ given by τ^{NJ} which results in that $(p^{NJ}, \tau^{NJ}) = (P, \tau^{NE})$. ■

C. Stackelberg Equilibrium

After investigating the NE solution of the strategic interaction in which the legitimate nodes and the jammer choose their optimal strategies simultaneously, a natural rising issue is whether the solution of the game changes assuming a hierarchy in the players' choices [20], [22], [32]. To tackle this issue, we study the SE and compare it to the NE and the jammer neutralization (NJ) states in Sec. III-B and III-A, respectively. We assume that the leader of the game L is playing first by choosing their best action (p^{SE}, τ^{SE}) while anticipating the response of player J. The follower, player J, observes the choice of the leader and reacts optimally (or best-responds) by choosing γ^{SE}.

To be specific, for an arbitrary choice of player L (p, τ), the best-response of the jammer is defined as:

$$\gamma^{BR}(p, \tau) = \arg \min_{\gamma \in [0,\Gamma]} \tilde{u}(p, \tau, \gamma). \tag{17}$$

The leader, anticipating the jammer's reaction described above, can choose their optimal strategy as follows

$$(p^{SE}, \tau^{SE}) = \arg \max_{p,\tau} \tilde{u}(p, \tau, \gamma^{BR}(p, \tau)). \tag{18}$$

The optimal strategy of the jammer is the best response $\gamma^{SE} = \gamma^{BR}(p^{SE}, \tau^{SE})$ given the optimal leader's strategy above. The solution is described in the next Theorem.

Theorem 3: Assuming the hierarchy described above, if $\tau^{NE} < \tau_{th}(P)$ where τ^{NE} is given in Theorem 1, the SE of the game $\tilde{\mathcal{G}}$ is unique (almost surely) and identical to the NE (P, τ^{NE}, Γ). Otherwise, if $\tau^{NE} = \tau_{th}(P)$, both the NJ state in Proposition 1 and the NE (P, τ^{NE}, Γ) are SE solutions providing identical SKG utility.

The proof is included in Appendix VI-C. Notice that in all possible cases $\tau^{NE} \leq \tau_{th}(P)$ (see Theorem 1). The above result shows that neutralizing the jammer is a rational solution when the strategic decisions are not taken simultaneously and the legitimate nodes play first. However, since the NJ state cannot provide a strictly better utility than the NE state (see Proposition 2), the hierarchical play does not bring an actual benefit to player L when compared with the NE.

Finally, we note that as opposed to the NE, the SE requires the leader to be able to anticipate precisely the response of the follower. For this reason, the leader cannot actually choose a strategy such that $p = p_{th}(\tau)$ which renders the follower indifferent between all its actions $\gamma \in [0, \Gamma]$ (and may choose any jamming power in an unpredictable way). A simple way to overcome this issue is for the leader to transmit at $p = p_{th}(\tau) - \varepsilon$ whenever it wants to silence the jammer (at the NJ), and to transmit at $p = p_{th}(\tau) + \varepsilon$ whenever it wants the jammer to transmit at full power (at the NE), with $\varepsilon > 0$ and $\varepsilon \ll 1$ chosen arbitrarily small, with little or no practical impact. Furthermore, this also the excludes other SE solutions (e.g., the NEs in Theorem 2 cannot be SEs).

IV. Channel Hopping vs. Power Spreading in BF AWGN Channels

If the legitimate nodes do not have EH capabilities, we investigate yet another way to defend against jamming by assuming that the legitimate nodes can employ channel hopping or power spreading strategies over multiple orthogonal subcarriers. For this, we generalize the system model (6) and (7) to an N-BF AWGN channel. Alice's and Bob's observations on the i-th subcarrier – denoted by $\hat{Y}_{A,i}$ and $\hat{Y}_{B,i}$ respectively – are expressed as:

$$\hat{Y}_{A,i} = \sqrt{p_i} H_i + \sqrt{\gamma_i} G_{A,i} + Z_{A,i}, \tag{19}$$

$$\hat{Y}_{B,i} = \sqrt{p_i} H_i + \sqrt{\gamma_i} G_{B,i} + Z_{B,i}, \tag{20}$$

where the fading coefficient in the link between Alice and Bob on the i-th subcarrier is denoted by H_i, in the link between Eve and Alice by $G_{A,i}$ and in the link between Eve and Bob by $G_{B,i}$. We assume that the fading coefficients are i.i.d. Gaussian random variables with $H_i \sim \mathcal{N}(0, \sigma_H^2)$, $G_{A,i} \sim \mathcal{N}(0, \sigma_A^2)$ and $G_{B,i} \sim \mathcal{N}(0, \sigma_B^2)$. Notice that the fading coefficients are assumed to have the same statistics. This assumption is justified, since, broadly speaking, narrowband fading depends on the bandwidth (which is the same for all subcarriers) and not on the central frequency (unlike wideband fading or large scale fading) [33]. Furthermore, the noise variables $Z_{A,i}$ and $Z_{B,i}$ are assumed to be i.i.d. Gaussian zero mean unit variance random variables. Finally, Alice and Bob exchange constant probe signals [8] with power p_i and that Eve transmits constant jamming signals [14] with power γ_i on the i-th subcarrier so that the following average power constraints are satisfied[2] [16], [18]:

$$\frac{1}{N}\sum_{i=1}^{N} p_i \leq P, \quad \frac{1}{N}\sum_{i=1}^{N} \gamma_i \leq \Gamma. \tag{21}$$

Given the above model, an easy calculation reveals that the SKG capacity over the i-th subcarrier can be expressed as a function of p_i and γ_i as:

$$\begin{aligned} C(p_i, \gamma_i) &= I(\hat{Y}_{A,i}; \hat{Y}_{B,i}) \\ &= \frac{1}{2}\log_2\left(1 + \frac{\sigma_H^2 p_i}{N_{A,i} + N_{B,i} + \frac{N_{A,i} N_{B,i}}{\sigma_H^2 p_i}}\right), \end{aligned}$$

with

$$N_{A,i} = 1 + \sigma_A^2 \gamma_i, \quad N_{B,i} = 1 + \sigma_B^2 \gamma_i. \tag{22}$$

In order to evaluate the overall SKG capacity, we formalize the channel hopping vs. power spreading techniques similarly to [16] and [18]. When channel hopping is employed, all of the available power is used to transmit on a *single* randomly chosen subcarrier i. Therefore, when the legitimate nodes employ channel hopping on subcarrier i, then $p_i = NP$ and $p_k = 0$ for $k \neq i$, while when the jammer hops on subcarrier i then $\gamma_i = N\Gamma$ and $\gamma_k = 0, k \neq i$. On the other hand, when power spreading is used, the available power is equally distributed across all subcarriers so that $p_i = P$ and $\gamma_i = \Gamma$ $\forall i \leq N$.

When transmitting over the entire spectrum, the choice of the uniform power allocation is motivated by the fact that the nodes do not know their actual channel gains and that their statistics are identical across all frequency carriers. Moreover, assuming that player L transmits with uniform power allocation and from the convexity of the SKG function in (22) w.r.t. γ_i, it turns out that the uniform power allocation for the jammer is optimal and minimizes the overall SKG utility. More general power allocation policies can be considered in future investigations.

From an implementation point of view for the proposed channel hopping and power spreading strategies, we consider that an OFDM transmitter with a standard inverse fast Fourier transform (IFFT) block is employed. In channel hopping mode, all but a randomly chosen IFFT input are set to zero. No coordination regarding the chosen channel hopping or spreading options is required between transmitting and receiving terminals. This is possible if wideband reception is employed by all parties, allowing transmitting terminals to independently choose their strategies without coordination with the receiving terminals. Such a wideband reception of the N orthogonal subcarriers can be efficiently implemented using a standard FFT based OFDM receiver.

Using this framework in the following, for Alice and Bob the probability of channel hopping on subcarrier i is

[2]Using constant probe signals preserves the Gaussianity of the inputs $\sqrt{p_i} H_i$, $\sqrt{\gamma_i} G_{A,i}$ and $\sqrt{\gamma_i} G_{B,i}$, which is optimal for the legitimate nodes and the jammer in our AWGN setting.

denoted by α_i $\forall i \leq N$, while α_{N+1} denotes the probability of spreading the available power uniformly over the whole spectrum. Similarly, we define β_i for the jammer. Since $\alpha = [\alpha_1, \ldots, \alpha_{N+1}]$ and $\beta = [\beta_1, \ldots, \beta_{N+1}]$ are discrete probability distributions, we have the constraints $\alpha_j \geq 0$, $\forall j$, $\sum_{i=1}^{N+1} \alpha_i = 1$, $\beta_j \geq 0$, $\forall j$, and $\sum_{i=1}^{N+1} \beta_i = 1$.

Given the above, the SKG capacity over the N orthogonal subcarriers is given by:

$$\begin{aligned}\hat{u}(\alpha, \beta) = \frac{1}{N}\Bigg\{ \sum_{i=1}^{N} \{&\alpha_i(1-\beta_i-\beta_{N+1})C(NP,0) \\ &+ \alpha_i\beta_i C(NP, N\Gamma) + \alpha_i\beta_{N+1}C(NP,\Gamma) \\ &+ \alpha_{N+1}\beta_i[(N-1)C(P,0) + C(P,N\Gamma)]\} \\ &+ \alpha_{N+1}\beta_{N+1}NC(P,\Gamma)\Bigg\}, \end{aligned} \tag{23}$$

where the normalization $\frac{1}{N}$ accounts for measuring the SKG capacity in bits/s/Hz. In (23), the first term corresponds to the case in which Alice (resp. Bob) hops on subcarrier i and the jammer hops on a different subcarrier; the second term to the case in which Alice (resp. Bob) and the jammer both hop on subcarrier i; the third term to the case in which Alice (resp. Bob) hops on subcarrier i and the jammer spreads; the fourth term to the case in which the Alice (resp. Bob) spreads and the jammer hops on subcarrier i. Finally, the last term corresponds to the case in which they both spread their power.

A. Game Formulation and Nash Equilibria

We model the competitive interaction between player L and J as the following zero-sum game $\hat{\mathcal{G}} = \{\hat{\mathcal{A}}_L, \hat{\mathcal{A}}_J, \hat{u}(\alpha, \beta)\}$, where the payoff $\hat{u}(\alpha, \beta)$ is given in (23). The action sets of the players are the probabilities of channel hopping and power spreading:

$$\hat{\mathcal{A}}_L = \left\{ \alpha \in [0,1]^{N+1} \left| \sum_{i=1}^{N+1} \alpha_i = 1 \right. \right\},$$
$$\hat{\mathcal{A}}_J = \left\{ \beta \in [0,1]^{N+1} \left| \sum_{i=1}^{N+1} \beta_i = 1 \right. \right\}.$$

As we have argued in the previous section, the natural solution in such a strategic interaction without cooperation among the opponents is the NE.

To derive the game's NE, let us introduce a finite discrete game $\hat{\mathcal{G}}^D = \{\hat{\mathcal{E}}_L, \hat{\mathcal{E}}_J, \hat{u}(\alpha, \beta)\}$ with action sets defined as $\hat{\mathcal{E}}_L \equiv \hat{\mathcal{E}}_J = \{e_1, \ldots, e_N, e_{(N+1)}\}$, where $e_i \in \{0,1\}^{N+1}$ is the canonical vector containing 1 on the i-th position and 0 otherwise. The i-th action e_i represents channel hopping on subcarrier i for all $i \leq N$ and e_{N+1} represents spreading the power across the spectrum. Such finite discrete games always have at least one NE in mixed strategy (α^*, β^*) [32, Sec. 1.3.1]. We observe that our game $\hat{\mathcal{G}}$ represents the mixed strategy extension of $\hat{\mathcal{G}}^D$ and thus $\hat{\mathcal{G}}$ has at least one NE.

Corollary 1 [32, Th. 1.1]: Game $\hat{\mathcal{G}}$ has at least one NE.

To compute the NEs, one possibility is to use the Minimax Theorem of von Neumann and Morgenstern [34] which allows us to compute mixed NEs of any two-player zero-sum game via linear programming (i.e., by solving two dual linear optimization problems). In our case, we show that the NEs can be characterized in an analytical closed-form manner without the need of solving any optimization problem. To this aim, an alternative characterization of the NE (see Definition 1.2 in [32, Sec.1.2.1]) is used:

Definition 1: A strategy profile $(\alpha^*, \beta^*) \in \hat{\mathcal{A}}_L \times \hat{\mathcal{A}}_J$ is a NE of the game $\hat{\mathcal{G}}$ if the following hold:

i) *both players are indifferent among the pure actions that are played with positive probability at the NE*

$$\hat{u}(\alpha^*, e_i) = \hat{u}(\alpha^*, e_k), \quad \forall i,k, \in \mathcal{I}_J,$$
$$\hat{u}(e_i, \beta^*) = \hat{u}(e_k, \beta^*), \quad \forall i,k, \in \mathcal{I}_L,$$

ii) *the pure actions that result in strictly smaller payoffs are played with zero probability at the NE*

$$\text{if } \hat{u}(\alpha^*, e_i) < \hat{u}(\alpha^*, e_k), \quad i \in \mathcal{I}_J, \quad \text{then } k \in \mathcal{N}_J,$$
$$\text{if } \hat{u}(e_i, \beta^*) > \hat{u}(e_k, \beta^*), \quad i \in \mathcal{I}_L, \quad \text{then } k \in \mathcal{N}_L,$$

where the sets $\mathcal{N}_L, \mathcal{I}_L \subseteq \{1, \ldots, N+1\}$ denote, respectively, the indices of the pure actions that are not played at the NE and those that are played at the NE by player L: $\mathcal{N}_L = \{i | \alpha_i^* = 0\}$, $\mathcal{I}_L = \{1, \ldots, N+1\} \setminus \mathcal{N}_L$; similarly, the sets $\mathcal{N}_J, \mathcal{I}_J \subseteq \{1, \ldots, N+1\}$ denote, respectively, the set of indices of the pure actions that are not used or are used by player J at the NE: $\mathcal{N}_J = \{i | \beta_i^* = 0\}$, and $\mathcal{I}_J = \{1, \ldots, N+1\} \setminus \mathcal{N}_J$.

At a first glance, Definition 1 provides a simple way to compute the NEs of the game $\hat{\mathcal{G}}$ by solving a system of linear equations and checking some conditions. Still, in order to use Definition 1, one would have to know in advance the faces of the simplex $\hat{\mathcal{A}}_L \times \hat{\mathcal{A}}_J$ on which the NEs lie, i.e., one would have to know $\mathcal{I}_L$, $\mathcal{I}_J$ for all NEs. An exhaustive search has an exponential complexity (the $N+1$-simplex has $2^{N+1} - 1$ faces). Nevertheless, the NEs of our game $\hat{\mathcal{G}}$ have a special structure which allows us to exploit Definition 1 and fully characterize the set of NEs in a simple manner.

To characterize the set of NEs as a function of the system's parameters we begin by examining the matrix structure of the discrete game $\hat{\mathcal{G}}^D$ given in Table I. We notice that there is a symmetry between the channel hopping strategies. In particular, the payoff does not depend on the particular index of the chosen subcarrier but only on whether both players hop on the same subcarrier or not. This symmetry allows us to show that the NE of the game $\hat{\mathcal{G}}$ have a particular structure specified in the following propositions.

Proposition 3: At the NE (α^*, β^*), a player uses either all channel hopping actions with non-zero probability or none of them: either $\alpha_i^* = 0$, $\forall i \leq N$ or $\alpha_i^* \neq 0$, $\forall i \leq N$, and similarly, either $\beta_i^* = 0$, $\forall i \leq N$ or $\beta_i^* \neq 0$, $\forall i \leq N$.

Proposition 4: If both players employ channel hopping with non-zero probability at the NE, i.e., $\alpha_i^* > 0$ and $\beta_i^* > 0$ $\forall i \leq N$, then the players will hop uniformly across all channels and the NE will have the following structure: $\alpha^* = (a, \ldots, a, (1-Na))$, $\beta^* = (b, \ldots, b, (1-Nb))$ for some $0 \leq a \leq 1/N$, $0 \leq b \leq 1/N$.

Propositions 3 and 4 are proven in Appendices VI-D and VI-E. These results shape the special structure of the NEs of $\hat{\mathcal{G}}$, which, alongside Definition 1 and the strict convexity of $C(p, \gamma)$ w.r.t. γ, allows us to fully

TABLE I
TWO PLAYER ZERO-SUM DESCRIPTION OF $\hat{\mathcal{G}}_d$

	$e_i,\ i \leq N$	$e_k,\ k \leq N, k \neq i$	e_{N+1}
$e_i,\ i \leq N$	$\frac{1}{N}C(NP, N\Gamma)$	$\frac{1}{N}C(NP, 0)$	$\frac{1}{N}C(NP, \Gamma)$
$e_k,\ k \leq N, k \neq i$	$\frac{1}{N}C(NP, 0)$	$\frac{1}{N}C(NP, N\Gamma)$	$\frac{1}{N}C(NP, \Gamma)$
e_{N+1}	$\frac{N-1}{N}C(P, 0) + \frac{1}{N}C(P, N\Gamma)$	$\frac{N-1}{N}C(P, 0) + \frac{1}{N}C(P, N\Gamma)$	$C(P, \Gamma)$

characterize the set of NEs in a very simple and explicit manner as function of the system parameters.

Theorem 4: The set of NEs of the game $\hat{\mathcal{G}}$ is characterized as follows:

1. *If $C(NP, \Gamma) < NC(P, \Gamma)$, then the game has a unique pure-strategy NE: both players spread their powers, $\alpha^* = \beta^* = e_{N+1}$.*
2. *If $C(NP, \Gamma) > NC(P, \Gamma)$, then player L hops and player J spreads at the NE: $\alpha^* = (\alpha_1, \ldots, \alpha_N, 0)$ and $\beta^* = e_{N+1}$. The NE strategies of player L are given by the (infinite number of) solutions to the following system of linear inequalities:*

$$\begin{cases} 0 \leq \alpha_i \leq 1, \quad \forall i \leq N, \\ \sum_{j=1}^{N} \alpha_j = 1, \\ \alpha_i < \dfrac{C(NP, 0) - C(NP, \Gamma)}{C(NP, 0) - C(NP, N\Gamma)}, \quad \forall i \leq N. \end{cases}$$

   In particular, the uniform probability over the channels is one of the NE solutions: $\alpha^* = (1/N, \ldots, 1/N, 0)$. All NEs are equivalent in terms of achieved utility.
3. *If $C(NP, \Gamma) = NC(P, \Gamma)$, player L employs all their actions and player J spreads at the NE: $\alpha^* = (\alpha_1, \ldots, \alpha_N, \alpha_{N+1})$ and $\beta^* = e_{N+1}$. The NE strategies of player L are the (infinite number of) solutions to the following linear system of inequalities:*

$$\begin{cases} \alpha_i \geq 0, \quad \forall i \leq N, \\ \sum_{j=1}^{N} \alpha_j = 1, \\ \alpha_i[C(NP, N\Gamma) - C(NP, 0)] + \alpha_{N+1}[(N-1)C(P, 0) \\ +C(P, N\Gamma) - C(NP, 0) + C(NP, \Gamma) - NC(P, \Gamma)] \\ > C(NP, \Gamma) - C(NP, 0), \quad \forall i \leq N. \end{cases}$$

   In this case, both players spreading (case 1) is an NE. Also, player J spreading and player L hopping strategies (case 2) are all NEs. All NEs are equivalent in terms of achieved utility.

The proof is provided in Appendix VI-F. We remark that the NE can be unique and in pure strategies if $C(NP, \Gamma) < NC(P, \Gamma)$ and the outcome of the game provides a utility equal to $\hat{u}(\alpha^*, \beta^*) = C(P, \Gamma)$. On the contrary, if $C(NP, \Gamma) \geq NC(P, \Gamma)$, there are an infinite number of NEs which are generally in mixed strategies. All these NEs are equivalent in terms of achieved utility, which equals $\hat{u}(\alpha^*, \beta^*) = \frac{1}{N}C(NP, \Gamma)$. Hence, the outcome of the game can be predicted without the need for implementing iterative or learning procedures.

Theorem 4 also shows that the optimal strategy for the jammer is always spreading their power across the entire spectrum. Intuitively, if the jammer were to use channel hopping, player L would exploit this fact and would also hop; this scenario is unfavorable for the jammer as the probability that both players hop on the same subcarrier equals $\frac{1}{N^2}$ (due to Proposition 3, when both players hop at the NE, they use uniform probabilities). Thus, the jammer's payoff from hopping cannot exceed that gained from spreading, assuming that the legitimate nodes play their optimal strategy. On the contrary, for player L the best strategy can be either channel hopping or power spreading depending on which option provides higher utility against a spreading jammer.

B. Stackelberg Equilibrium

In Sec. III-C, we have shown that the hierarchy of play among the adversaries does not bring an advantage to the legitimate nodes assuming they have EH capabilities. Here, we investigate whether this remains true in N-BF AWGN systems in which the players choose between channel hopping and power spreading strategies. The leader, player L, is assumed to play first and to choose α^{SE} anticipating the jammer's response. The follower, player J, observes α^{SE} and best-responds by choosing β^{SE}.

More precisely, the best-response of the jammer for an arbitrary choice of α is defined as: $\beta^{BR}(\alpha) = \arg\min_{\beta} \hat{u}(\alpha, \beta)$. Thus, the leader chooses their optimal strategy as follows

$$\alpha^{SE} = \arg\max_{\alpha} \ \hat{u}(\alpha, \beta^{BR}(\alpha)) \tag{24}$$

and the resulting best-response or SE strategy of the jammer is $\beta^{SE} = \beta^{BR}(\alpha^{SE})$.

To characterize the SE in closed-form, we use a similar approach as for the NE: we show first that the leader's strategy at the SE has a special form described below. Then, we exploit this structure to provide the SE solution.

Proposition 5: At the SE, the legitimate player uses either all hopping strategies with uniform probability or none of them, i.e., $\alpha^{SE} = (a, \ldots, a, 1 - Na)$ for some $a \in [0, 1/N]$.

The proof is provided in Appendix VI-G. The above structure of α^{SE} allows us to analyze the optimal response of the jammer β^{SE} and to prove that, in all cases, the jammer's best strategy is to spread: $\beta^{SE} = (0, \ldots, 0, 1)$. On the other hand, depending on the channel parameters, the leader will either channel hop or spread their power, identically to the NE.

Theorem 5: The set of SEs of the game $\hat{\mathcal{G}}$ is identical to the set of NEs.

The proof is provided in Appendix VI-H. Therefore, the legitimate nodes do not gain in utility by choosing first their strategy as opposed to the NE where both players choose their strategies simultaneously.

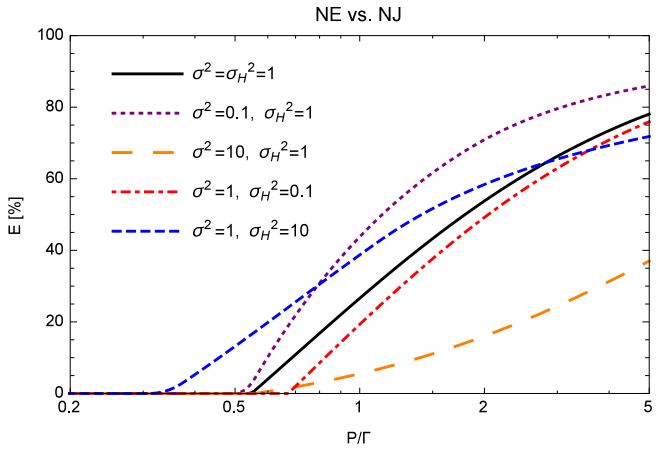


Fig. 2. Relative utility gain at the NE vs. NJ $E = (C^{NE} - C^{NJ})/C^{NE}$ as a function of $P/\Gamma \geq 0$ for $\zeta = 0.7$.

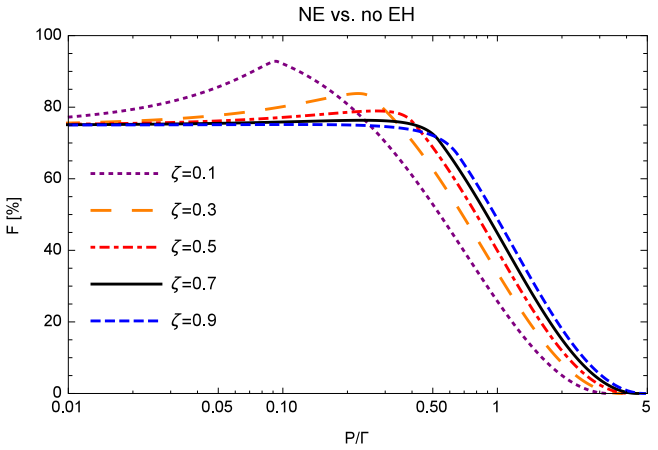


Fig. 3. Relative utility gain at the NE vs. no EH: $F = (C^{NE} - C^{noEH})/C^{NE}$ as a function of $P/\Gamma \geq 0$.

V. Numerical Illustrations and Discussion

In this Section, several representative illustrations are chosen allowing the deduction of generic conclusions that carry over most setups. The benchmark setting is chosen as follows: unit jamming power $\Gamma = 1$, unit variance Rayleigh channel coefficients $\sigma_A^2 = \sigma_B^2 = \sigma^2 = \sigma_H^2 = 1$.

A. EH at the Legitimate Nodes

We start by evaluating the SKG capacity at the NJ in Proposition 1 and NE in Theorem 1 as functions of the system parameters for a harvesting efficiency $\zeta = 0.7$. In Fig. 2, the relative gain in utility obtained at the NE ($C^{NE} = \tilde{u}(P, \tau^{NE}, \Gamma)$) compared with the NJ ($C^{NJ} = \tilde{u}(p^{NJ}, \tau^{NJ}, 0)$) defined by $E \triangleq \frac{C^{NE} - C^{NJ}}{C^{NE}}$, is depicted as a function of the signal to interference ratio (SIR) P/Γ for different values of σ^2 and σ_H^2. In the investigated settings, the NJ strategy never outperforms the NE in terms of utility, which is consistent with Proposition 2. When the SIR P/Γ is relatively low, both the NE and the NJ provide identical utilities. In this case, $p^{NJ} = P$ and $\tau^{NJ} = \tau^{NE} = \tau_{th}(P)$, the jammer is indifferent between $\{0, \Gamma\}$ and both states are SE solutions. With increasing SIR P/Γ, it is no longer optimal for the legitimate nodes to harvest energy for a fraction of time $\tau_{th}(P)$ in order to neutralize the jammer. Instead, by limiting the duration of EH to $\tau^{NE} = \tau_{max} < \tau_{th}(P)$ the SKG capacity increases in spite of the full power jamming $\gamma = \Gamma$ and only the NE is also a SE solution. Moreover, as the SIR increases, e.g., for $P/\Gamma \gg 1$, the legitimate nodes should not harvest energy at all as the jammer's interference is relatively negligible.

Notice that Fig. 2 also illustrates the SE solution described in Theorem 3. Indeed, at low SIR, when both NE and NJ provide equal SKG capacity, they are both SE solutions. At high SIR, the SE is unique and identical to the NE.

Subsequently, we evaluate the impact of the EH capability on the SKG capacity at the NE. The relative gain in utility obtained at the NE compared with the case in which there is no EH capability $C^{NoEH} = \tilde{u}(P, 0, \Gamma) = C(P, \Gamma)$, defined as $F \triangleq \frac{C^{NE} - C^{NoEH}}{C^{NE}}$, is depicted as a function of P/Γ in Fig. 3. The benchmark setup is considered and the different curves correspond to harvesting efficiencies $\zeta \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. At low SIR, F decreases with the harvesting efficiency ζ. Although counter-intuitive, this is explained by the fact that, at the NE, the harvesting period $\tau^{NE} = \tau_{th}(P) = P/\zeta$ decreases in ζ in this regime. The peaks represent the transition points to the second regime, in which harvesting energy at the threshold is no longer optimal and $\tau^{NE} = \tau_{max} < \tau_{th}(P)$. Also, as the SIR increases, the curves progressively switch ranks and F becomes increasing in ζ as expected. For $P/\Gamma = 1$ and $\zeta = 0.3$ the gain in using EH is around 30 % while it increases to 40 % for $\zeta = 0.7$. At low SIR P/Γ the gain observed can reach 90 %, while at the high SIR it is negligible: in the third regime $\tau^{NE} = 0$, as harvesting energy becomes time-consuming and inefficient in terms of SKG capacity.

Finally, the relative utility F defined above is depicted in Fig. 4 for $\zeta = 0.7$ and various channel parameters. For low SIR P/Γ, there is a significant gain in utility when employing EH. This gain becomes significantly large at very low SIR, exceeding 97.5 % when the legitimate nodes experience poor channel conditions as opposed to the jammer. When both parties experience similar channel conditions the gain is in the range of 60 % in the medium SIR. Overall, the numerical results demonstrate that using EH as a counter-jamming technique is of particular interest in the low and medium SIR regimes but, as expected, does not increase the utility in the high SIR. The peaks represent here as well the transition from the $\tau^{NE} = \tau_{th}(P)$ regime (at low SIR) to the second regime in which $\tau^{NE} = \tau_{max} < \tau_{th}(P)$.

B. Channel Hopping vs. Power Spreading

First, we analyze the NE as function of N and the ratio P/Γ for the benchmark scenario in Fig. 5. There exist two regions delimited by the curve $C(NP, \Gamma) = NC(P, \Gamma)$: a region in which the NE is unique and both players spread their power, and a region in which the jammer spreads their power and the legitimate nodes employ channel hopping.

Player L hops at the NE below the curve, when the SIR P/Γ is relatively small. This is intuitive since, in the low

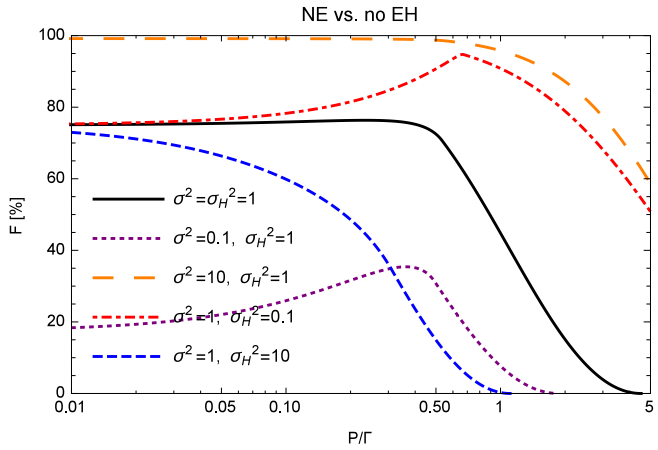


Fig. 4. Relative utility gain at the NE vs. no EH: $F = (C^{NE} - C^{noEH})/C^{NE}$ as a function of $P/\Gamma \geq 0$ for $\zeta = 0.7$ and different channel parameters.

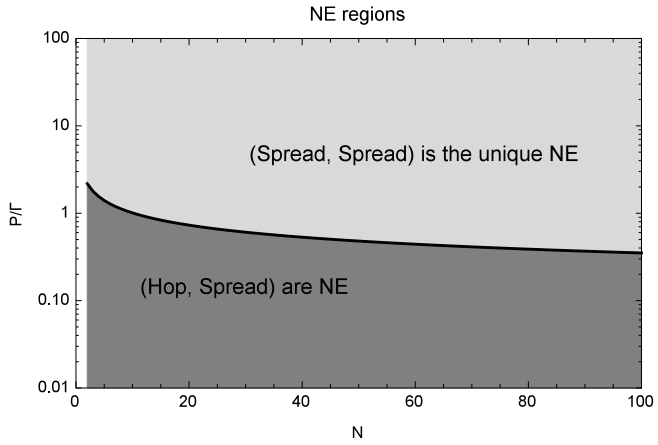


Fig. 5. NE regions as a function of $P/\Gamma \geq 0$ and $N \geq 2$ for $\Gamma = \sigma_A^2 = \sigma_B^2 = \sigma_H^2 = 1$.

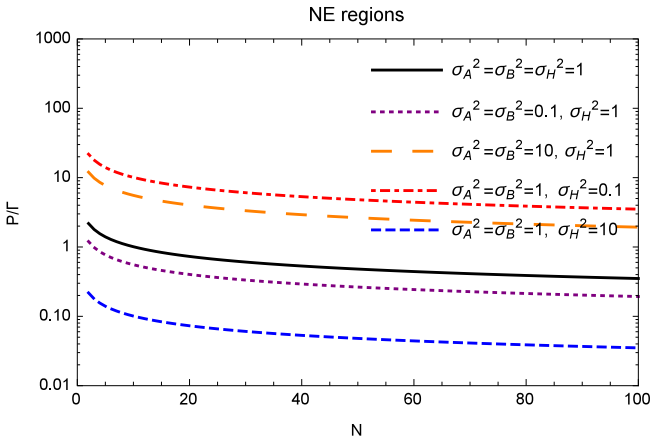


Fig. 6. NE regions as a function of $P/\Gamma \geq 0$ and $N \geq 2$ for $\Gamma = 1$ and different channel parameters.

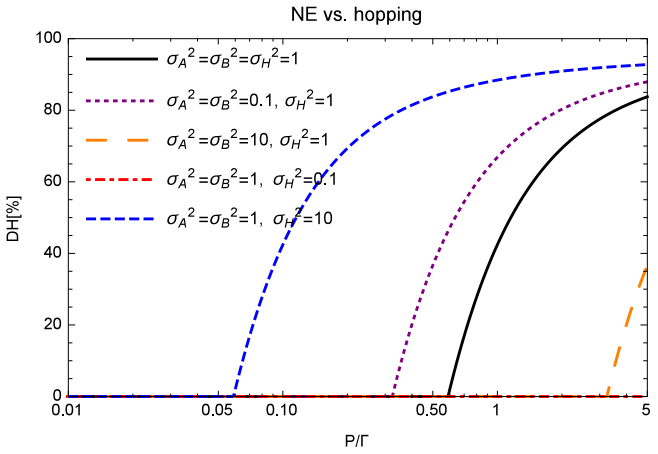


Fig. 7. Relative utility gain between the NE vs. always hopping: $D_H = (u^{NE} - u^{Hop,Spread})/u^{NE}$ as a function of P/Γ for $N = 32$, $\Gamma = 1$ and different channel parameters.

transmit power regime, the legitimate nodes should avoid as much jamming interference as possible by transmitting on a single subcarrier, which also means that their available power is concentrated on a single channel.

In Fig. 6, the NE regions are illustrated for different channel parameters. When σ_H^2 increases, the region in which player L should employ channel hopping at the NE shrinks down while when σ_A^2, σ_B^2 increase, this region expands.

In Fig. 7, the relative gain obtained by player L when employing the NE strategy as opposed to a naive hopping strategy is depicted. The relative utility gain $D_H = (u^{NE} - u^{Hop,Spread})/u^{NE}$, where $u^{Hop,Spread} = 1/NC(NP, \Gamma)$ is relatively large (up to 80%) in the high SIR regime, in which case the optimal strategy for player L is to use the entire spectrum in spite of the jammer's interference.

Finally, in Fig. 8, the relative utility gain when using the NE strategy over N subcarriers as opposed to a single subcarrier ($u^{single} = C(P, \Gamma)$) is investigated for $\Gamma = \sigma_H^2 = \sigma_A^2 = \sigma_B^2 = 1$ as a function of P/Γ for $N \in \{2, 4, 8, 16, 32, 64\}$. At low SIR, when the channel hopping strategy is optimal for the legitimate nodes, the higher the number of subcarriers N, the lower the jammer's interference in each subcarrier, and hence, the higher the SKG capacity. At last, in the high SIR regime, when spreading is optimal the SKG utility becomes $C(P, \Gamma)$, which is identical to transmitting over a single channel with powers P and Γ.

Remark that all figures illustrating the NE, in this subsection, also illustrate the SE solution, since the SE is identical to the NE as per Theorem 5.

C. Discussion and Perspectives

We discuss here the differences and similarities between the two approaches: a) EH at the legitimate nodes, and b) employing channel hopping or power spreading techniques.

EH at the legitimate nodes enables them to completely neutralize the jammer. By harvesting the jamming power in a first phase and exploiting it for SKG in a second phase, the jammer's attacks may increase the SKG capacity; in this case, the jammer should not launch the attack, i.e., it is neutralized. However, it is not always optimal for the legitimate nodes to neutralize the jammer. Indeed, using EH can reduce the SKG capacity since, for a non-trivial fraction of time, there is no secret bits generation; when the jammer is neutralized the

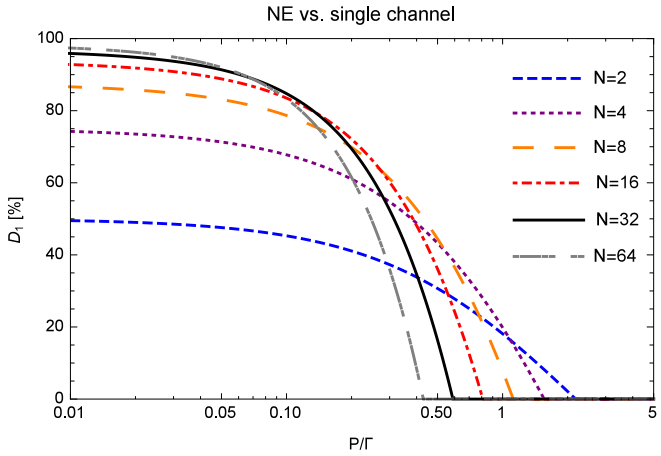


Fig. 8. Relative utility gain between the NE vs. single channel SKG: $D_1 = (u^{NE} - u^{single})/u^{NE}$ as a function of P/Γ for $\Gamma = \sigma_H^2 = \sigma_A^2 = \sigma_B^2 = 1$ and $N \in \{2, 4, 8, 16, 32, 64\}$.

penalty in terms of utility might become too high, depending on the system parameters (e.g., high SIR regime). In such cases, the obtained utility at the NE is strictly higher than the one at the NJ state.

On the other hand, in the case of BF AWGN channels (i.e., in systems with multiple orthogonal subcarriers), the idea is to use channel hopping in a random fashion and avoid most of the jammer's interference as opposed to completely neutralizing it. Since potential jammers cannot predict the subcarrier used by the legitimate nodes, they will always spread their powers over the entire spectrum: the larger the number of subcarriers, the smaller the jammer's interference on each subcarrier. However, channel hopping is not always optimal since only a fraction of the entire spectrum is used for SKG. Depending on the system parameters (high SIR), it can be preferable for the legitimate nodes to spread the available power across the entire spectrum rather than concentrate it on a single subcarrier. In this case, the SKG capacity (measured in bits/s/Hz) is identical to that of single channel with the same average power constraints.

In the critical cases of low and medium SIR regimes ($P/\Gamma < 1$), both approaches turn out to be advantageous in terms of SKG capacity compared to a single channel SKG system without EH capabilities; the gains in SKG capacity depend on the harvesting duration or the number of subcarriers N. On the contrary, in the high SIR regime ($P/\Gamma \gg 1$), the jammer's impact and interference become relatively low or even negligible and the cost of counter-jamming measures might not be justified compared to simply tolerating it. However, the interesting cases are indeed the former ones in which the jamming power is higher or of the same order as the legitimate nodes' transmit powers, in which overcoming the attack becomes critical.

For both approaches it turns out that a hierarchical decision model that in principle could favor the legitimate nodes compared to a simultaneous decision model does not bring an actual benefit. Indeed, the SKG utility obtained at the SE is identical to the SKG utility at the NE (even though the set of SEs is not necessarily identical to the set of NEs as in the EH approach).

Several questions arise for future work. First, an interesting issue would be to study reactive vs. proactive jamming [35] as well as the joint use of EH and multi-carrier transmission against jamming attacks. Second, in the EH case, the study of more realistic models accounting for finite storage capabilities, asymmetries in the legitimate nodes' parameters and EH at the jammer side, are interesting future extensions. Moreover, the study of multi-user and multi-jammer interactions as well as games of incomplete information are challenging open issues.

VI. Conclusions

In this work, the adversarial interaction between a pair of legitimate nodes and a malicious jammer in a wireless secret key generation (SKG) framework was investigated. Two different counter-jamming approaches were proposed and studied. First, energy harvesting at the legitimate nodes, and, second, channel hopping vs. power spreading in block fading AWGN channels. In either approach, a zero-sum game was introduced as the objectives of the two parties involved were opposed. Complete characterizations of the Nash and Stackelberg equilibria in closed-form were provided in both cases. It was demonstrated that either approach may offer significant gains in utility, particularly in the low signal-to-interference ratio regime, in which counteracting the jamming interference becomes crucial. As a result, viable and low complexity alternatives for defending SKG systems may be developed by exploiting either novel transceiver features or available spectral resources.

Appendix

A. Proof of Proposition 1

Let us assume that the legitimate nodes neutralize the jammer by transmitting at power $p \in [0, \min\{p_{th}(\tau), P\}]$. The jammer observes player L's choice and from Remark 2, decides to stay silent. Notice that player L can force the jammer to remain silent by transmitting at $p \in [0, \min\{p_{th}(\tau) - \varepsilon, P\}]$ for an arbitrarily small $\varepsilon > 0$. For simplicity, $\varepsilon \simeq 0$ is assumed in the following.

The remaining question is: how will player L choose $\tau \in [0, 1)$ and $p \in [0, \min\{p_{th}(\tau), P\}]$ to maximize the resulting SKG utility

$$\tilde{u}(p, \tau, 0) = \frac{1-\tau}{2} \log_2 \left(1 + \frac{p\sigma_H^2}{2(1-\tau) + \frac{(1-\tau)^2}{p\sigma_H^2}} \right), \quad (25)$$

while ensuring that the jammer stays silent and cannot decrease the utility by transmitting with non-zero power? Since the feasible set of p depends on τ, we first have to find the maximum of $\tilde{u}(p, \tau, 0)$ w.r.t. p for any fixed τ. The function $\tilde{u}(p, \tau, 0)$ is strictly increasing in p and, hence, the optimal power is given by $\tilde{p}(\tau) = \min\{P, p_{th}(\tau)\}$. Now, we need to maximize $\tilde{u}(\tilde{p}(\tau), \tau, 0)$ w.r.t. $\tau \in [0, 1]$:

$$\tilde{u}(p_{th}(\tau), \tau, 0) = \frac{1-\tau}{2} \log_2 \left(1 + \frac{\zeta\tau\sigma_H^2}{(2 + \frac{1-\tau}{\zeta\tau\sigma_H^2})(1-\tau)} \right).$$

At the extremes $\tau = 0$ and $\tau \to 1$ the utility goes to zero. By investigating its second order derivatives w.r.t. τ, which amounts to the following quadratic equation:

$$(1-\tau)^2 - 2\sigma_H^4 \zeta^2 \tau^2 = 0, \tag{26}$$

it can be shown that $\tilde{u}(p_{th}(\tau), \tau, 0)$ always has an inflexion point in between $(0, 1)$ and starts as convex and then becomes concave. Knowing that the the utility is always positive, we can conclude that $\tilde{u}(p_{th}(\tau), \tau, 0)$ has a unique critical point that is the global maximizer $\tau^* \in (0, 1)$ and which is the solution to $\frac{d\tilde{u}(p_{th}(\tau),\tau,0)}{d\tau} = 0$. This implies that, if $p_{th}(\tau^*) \leq P$, then the optimal solution that neutralizes the jammer is $\tau^{NJ} = \tau^*$ and $p^{NJ} = p_{th}(\tau^*)$. If $p_{th}(\tau^*) > P$ (or equivalently $\tau^* > \tau_{th}(P)$), then the optimal solution that neutralizes the jammer is $p^{NJ} = P$ and $\tau^{NJ} = \tau_{th}(P) = \frac{P}{\zeta}$.

B. Proof of Theorem 1 and Theorem 2

From Remark 1, we know that transmitting at maximum power is a strictly dominant strategy for player L and, hence, $p^{NE} = P$. We first prove that at the NE, player L will not operate in EH mode for longer than the threshold $\tau_{th}(P)$. Let's suppose by absurdum that $\tau^{NE} > \tau_{th}(P)$, then the jammer's best response would be to remain silent $\gamma^{NE} = 0$ (as the energy harvested from the jammer in the EH phase is enough to overcome the interference inflicted by the jammer in the SKG phase). Then, the optimal τ^{NE} maximizing the utility $\tilde{u}(P, \tau, 0)$ (which is decreasing in τ) would be $\tau^{NE} \to \tau_{th}(P)$ obtaining the utility $\tilde{u}^{NE} \to \tilde{u}(P, \tau_{th}(P), 0)$. However, this state cannot be an NE. Indeed, if the jammer stays silent $\gamma^{NE} = 0$, no energy is harvested during τ^{NE} and player L gains in utility by deviating to $\tau = 0$. This will also cause the jammer to deviate to $\gamma = \Gamma$.

The above implies that player L can only choose an EH strategy such that $\tau^{NE} \leq \tau_{th}(P)$ at the NE. This condition is equivalent to $P \geq p_{th}(\tau^{NE})$, which means that the utility is either decreasing or simply a constant in γ (see Remark 2). This further implies that if the jammer uses maximum power $\gamma^{NE} = \Gamma$, then it cannot benefit by deviating unilaterally. Hence, to find the NE of the form (P, τ^{NE}, Γ), we only need to find the optimal value or values of $\tau \in [0, \tau_{th}(P)]$ that maximizes the function $\tilde{u}(P, \tau, \Gamma)$ given by:

$$\tilde{u}(P, \tau, \Gamma) = \frac{1-\tau}{2} \log_2 \left(1 + \frac{\left(\frac{P}{1-\tau} + \kappa\Gamma\right)\sigma_H^2}{2(1+\sigma^2\Gamma) + \frac{(1+\sigma^2\Gamma)^2}{\left(\frac{P}{1-\tau}+\kappa\Gamma\right)\sigma_H^2}} \right),$$

where κ depends on τ: $\kappa(\tau) = \frac{\zeta\tau\sigma^2}{1-\tau}$. At $\tau = 0$, this function is strictly positive $\tilde{u}(P, 0, \Gamma) > 0$ equal to the SKG capacity without EH and, when $\tau \to 1$ the function goes to 0. By investigating the second order derivative of $\tilde{u}(P, \tau, \Gamma)$ w.r.t. τ, which amounts to the analysis of the following quadratic equation

$$(1-\tau)^2(1+\sigma^2\Gamma)^2 - 2\sigma_H^4(P + \sigma^2\zeta\Gamma\tau)^2 = 0, \tag{27}$$

two different cases arise:

- *Case A:* If $1+\sigma^2\Gamma \geq \sqrt{2}\sigma_H^2 P$, this function has a unique inflexion point that lies in $(0, 1)$ and the function starts as convex and then becomes concave. Thus, $\tilde{u}(P, \tau, \Gamma)$ has a critical point that is a local maximum $\tau_{max} \in (0, 1)$, which is a solution of the equation $\frac{d\tilde{u}(P,\tau,\Gamma)}{d\tau} = 0$. Hence, the optimal strategy is given by either the maximal point τ_{max} or by one (or both) of the borders of the interval $[0, \tau_{th}(P)]$ depending on the system parameters:

$$\tau^{NE} = \arg \max_{\tau \in \{0, \min\{\tau_{th}(P), \tau_{max}\}\}} \tilde{u}(P, \tau, \Gamma). \tag{28}$$

- *Case B:* If $1 + \sigma^2\Gamma < \sqrt{2}\sigma_H^2 P$, then the function is always concave (and it does not have an inflexion point) in $(0, 1)$. If the function has a critical point in $(0, 1)$, then this critical point is a maximum point denoted by τ_{max} and $\tau^{NE} = \min\{\tau_{th}(P), \tau_{max}\}$. Otherwise, the function is concave decreasing and $\tau^{NE} = 0$.

Remark that, at least in theory, Case A can lead to the existence of two NEs whenever the additional equality condition is met: $\tilde{u}(P, 0, \Gamma) = \tilde{u}(P, \min\{\tau_{th}(P), \tau_{max}\}, \Gamma)$, i.e., when both borders of the interval $[0, \min\{\tau_{max}, \tau_{th}(P)\}]$ provide equal maximum utility. However, this can happen only in very special cases of the system parameters or with zero probability.

Aside from the zero probability case described above, this profile may not be the unique NE of the game $\tilde{\mathcal{G}}$ as there may exist other NEs such that $\gamma^{NE} \in [0, \Gamma)$. Such cases can only happen if the strategy of player L at the NE equals $(P, \tau_{th}(P))$ or equivalently if $(P, \tau_{th}(P), \Gamma)$ is the above NE. Otherwise, whenever $\tau^{NE} < \tau_{th}(P)$, the utility is strictly decreasing in γ and the only strategy of the jammer at the NE is Γ (case discussed previously).

Now, whenever the legitimate user chooses their strategy $(P, \tau_{th}(P))$, the jammer becomes totally indifferent between all their strategies and, in particular, all jamming powers in $[0, \Gamma)$ provide the same utility (see Remark 2). Hence, in this case, there may be other NEs aside from $(P, \tau_{th}(P), \Gamma)$ that provide identical utilities to $\tilde{u}(P, \tau_{th}(P), \Gamma)$. We can disregard the state $(P, \tau_{th}(P), 0)$ for the same reasons for which the NJ state is not a NE.

In order to find all NE of the form $(P, \tau_{th}(P), \gamma^{NE})$, we need to find all $\gamma^{NE} \in (0, \Gamma)$ such that the legitimate user cannot deviate from $(P, \tau_{th}(P))$ or it will lose in terms of utility. Stated otherwise, all $\gamma^{NE} \in (0, \Gamma)$ such that $\tau_{th}(P) = \arg\max_\tau \tilde{u}(P, \tau, \gamma^{NE})$ provide additional NE profiles of the form $(P, \tau_{th}(P), \gamma^{NE})$.

The analysis of the utility $\tilde{u}(P, \tau, \gamma^{NE})$ as a function of τ is very similar to $\tilde{u}(P, \tau, \Gamma)$ above. There are two cases in function of the system parameters.

- *Case A:* If $1 + \sigma^2\Gamma \geq \sqrt{2}\sigma_H^2 P$, for all $\gamma^{NE} \in \left[\frac{\sqrt{2}\sigma_H^2 P - 1}{\sigma^2}, \Gamma\right)$, the function $\tilde{u}(P, \tau, \gamma^{NE})$ has a unique inflexion point that lies in $(0, 1)$ and starts as convex and then becomes concave. Thus, $\tilde{u}(P, \tau, \gamma^{NE})$ has a critical point that is a local maximum $\tau_{max}(\gamma^{NE}) \in (0, 1)$, which is a solution of the equation $\frac{d\tilde{u}(P,\tau,\gamma^{NE})}{d\tau} = 0$. The additional conditions for the strategy $(P, \tau_{th}(P))$ to be optimal for player L are:

$$\begin{aligned} \tau_{th}(P) &\leq \tau_{max}(\gamma^{NE}) \\ \tilde{u}(P, 0, \gamma^{NE}) &\leq \tilde{u}(P, \tau_{th}(P), \gamma^{NE}) \end{aligned} \tag{29}$$

- *Case B:* If $1+\sigma^2\Gamma < \sqrt{2}\sigma_H^2 P$, for all $\gamma^{NE} \in \left(0, \frac{\sqrt{2}\sigma_H^2 P-1}{\sigma^2}\right)$, the function $\tilde{u}(P,\tau,\gamma^{NE})$ is always concave in $\tau \in (0,1)$. If the function has a critical point in $(0,1)$, then this critical point is a maximum point denoted by $\tau_{max}(\gamma^{NE})$. The additional condition for the strategy $(P,\tau_{th}(P))$ to be optimal is $\tau_{th}(P)=\tau_{max}(\gamma^{NE})$. Otherwise, the function is concave decreasing in τ and $(P,\tau_{th}(P))$ cannot be optimal for player L.

C. Proof of Theorem 3

Let us first find the best-response of the jammer defined in (17). Given the second remark, it is easy to see that:

$$\gamma^{BR}(p,\tau)=\begin{cases}0, & \text{if } p<p_{th}(\tau)\\ \in[0,\Gamma], & \text{if } p=p_{th}(\tau)\\ \Gamma, & \text{if } p>p_{th}(\tau),\end{cases} \tag{30}$$

where $p_{th}(\tau)=\zeta\tau$. Notice that whenever $p=p_{th}(\tau)$ the best response of the jammer can be anything and cannot in fact be predicted by player L. However, the obtained payoff is anticipated by player L as it does not depend on the actual choice of the jammer: $\tilde{u}(p_{th}(\tau),\tau,\gamma)=\tilde{u}(p_{th}(\tau),\tau,0)$, for all γ.

The SE action of the leader, anticipating that the jammer will best respond to their own choice is given by:

$$(p^{SE},\tau^{SE})=\arg\max_{p,\tau}\ \tilde{u}(p,\tau,\gamma^{BR}(p,\tau)) \tag{31}$$

From (30), we see that player L can either neutralize the jammer or allow it to transmit, knowing that the jammer will transmit with full power Γ. The situation that proves to be mostly advantageous to the legitimate player will be chosen.

- *Case A:* Assume the legitimate player neutralizes the jammer by choosing a strategy such that $p \le p_{th}(\tau)$. Player L has to find the best pair (p,τ) that maximizes $\tilde{u}(p,\tau,0)$ knowing that $p\in[0,\min\{P,p_{th}(\tau)\}]$ and that $\tau\in[0,1]$; the solution equals (p^{NJ},γ^{NJ}) in Proposition 1.

- *Case B:* Assume now that the legitimate player does not neutralize the jammer and $p \ge p_{th}(\tau)$. Player L has to find the best pair (p,τ) that maximizes $\tilde{u}(p,\tau,\Gamma)$ knowing that $p\in[0,P]\cap[p_{th}(\tau),\infty)$ and $\tau\in(0,1)$. By fixing τ first and optimizing with respect to p, we have that $u(p,\tau,\Gamma)$ is increasing in p and hence, the optimal power equals P and the value of τ will be constrained by $P\ge p_{th}(\tau)$ or equivalently $\tau\le\tau_{th}(P)$. This analysis is identical to the analysis of the NE and one possible SE solution is the NE in Theorem 1.

At the SE, the legitimate user will choose one of the two possibilities which provides a higher SKG utility. From Proposition 2, we know that the NJ state cannot provide a strictly higher utility than the NE state. Hence, whenever $\tau^{NE}<\tau_{th}(P)$, the utility of the unique NE is strictly higher than that of the NJ state. This implies a unique SE that is identical to the NE. If $\tau^{NE}=\tau_{th}(P)$, this means that $(p^{NJ},\tau^{NJ})=(P,\tau^{NE})$ which implies that the utilities at both states NJ and NE are identical. Both the NE (in Theorem 1) and NJ (in Proposition 1) states are SE solutions: $(P,\tau_{th}(P),\Gamma)$ and $(P,\tau_{th}(P),0)$.

The remaining question is whether there exist other solutions when player L chooses the strategy $(p^{SE},\tau^{SE})=(P,\tau_{th}(P))$. In this case, the jammer is rendered indifferent between all of its actions $\gamma\in[0,\Gamma]$, which means that it is also rendered unpredictable. As opposed to the NE, the SE requires the legitimate user to be able to anticipate precisely the jammer's response. To avoid this problem, the leader can silence the jammer by transmitting with power $p=P-\varepsilon$ or ensures that the jammer transmits with full power by transmitting at power $p=P+\varepsilon$, where ε could be made arbitrarily small and, hence, has no practical impact. None of the other NE in Theorem 2 can be SEs, since the jammer's response cannot be predictable.

In conclusion, if $\tau^{NE}<\tau_{th}(P)$, then the SE is unique and identical to the NE in Theorem 2. Otherwise, both the NE and the NJ states are SE solutions.

D. Proof of Proposition 3

Assume by absurdum and WLOG that player J has an NE strategy such that the first channel is left unused $\beta^*=(0,\beta_2,\ldots,\beta_{N+1})$ while other channels are used $\beta_i>0$ for some $2\le i\le N$. Exploiting this knowledge, player L will only employ channel hopping on channel 1 and maybe spreading with non zero probability at the NE. To see this, we write the expected payoff of player L assuming $\beta_1=0$

$$\begin{aligned}&2N\hat{u}(\alpha^*,\beta^*)\\ &=\sum_{i=2}^{N}\{\alpha_i(1-\beta_i-\beta_{N+1})C(NP,0)\\ &\quad+\alpha_i\beta_i C(NP,N\Gamma)+\alpha_i\beta_{N+1}C(NP,\Gamma)\}\\ &\quad+\alpha_{N+1}(1-\beta_{N+1})[(N-1)C(P,0)+C(P,N\Gamma)]\\ &\quad+\alpha_{N+1}\beta_{N+1}NC(P,\Gamma)\alpha_1(1-\beta_{N+1})(N-1)C(NP,0).\end{aligned}$$

Since $C(NP,0)>C(NP,N\Gamma)$ and there exists some $\beta_i>0$, we have that:

$$\begin{aligned}&(1-\beta_{N+1})(N-1)C(NP,0)\\ &\qquad>\sum_{i\neq1}[\beta_i C(NP,N\Gamma)+(1-\beta_i-\beta_{N+1})C(NP,0)].\end{aligned}$$

This means that, if the jammer does not use channel 1, the legitimate ndes will only employ this channel and none of the other channel hopping strategies and the NE will be of the form $\alpha^*=(1-\alpha_{N+1},0,\ldots,0,\alpha_{N+1})$. The utility becomes:

$$\begin{aligned}&2N\hat{u}(\alpha^*,\beta^*)\\ &\quad=(1-\alpha_{N+1})(1-\beta_{N+1})(N-1)C(NP,0)\\ &\qquad+\alpha_{N+1}(1-\beta_{N+1})[(N-1)C(P,0)+C(P,N\Gamma)]\\ &\qquad+\alpha_{N+1}\beta_{N+1}NC(P,\Gamma).\end{aligned}$$

But now, if the jammer uses all channel hopping probabilities back in channel 1, he can strictly decrease the utility. Assume that the jammer switches from the initial β^* to $\delta=(1-\beta_{N+1},0,\ldots,0,\beta_{N+1})$. The payoff becomes:

$$\begin{aligned}&2N\hat{u}(\alpha^*,\delta)\\ &\quad=(1-\alpha_{N+1})(1-\beta_{N+1})(N-1)C(NP,N\Gamma)\\ &\qquad+\alpha_{N+1}(1-\beta_{N+1})[(N-1)C(P,0)+C(P,N\Gamma)]\\ &\qquad+\alpha_{N+1}\beta_{N+1}NC(P,\Gamma).\end{aligned}$$

Since $\hat{u}(\alpha^*, \beta^*) > \hat{u}(\alpha^*, \delta)$, the jammer has an incentive to deviate from the NE which is a contradiction. Thus, the jammer uses either all or none of the channel hopping actions. For player L, the proof follows similarly.

E. Proof of Proposition 4

Let us write the linear equations obtained when the players are indifferent among their channel hopping actions. There are four very similar cases depending on whether the players use spread with zero probability at the NE or not. We only detail one case here below. If both players use spread at the NE, the following conditions must be met:

$$\begin{aligned}&\alpha_i C(NP, N\Gamma) + (1 - \alpha_i - \alpha_{N+1})C(NP, 0)\\ &\qquad + \alpha_{N+1}[(N-1)C(P,0) + C(P, N\Gamma)] = c_\alpha,\\ &\beta_i C(NP, N\Gamma) + (1 - \beta_i - \beta_{N+1})C(NP, 0)\\ &\qquad + \beta_{N+1} C(NP, \Gamma) = c_\beta.\end{aligned}$$

The equations in α illustrate that player J becomes indifferent among their pure channel hopping actions at the NE. Similarly, the equations in β make player L indifferent among their pure channel hopping actions at the NE. We remark that all these equations are identical in the sense that their coefficients do not depend on the index i of the α and β variables. This means that their solutions are of the form: $\alpha_i = a$ and $\beta_i = b$ for any $i \leq N$. Therefore – irrespective of whether the players employ or not spreading at the NE – if both players employ the channel hopping strategy, then the NE takes on the special form $\alpha^* = (a, \ldots, a, (1 - Na))$, $\beta^* = (b, \ldots, b, (1 - Nb))$ for some $0 \leq a \leq 1/N$, $0 \leq b \leq 1/N$.

F. Proof of Theorem 4

If $N = 1$, the NE analysis is trivial and both players transmit at full powers $(NP, N\Gamma)$. If $N > 1$ and given the strict convexity of $C(p, \gamma)$ in γ, we have the following inequality for all p, $\gamma_1 \neq \gamma_2$ and $\lambda \in (0, 1)$:

$$C(p, \lambda\gamma_1 + (1-\lambda)\gamma_2) < \lambda C(p, \gamma_1) + (1-\lambda)C(p, \gamma_2).$$

By taking $p = P$, $\gamma_1 = 0, \gamma_2 = N\Gamma, \lambda = \frac{N-1}{N}$, we obtain:

$$NC(P, \Gamma) < (N-1)C(P, 0) + C(P, N\Gamma) \tag{32}$$

Similarly, by taking $p = NP$, $\gamma_1 = 0, \gamma_2 = N\Gamma, \lambda = \frac{N-1}{N}$, we obtain:

$$NC(NP, \Gamma) < (N-1)C(NP, 0) + C(NP, N\Gamma). \tag{33}$$

From Proposition 3 and Proposition 4, the NE can only take nine forms which are not mutually exclusive. Each case is studied using Definition 1 and for which necessary and sufficient conditions are provided. Then, using (32) and (33), we show that only three of the nine cases can occur. The proof is rather long and tedious and only a sketch containing the main ideas is provided below. 1) *Both players spread at the NE* (i.e., $\alpha^* = \beta^* = e_{N+1}$), iff $C(NP, \Gamma) < NC(P, \Gamma)$ and $(N-1)C(P, 0) + C(P, N\Gamma) > NC(P, \Gamma)$. The second condition is always true due to (32).

2) *Both players use only channel hopping at the NE* (i.e., $\alpha^* = \beta^* = (1/N, \ldots, 1/N, 0)$), iff $C(NP, N\Gamma) + (N-1)C(NP, 0) > N(N-1)C(P, 0) + NC(P, N\Gamma)$ and $C(NP, N\Gamma) + (N-1)C(NP, 0) < NC(NP, \Gamma)$. This case is impossible because of (33).

3) *The game has a strictly mixed NE*, i.e., all actions are used with non-zero probability, of the form $\alpha^* = (a, \ldots, a, (1 - Na))$, $\beta^* = (b, \ldots, b, (1 - Nb))$ iff there exist $0 < a < 1/N$ and $0 < b < 1/N$ such that both players are indifferent among all their pure strategies. Let us write the condition for $(a, \ldots, a, 1 - Na)$ to be a NE and for which the jammer is indifferent among their pure strategies by Definition 1. This yields the following linear equation:

$$\begin{aligned}&a[NC(NP, \Gamma) - C(NP, N\Gamma) - (N-1)C(NP, 0)]\\ &\qquad = (1 - Na)[(N-1)C(P,0) + C(P, N\Gamma) - NC(P, \Gamma)],\end{aligned}$$

where the term on the LHS is a strictly negative value from $a > 0$ and (33) and the RHS is a strictly positive value from $a < 1/N$ and (32). Thus, this case can never occur.

4) *Player L only channel hops and player J uses both channel hopping and spreading at the NE*: $\alpha^* = (1/N, \ldots, 1/N, 0)$ and $\beta^* = (b, \ldots, b, (1 - Nb))$, iff $C(NP, N\Gamma) + (N-1)C(NP, 0) = NC(NP, \Gamma)$, $0 < b < 1/N$, and $Nb[(N-1)C(P,0) + C(P, N\Gamma)] + (1 - Nb)NC(P, \Gamma) < bC(NP, N\Gamma) + (N-1)bC(NP, 0) + (1 - Nb)C(NP, \Gamma)$, where b is chosen such that player L is indifferent among their pure strategies. Given (33), the above equality never holds.

5) *Player J only channel hops and player L uses both channel hopping and spreading at the NE* (i.e., $\alpha^* = (a, \ldots, a, (1 - Na))$ and $\beta^* = (1/N, \ldots, 1/N, 0)$), iff $C(NP, N\Gamma) + (N-1)C(NP, 0) = N(N-1)C(P, 0) + C(P, N\Gamma)$, $0 < a < 1/N$, and $MaC(NP, \Gamma) + (1 - Na)NC(P, \Gamma) > aC(NP, N\Gamma) + (N-1)aC(NP, 0) + (1 - Na)[(N-1)C(P, 0) + C(P, N\Gamma)]$ where a is chosen such that player J is indifferent among their pure strategies. The last inequality condition becomes:

$$\begin{aligned}&a[NC(NP, \Gamma) - C(NP, N\Gamma) - (N-1)C(NP, 0)]\\ &\qquad > (1 - Na)[(N-1)C(P,0) + C(P, N\Gamma) - NC(P, \Gamma)]\end{aligned}$$

where the term on the LHS is a strictly negative value from $a > 0$ and (33) and the RHS is a strictly positive value from $a < 1/N$ and (32). Thus, this case can never occur.

6) *Player L spreads and player J channel hops at the NE* (i.e., $\alpha^* = e_{N+1}$ and $\beta^* = (\beta_1, \ldots, \beta_N, 0)$), iff $NC(P, \Gamma) > (N-1)C(P, 0) + C(P, N\Gamma)$, $NC(NP, 0) - N(N-1)C(P, 0) - NC(P, N\Gamma) < C(NP, 0) - C(NP, N\Gamma)$ and β_i meet some additional constraints. Because of (32) this case never occurs as the first condition is never satisfied.

7) *Player J spreads and player L channel hops at the NE* (i.e., $\beta^* = e_{N+1}$ and $\alpha^* = (\alpha_1, \ldots, \alpha_N, 0)$), iff $C(NP, \Gamma) > NC(P, \Gamma)$ and $NC(NP, 0) - NC(NP, \Gamma) > C(NP, 0) - C(NP, N\Gamma)$. The NE strategies of player L are given by the (infinite number) of solutions to the following system of linear inequalities:

$$\begin{cases} 0 \leq \alpha_i \leq 0, \ \forall i, \ \sum_{j=1}^{N} \alpha_j = 1 \\ \alpha_i < \frac{C(NP,0) - C(NP,\Gamma)}{C(NP,0) - C(NP,N\Gamma)}, \ \forall i \leq N. \end{cases}$$

The second condition is always true (33). From (33), the above system of inequality always has the uniform probability over the channels solution $\alpha^* = (1/N, \ldots, 1/N, 0)$.

8) *Player L spreads and player J employs all their actions at the NE* (i.e., $\alpha^* = e_{N+1}$, $\beta^* = (\beta_1, \ldots, \beta_{N+1})$), iff $(N-1)C(P,0) + C(P, N\Gamma) = NC(P,\Gamma)$ and $\beta_i, \forall i$ meet some additional constraints that are not detailed here. The reason is that, given (32), the equality condition never holds and, hence, this case is impossible.

9) *Player J spreads and player L employs all their actions at the NE* (i.e., $\beta^* = e_{N+1}$ and $\alpha^* = (\alpha_1, \ldots, \alpha_N, \alpha_{N+1})$), iff $C(NP,\Gamma) = NC(P,\Gamma)$ and the solutions to the following linear system of inequalities are NE strategies for player L:

$$\begin{cases} 0 \le \alpha_i \le 1, \ \forall i, \ \sum_{j=1}^{N} \alpha_j = 1 \\ \alpha_i[C(NP, N\Gamma) - C(NP,0)] + \alpha_{N+1}[(N-1)C(P,0) \\ +C(P, N\Gamma) - C(NP,0) + C(NP,\Gamma) - NC(P,\Gamma)] \\ > C(NP,\Gamma) - C(NP,0), \ \forall i \le N. \end{cases}$$

Notice that, by taking $\alpha_{N+1} = 0$, the above system of linear equations is precisely the one in case 7 which has an infinite number of solutions, and in particular $\alpha_i = 1/N, \ \forall i \le N$. Similarly, $\alpha_i = 0$ for all $i \le N$ and $\alpha_{N+1} = 1$ (player L spreads) is also a solution, which follows directly from (32).

G. Proof of Proposition 5

The best-response for the jammer is defined as $\beta^{BR}(\alpha) = \arg\min_\beta \hat{u}(\alpha, \beta)$, where $\beta^{BR}(\alpha)$ represents the best action the jammer can take knowing that the legitimate player choses α. The payoff is affine in β and can be rewritten it as $\hat{u}(\alpha,\beta) = \sum_{i=1}^{N+1} \beta_i c_i(\alpha) + c_0(\alpha)$, with the coefficients:

$$\begin{aligned} c_i(\alpha) &= \alpha_i[C(NP, N\Gamma) - C(NP,0)] \\ &\quad + \alpha_{N+1}[C(P, N\Gamma) - C(P,0)], \quad i \le N, \\ c_{N+1}(\alpha) &= \sum_{j=1}^{N} \alpha_i[C(NP,\Gamma) - C(NP,0)] \\ &\quad + N\alpha_{N+1}[C(P,\Gamma) - C(P,0)], \\ c_0(\alpha) &= \sum_{j=1}^{N} \alpha_i C(NP,0) + N\alpha_{N+1} C(P,0). \end{aligned} \tag{34}$$

Thus, we observe that to find the best-response function $\beta^{BR}(\alpha)$, the jammer has to solve a linear program under the constraints: $\beta_i \ge 0, \ \forall\, i$ and $\sum_{j=1}^{N+1} \beta_j = 1$. The SE action of the leader, anticipating that the jammer will best respond to their own choice is given by:

$$\begin{aligned} \alpha^{SE} &= \arg\max_\alpha \ \hat{u}(\alpha, \beta^{BR}(\alpha)) \\ &= \arg\max_\alpha \ \left\{ \min_{j>0} c_j(\alpha) + c_0(\alpha) \right\}. \end{aligned}$$

Player L can anticipate the response of the jammer, who seeks to minimize the coefficients $c_j(\alpha)$. We remark that: $c_{N+1}(\alpha) = (1-\alpha_{N+1})[C(NP,\Gamma) - C(NP,0)] + N\alpha_{N+1}[C(P,\Gamma) - C(P,0)]$ and $c_0(\alpha) = (1-\alpha_{N+1})C(NP,0) + N\alpha_{N+1}$ do not depend on the way in which the load $1-\alpha_{N+1}$ is spread over the channel hopping actions. Therefore we can only focus on $c_i(\alpha), \ 1 \le i \le N$.

If player L uses channel hopping strategies with uniform probability $\alpha^{(1)} = (a, \ldots, a, 1-Na)$, all coefficients will be equal $c_i(\alpha^{(1)}) = a[C(NP, N\Gamma) - C(NP,0)] + (1-Na)[C(P, N\Gamma) - C(P,0)]$. This means that the jammer is indifferent between the different channels $\min_{1\le j\le N} c_j(\alpha) = a[C(NP, N\Gamma) - C(NP,0)] + (1-Na)[C(P, N\Gamma) - C(P,0)]$.

Now, if player L has a preference for a certain channel, say for channel 1: $\alpha^{(2)} = (a+\delta_1, a-\delta_2, \ldots, a-\delta_N, 1-Na)$, with $\sum_{j=2}^{N} \delta_j = \delta_1 > 0$, the coefficients will be: $c_1(\alpha^{(2)}) = (a+\delta)[C(NP, N\Gamma) - C(NP,0)] + (1-Na)[C(P, N\Gamma) - C(P,0)]$, $c_i(\alpha^{(2)}) = (a-\delta_i)[C(NP, N\Gamma) - C(NP,0)] + (1-Na)[C(P, N\Gamma) - C(P,0)]$. In this case, the jammer will profit from this information and will put all their channel hopping load on channel 1 alone: $\beta_1^{BR}(\alpha^{(2)}) = 1 - \beta_{N+1}^{BR}(\alpha^{(2)})$, $\beta_i(\alpha^{(2)}) = 0, \forall 2 \le N$ and $\min_{1\le j\le N} c_j(\alpha^{(2)}) = (a+\delta)[C(NP, N\Gamma) - C(NP,0)] + (1-Na)[C(P, N\Gamma) - C(P,0)]$. But this means that $\min_{1\le j\le N} c_j(\alpha^{(2)}) < \min_{1\le j\le N} c_j(\alpha^{(1)})$, which further implies that $\hat{u}(\alpha^{(1)}, \beta^{BR}(\alpha^{(1)})) < \hat{u}(\alpha^{(2)}, \beta^{BR}(\alpha^{(2)}))$. This means that player L will lose in utility by not assigning uniform probability to the channel hopping strategies.

H. Proof of Theorem 5

Proposition 1 tells us that the SE strategy of player L is of the form: $\alpha^{SE} = (a, \ldots, a, (1-Na))$ for some $a \in [0, 1/N]$, which is to be determined. The coefficients in (34) become:

$$\begin{aligned} c_i(\alpha^{SE}) &= a[C(NP, N\Gamma) - C(NP,0)] \\ &\quad + (1-Na)[C(P, N\Gamma) - C(P,0)], \quad i \le N \\ c_{N+1}(\alpha^{SE}) &= Na[C(NP,\Gamma) - C(NP,0)] \\ &\quad + N(1-Na)[C(P,\Gamma) - C(P,0)]. \end{aligned}$$

Using the fact that $C(p,\gamma)$ is convex w.r.t. γ for a fixed p, we have the following inequalities: $NC(P,\Gamma) < (N-1)C(P,0) + C(P, N\Gamma)$ and $NC(NP,\Gamma) < (N-1)C(NP,0) + C(NP, N\Gamma)$ which imply that $c_i(\alpha^{SE}) < c_{N+1}(\alpha^{SE})$. This means that the jammer's strategy is to spread always: $\beta^{SE} = (0, \ldots, 0, 1)$. The SE utility becomes:

$$\hat{u}(\alpha^{SE}, \beta^{SE}) = aC(NP,\Gamma) + (1-Na)C(P,\Gamma). \tag{35}$$

This implies that, if $C(NP,\Gamma) > NC(P,\Gamma)$ player L will only channel hop with uniform probability $a = 1/N$. If $C(NP,\Gamma) < NC(P,\Gamma)$ player L will only spread $a = 0$. If $C(NP,\Gamma) = NC(P,\Gamma)$ then the legitimate user is indifferent between spreading and channel hopping and all $a \in [0, 1/N]$ are solutions.

REFERENCES

[1] E. V. Belmega and A. Chorti, "Energy harvesting in secret key generation systems under jamming attacks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017.

[2] A. Chorti and E. V. Belmega, "Secret key generation in Rayleigh block fading AWGN channels under jamming attacks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–7.

[3] R. Ahlswede and I. Csiszár, "Common randomness in information theory and cryptography—Part I: Secret sharing," *IEEE Trans. Inf. Theory*, vol. 39, no. 4, pp. 1121–1132, Jul. 1993.

[4] U. M. Maurer, "Secret key agreement by public discussion from common information," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 733–742, May 1993.

[5] R. Ahlswede and I. Csiszár, "Common randomness in information theory and cryptography—Part II: CR capacity," *IEEE Trans. Inf. Theory*, vol. 44, no. 1, pp. 225–240, Jan. 1998.

[6] C. Bennett, G. Brassard, C. Crépeau, and U. Maurer, "Generalized privacy amplification," *IEEE Trans. Inf. Theory*, vol. 41, no. 6, pp. 1915–1923, Nov. 1995.

[7] C. Ye, A. Reznik, and Y. Shah, "Extracting secrecy from jointly Gaussian random variables," in *Proc. Int. Symp. Inf. Theory (ISIT)*, Seattle, WA, USA, Jul. 2006, pp. 2593–2597.

[8] C. Ye, S. Mathur, A. Reznik, Y. Shah, W. Trappe, and N. B. Mandayam, "Information-theoretically secret key generation for fading wireless channels," *IEEE Trans. Inf. Forensics Security*, vol. 5, no. 2, pp. 154–240, Jun. 2010.

[9] T.-H. Chou, S. C. Draper, and A. M. Sayeed, "Key generation using external source excitation: Capacity, reliability, and secrecy exponent," *IEEE Trans. Inf. Theory*, vol. 58, no. 4, pp. 2455–2474, Apr. 2012.

[10] A. Mukherjee, S. A. A. Fakoorian, J. Huang, and A. L. Swindlehurst, "Principles of physical layer security in multiuser wireless networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 3, pp. 1550–1573, Aug. 2014.

[11] U. Maurer and S. Wolf, "Secret-key agreement over unauthenticated public channels—Part III: Privacy amplification," *IEEE Trans. Inf. Theory*, vol. 49, no. 4, pp. 839–851, Apr. 2003.

[12] V. Yakovlev, V. Korzhik, and G. Morales-Luna, "Key distribution protocols based on noisy channels in presence of an active adversary: Conventional and new versions with parameter optimization," *IEEE Trans. Inf. Theory*, vol. 54, no. 6, pp. 2535–2549, Jun. 2008.

[13] C. Saiki and A. Chorti, "A novel physical layer authenticated encryption protocol exploiting shared randomness," in *Proc. IEEE Conf. Commun. Netw. Secur. (CNS)*, Sep. 2015, pp. 113–118.

[14] M. Zafer, D. Agrawal, and M. Srivatsa, "Limitations of generating a secret key using wireless fading under active adversary," *IEEE/ACM Trans. Netw.*, vol. 20, no. 5, pp. 1440–1451, Oct. 2012.

[15] R. Molière, F. Delaveau, C. L. K. Ngassa, C. Lemenager, T. Mazloum, and A. Sibille, "Tag signals for early authentication and secret key generation in wireless public networks," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Jun. 2015, pp. 108–112.

[16] G. T. Amariucai, S. Wei, and R. Kannan, "Gaussian jamming in block-fading channels under long term power constraints," in *Proc. Int. Symp. Inf. Theory (ISIT)*. Nice, France Jun. 2007, pp. 1001–1005.

[17] X. Song, P. Willett, S. Zhou, and P. Luh, "The MIMO radar and jammer games," *IEEE Trans. Signal Process.*, vol. 60, no. 2, pp. 687–699, Feb. 2012.

[18] S. Wei, R. Kannan, V. Chakravarthy, and M. Rangaswamy, "CSI usage over parallel fading channels under jamming attacks: A game theory study," *IEEE Trans. Commun.*, vol. 60, no. 4, pp. 1167–1175, Apr. 2012.

[19] R. El-Bardan, S. Brahma, and P. Varshney, "Strategic power allocation with incomplete information in the presence of a jammer," *IEEE Trans. Commun.*, vol. 64, no. 8, pp. 3467–3479, Aug. 2016.

[20] L. Xiao, T. Chen, J. Liu, and H. Dai, "Anti-jamming transmission Stackelberg game with observation errors," *IEEE Commun. Lett.*, vol. 19, no. 6, pp. 949–952, Jun. 2015.

[21] J. Guo, N. Zhao, F. R. Yu, X. Liu, and V. C. M. Leung, "Exploiting adversarial jamming signals for energy harvesting in interference networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 2, pp. 1267–1280, Feb. 2017.

[22] H. Fang, L. Xu, and K.-K. R. Choo, "Stackelberg game based relay selection for physical layer security and energy efficiency enhancement in cognitive radio networks," *Appl. Math. Comput.*, vol. 296, pp. 153–167, Mar. 2017.

[23] A. Mukherjee and A. L. Swindlehurst, "Jamming games in the MIMO wiretap channel with an active eavesdropper," *IEEE Trans. Signal Process.*, vol. 61, no. 1, pp. 82–91, Jan. 2013.

[24] M. Bloch and J. Barros, *Physical-Layer Security: From Information Theory to Security Engineering*. Cambridge, U.K: Cambridge Univ. Press, 2011.

[25] M. K. Simon, J. K. Omura, R. A. Scholtz, and B. K. Levitt, *Spread Spectrum Communications Handbook. Electronic*. New York, NY, USA: McGraw-Hill, 2002.

[26] R. Poisel, *Modern Communications Jamming Priniples and Techniques*. Norwood, MA, USA: Artech House, 2003.

[27] M. Strasser, C. Pöpper, S. Čapkun, and M. Čagalj, "Jamming-resistant key establishment using uncoordinated frequency hopping," in *Proc. IEEE Symp. Secur. Privacy*, May 2008, pp. 64–78.

[28] C. Pöpper, M. Strasser, and S. Čapkun, "Anti-jamming broadcast communication using uncoordinated spread spectrum techniques," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 5, pp. 703–715, Jun. 2010.

[29] R. Rajesh, V. Sharma, and P. Viswanath, "Capacity of Gaussian channels with energy harvesting and processing cost," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2563–2575, May 2014.

[30] Y. Gu and S. Aïssa, "RF-based energy harvesting in decode-and-forward relaying systems: Ergodic and outage capacities," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6425–6434, Nov. 2015.

[31] S. Ulukus *et al.*, "Energy harvesting wireless communications: A review of recent advances," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 3, pp. 360–381, Mar. 2015.

[32] D. Fudenberg and J. Tirole, *Game Theory*. Cambridge, MA, USA: MIT Press, 1991.

[33] A. Goldsmith, *Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[34] J. V. Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*. Princeton, NJ, USA: Princeton Univ. Press, 2007.

[35] J. Xu, L. Duan, and R. Zhang, "Proactive eavesdropping via jamming for rate maximization over Rayleigh fading channels," *IEEE Wireless Commun. Lett.*, vol. 5, no. 1, pp. 80–83, Feb. 2016.

E. Veronica Belmega (S'08–M'10) received the M.Sc. (engineer diploma) degree from the University Politehnica of Bucharest, Romania, in 2007, the M.Sc. and Ph.D. degrees from the Université Paris-Sud 11, France, in 2007 and 2010, respectively. From 2010 to 2011, she was a Post-Doctoral Researcher in a joint project between Princeton University, USA, and the Alcatel-Lucent Chair on Flexible Radio in Supélec, France. She is currently an Assistant Professor with ETIS, UMR 8051, Université Paris Seine, Université Cergy-Pontoise, ENSEA, CNRS, France, and Inria. She was one of the ten recipients of the L'Oréal-UNESCO-French Academy of Science Fellowship: For young women doctoral candidates in science in 2009.

Arsenia Chorti (S'00–M'05) received the M.Eng. degree in electrical and electronics engineering from the University of Patras, Greece, the D.E.A. degree in electronics from the University Pierre et Marie Curie-Paris VI, France, and the Ph.D. degree in electrical engineering from Imperial College London, U.K. She holds the post-doctoral positions with the Universities of Southampton, U.K., TCU, Greece, and UCL, U.K., from 2005 to 2008. She has served as a Senior Lecturer in communications with Middlesex University, U.K., from 2008 to 2010. From 2010 to 2013, she was a Marie Curie IOF Researcher with Princeton University, USA, and ICS-FORTH, Greece. Since 2013, she has been holding a lecturer position in communications and networks with CSEE University Essex, U.K. She is currently a Visiting Research Collaborator with Princeton University.

Mitigating Jamming Attacks Using Energy Harvesting

Gada Rezgui, E. Veronica Belmega, *Member, IEEE*, and Arsenia Chorti, *Member, IEEE*

***Abstract*—The use of energy harvesting as a counter-jamming measure is investigated on the premise that part of the harmful interference can be harvested to increase the transmit power. We formulate the strategic interaction between a pair of legitimate nodes and a malicious jammer as a zero-sum game. Our analysis demonstrates that the legitimate nodes are able to neutralize the jammer. However, this policy is not necessarily a Nash equilibrium and hence is sub-optimal. Instead, harvesting the jamming interference can lead to relative gains of up to 95%, on average, in terms of Shannon capacity, when the jamming interference is high.**

***Index Terms*—Energy harvesting, jamming, game theory.**

I. Introduction

IN RECENT years, the simultaneous wireless information and power transfer has gained momentum in the realm of energy harvesting (EH) technologies. In this contribution, we focus on systems employing a time-splitting EH approach [1]–[4], i.e., during a first phase the receiver collects energy from RF microwave radiation and in a second phase it uses the harvested energy for data transfer.

An interesting application of such EH approaches arises for wireless systems under jamming attacks. In the past, two main counter-jamming approaches have been commonly considered: direct sequence spread spectrum (DSSS) and frequency hopping spread spectrum (FHSS), [5], [6]. More recently, the use of multiple antennas has been exploited in [7] against both jamming and eavesdropping. In the first two approaches, the impact of power constrained jammers can be limited by increasing the spectral resources because the optimal jamming strategy is to spread the jamming power over the entire bandwidth; whereas the former requires an increased number of antennas. In this letter, we alternatively explore the possibility of mitigating jamming attacks by using EH *without increasing the spectral or the spatial resources*. So far, there has only been a limited number of contributions in this area [8]–[11].

In [8], the jamming interference is harvested and exploited in a two-way channel assuming that the jammer's policy is fixed (i.e., the jammer is not strategic). Furthermore, in [9] a cooperative relay wiretap channel is studied, in which the helping nodes harvest energy from the legitimate link and then generate interference to the eavesdropper. On the other hand, in [11], it is shown that EH can be exploited to mitigate jamming attacks in wireless secret key generation (SKG) systems and that it is possible to neutralize the jammer, i.e., to fully compensate its impact on the SKG rate. Building on this idea, the objective of this letter is to investigate whether EH at the legitimate transmitter can be an efficient measure against jamming attacks. We investigate the strategic interaction between a pair of legitimate nodes and a jammer employing as utility function the Shannon capacity.

The main contributions of this letter can be summarized as follows. First, we demonstrate that the jamming attack can be prevented entirely by adjusting the EH duration, i.e., the jammer can be neutralized (or forced to remain silent). This is only possible when the quality of the channel in the harvesting link is higher than in the jamming link. Nevertheless, neutralizing the jammer imposes too stringent restrictions on the EH duration and on the legitimate transmit power and hence is not optimal.

Second, we formulate a zero-sum game between the legitimate users and the jammer and derive the Nash equilibrium (NE) analytically. At the NE, both players transmit at full power, while, the optimal EH duration depends on the system parameters. Interestingly, we show that the NE always outperforms neutralizing the jammer. At the NE, the jamming interference is not fully cancelled but rather exploited, particularly efficient in the high jamming interference regime.

This letter represents a proof of concept of the potential use of EH against jamming attacks, relying on widely used system model assumptions [1]–[3], [5], [6], [10], [11]. Demonstration of the proposed EH policy in a real testbed is left as future work, in which the effect of imperfect channel estimation, type I and II jamming detection errors, implementation aspects of EH, etc., will also be considered.

II. System Model

The system model, depicted in Fig. 1, comprises three nodes: a legitimate transmitter, Alice, its intended receiver, Bob, and a malicious jammer, Jay. The channel coefficients in the links Alice-Bob, Jay-Alice and Jay-Bob are denoted by H, G_A and G_B, respectively and model fading; they are assumed to remain constant during each EH and transmission cycle and to change independently from one cycle to the next. We assume that full channel state information is available at all nodes.

When Alice sends a message X_A to Bob, Jay can jam the transmission. Bob's observation Y_B can be expressed as:

$$Y_B = HX_A + G_B X_J + Z_B, \tag{1}$$

Manuscript received July 17, 2018; revised August 23, 2018 and September 11, 2018; accepted September 13, 2018. Date of publication September 19, 2018; date of current version February 19, 2019. The associate editor coordinating the review of this paper and approving it for publication was X. Zhou. *(Corresponding author: E. Veronica Belmega.)*

The authors are with ETIS, ENSEA, CNRS, Université Paris Seine, Université Cergy-Pontoise, 95014 Cergy-Pontoise, France (e-mail: gada.rezgui@ensea.fr; belmega@ensea.fr; arsenia.chorti@ensea.fr).

Digital Object Identifier 10.1109/LWC.2018.2871152

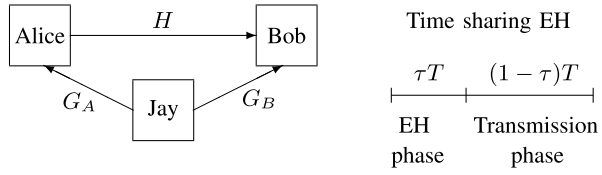


Fig. 1. System model and time sharing scheme at Alice's side.

where the message $X_A \sim \mathcal{N}(0, p)$ is drawn from a Gaussian codebook under a short-term power constraint $0 \leq p \leq P$. Finally, $Z_B \sim \mathcal{N}(0, N_B)$ models the effect of the additive white Gaussian noise (AWGN) in the Alice-Bob link. We consider that Jay transmits Gaussian jamming signals $X_J \sim \mathcal{N}(0, \gamma)$ and $0 \leq \gamma \leq \Gamma$ represents the jamming power.

This letter studies whether EH can be exploited to harvest the jamming interference and boost the transmit power. We assume a time sharing scheme with two phases: the first phase of duration τT, where T is the symbol period, is dedicated to EH. The second phase of duration $(1-\tau)T$ is dedicated to information transmission.

Alice's observation during the EH phase is then given by:

$$Y_A = G_A X_J + Z_A, \tag{2}$$

where $Z_A \sim \mathcal{N}(0, N_A)$ models the effect of AWGN noise in the link Jay-Alice. As commonly assumed [2], the harvested energy is proportional to the energy of the received signal:

$$E = \tau T \zeta \mathbb{E}\Big[|Y_A|^2\Big] = \tau T \zeta \Big(\gamma |G_A|^2 + N_A\Big), \tag{3}$$

where $\zeta \in [0, 1]$ is the harvesting efficiency parameter. The average power harvested during the EH phase is used in the information transmission phase and is given as follows:

$$p^{EH} = \frac{\tau}{1-\tau} \zeta \Big(\gamma |G_A|^2 + N_A\Big). \tag{4}$$

The multiplicative term $\frac{1}{1-\tau}$ in the above expression stems from keeping the energy constant during each transmission phase of duration $(1-\tau)\,T$ [11]. The initially available transmit power (aside from the harvested one) is also enhanced to $\frac{p}{1-\tau}$ for the same reason. Under the above assumptions and using standard time sharing arguments [12], the Shannon capacity of the Alice-Bob link is given by

$$C^{EH}(p, \tau, \gamma) = \frac{(1-\tau)}{2} \log\left(1 + \frac{\left(\frac{p}{1-\tau} + p^{EH}\right)|H|^2}{\gamma |G_B|^2 + N_B}\right). \tag{5}$$

Note that the multiplicative term $(1-\tau)$ in front of the logarithm represents the reduction of the Shannon capacity due to time sharing.

III. Jammer Neutralization

In this Section, we first investigate whether it is possible to neutralize the jammer. We note that C^{EH} is increasing with the transmit power p for fixed τ and γ. On the other hand, C^{EH} is not necessarily decreasing with the jamming power: since the interference is harvested and p^{EH} increases with γ, the Shannon capacity may even increase with the interfering power γ for specific system parameters. Although this may seem counter-intuitive, consider the case in which the interfering link is very poor $|G_B|^2 << 1$ (e.g., Jay is very far from Bob). In this case, the interference at Bob is negligible: $\gamma |G_B|^2 + N_B \simeq N_B$ and, hence, C^{EH} increases with γ due to p^{EH}. We prove this result rigorously by investigating the first-order derivatives of C^{EH}.

Proposition 1: For fixed p and τ, if $\frac{|G_A|^2}{N_A} > \frac{|G_B|^2}{N_B}$, then $C^{EH}(p, \tau, \gamma)$ is monotonically increasing w.r.t. γ if $p \leq p_{th}(\tau) \triangleq \tau K$, with $K \triangleq (\frac{|G_A|^2 N_B}{|G_B|^2} - N_A)\zeta$, and it is monotonically decreasing w.r.t. γ if $p > p_{th}(\tau)$. This implies:

$$\arg \max_{\gamma \in [0,\Gamma]} C^{EH}(p, \tau, \gamma) = 0, \quad \text{if } p \leq p_{th}(\tau), \tag{6}$$

$$\arg \max_{\gamma \in [0,\Gamma]} C^{EH}(p, \tau, \gamma) = \Gamma, \quad \text{if } p > p_{th}(\tau). \tag{7}$$

Otherwise, $C^{EH}(p, \tau, \gamma)$ is always monotonically decreasing w.r.t. γ for any fixed p and τ and

$$\arg \max_{\gamma \in [0,\Gamma]} C^{EH}(p, \tau, \gamma) = \Gamma. \tag{8}$$

Intuitively, if the quality of the harvesting link is higher than that of the jamming link $\frac{|G_A|^2}{N_A} > \frac{|G_B|^2}{N_B}$, then the legitimate users can neutralize the jammer by tuning the transmit power p and the EH policy τ such that $p \leq p_{th}(\tau)$. This highlights the existence of a power threshold $p_{th}(\tau)$ below which, harvesting the jamming interference in the first phase overcomes the harmful jamming in the second phase. The optimal strategy (p, τ) that neutralizes the jammer (NJ) is given below.

Theorem 1: If $\frac{|G_A|^2}{N_A} > \frac{|G_B|^2}{N_B}$, the strategy that maximizes the capacity while neutralizing the jammer (p^{NJ}, τ^{NJ}) is given as follows:

a) If $p_{th}^{-1}(P) > 1$, then $(p^{NJ}, \tau^{NJ}) = (p_{th}(\hat{\tau}), \hat{\tau})$, where $p_{th}^{-1}(p) = \frac{p}{K}$ is the inverse function of $p_{th}(\tau)$.

b) Otherwise, the optimal strategy is

$$(p^{NJ}, \tau^{NJ}) = \underset{(p,\tau) \in \{(p_1,\tau_1),(p_2,\tau_2)\}}{\arg\max} C^{EH}(p, \tau, 0),$$
$$(p_1, \tau_1) = (\min\{p_{th}(\hat{\tau}), P\}, \min\{\hat{\tau}, p_{th}^{-1}(P)\}),$$
$$(p_2, \tau_2) = (P, \max\{\tilde{\tau}, p_{th}^{-1}(P)\}),$$

where $\hat{\tau} \in (0, 1)$ and $\tilde{\tau} \in (0, 1)$ are the unique solutions of the equations

$$\frac{\partial C^{EH}(p_{th}(\tau), \tau, 0)}{\partial \tau} = 0 \text{ and } \frac{\partial C^{EH}(P, \tau, 0)}{\partial \tau} = 0$$

respectively, which can be easily computed numerically.[1]

Fig. 2 illustrates the Shannon capacity obtained while neutralizing the jammer, $C^{EH}(p^{NJ}, \tau^{NJ}, 0)$, as a function of the signal to interference power ratio (SIR) defined as $SIR = P/\Gamma$ in the range $P/\Gamma = -30$ dB to $P/\Gamma = 10$ dB, for various settings w.r.t. the channels gains for $N_A = -10$ dBm, $N_B = -7$ dBm, $\Gamma = 10$ dBm, $\zeta = 0.8$. These parameters

[1]Finding explicit expressions for $\hat{\tau}$ and $\tilde{\tau}$ is non trivial and involves solving nonlinear equations containing both logarithmic and fractional terms. Instead, we can exploit numerical methods based on iterative one-dimensional search (e.g., `fzero` in MATLAB) or methods relying on trust region or Levenberg-Marquardt techniques (e.g., `fsolve` in MATLAB).

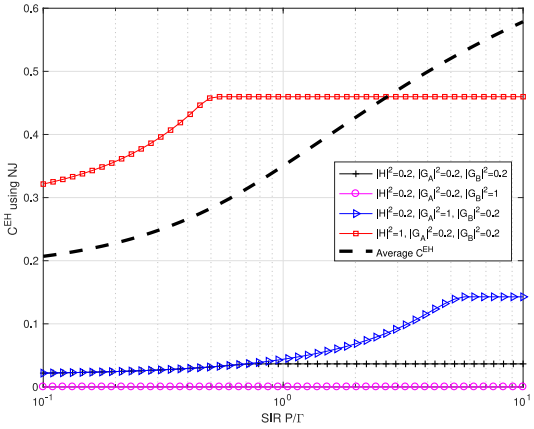


Fig. 2. Shannon capacity while neutralizing the jammer as a function of SIR. The jammer cannot be neutralized if the harvesting link is poor. Transmitting at full power P is not always optimal.

were chosen to showcase the different regimes in terms of the optimal strategies at the NJ and NE states. Nevertheless, all our remarks and observations are general and remain valid irrespective of the specific choice of the system parameters.

The capacity $C^{EH}(p^{NJ},\tau^{NJ},0)$ is zero for $|H|^2=0.2$, $|G_A|^2=0.2$, $|G_B|^2=1$ since the minimum condition on the harvesting link quality is not met and the jammer cannot be neutralized. In the other settings, we can identify two regimes depending on whether $p_{th}^{-1}(P)\leq 1$ or not. When the SIR is low, the system is in case b) of Theorem 1 and increasing P increases the feasible set of the optimization problem and, hence, the optimal value of the capacity. At higher SIR, the system shifts to case a), in which the optimal solution that neutralizes the jammer no longer depends on P. The dashed black curve depicts the average $C^{EH}(p^{NJ},\tau^{NJ},0)$ over 10,000 random realizations of the channel gains drawn from a standard Gaussian distribution.

We remark that to neutralize the jammer a non-zero EH duration is required, i.e., $\tau^{NJ}>0$, during which little energy can actually be harvested (as the jammer is silent). Also, the transmit power may be required to be below the maximum available power. In light of this, we next investigate the optimal strategies of both parties in this competitive interaction.

IV. Optimal Strategies: Game Theoretic Analysis

In this Section, we study the adversarial interaction between the legitimate users and the jammer using a non-cooperative game [13]. More specifically, we formulate a two-player zero-sum game defined as the triplet:

$\mathcal{G}=\{\{L,J\},\{A_L,A_J\},C^{EH}(a_L,a_J)\}$, where L and J denote the opposing players; $A_L=[0,P]\times[0,1]$ and $A_J=[0,\Gamma]$ denote the possible actions the two players can take; and $C^{EH}(a_L,a_J)$ is both the utility of L and the cost of J. The strategy of legitimate users is denoted by $a_L=(p,\tau)$ and that of the jammer by $a_J=\gamma$.

A natural solution of the game is the NE, denoted by (a_L^{NE},a_J^{NE}), which is stable to unilateral deviations:

$$C^{EH}(a_L^{NE},a_J^{NE})\geq C^{EH}(a_L,a_J^{NE}),\quad \forall a_L\neq a_L^{NE},$$
$$C^{EH}(a_L^{NE},a_J^{NE})\leq C^{EH}(a_L^{NE},a_J),\quad \forall a_J\neq a_J^{NE}.$$

Neither the legitimate player nor the jammer have any interest in deviating from the NE state knowing that their opponent is playing the NE strategy.

Proposition 2: The NJ state, $(p^{NJ},\tau^{NJ},0)$ in Theorem 1, is not a NE of the game $\mathcal{G}$.

Proof: If $p_{th}^{-1}(P)>1$ then $(p^{NJ},\tau^{NJ})=(p_{th}(\hat{\tau}),\hat{\tau})$, from Theorem 1. This cannot be a NE, because it implies that $p_{th}(\hat{\tau})<P$, and since C^{EH} is increasing in p then player L should transmit at maximum power at the NE: $p^{NE}=P$. Otherwise, we have two cases: $(p^{NJ},\tau^{NJ})=(P,\max\{\tilde{\tau},p_{th}^{-1}(P)\})$ and $(p^{NJ},\tau^{NJ})=(P,p_{th}^{-1}(P))$. Neither can be NE: since the jammer is silent ($\gamma^{NJ}=0$) and only the noise N_A is harvested, the legitimate user will deviate from $\tau^{NJ}>0$ to $\tau=0$. EH operates as a threat to neutralize the jammer, which results in an inefficient time sharing policy.

The game's NE is given in the following theorem.

Theorem 2: The NE of the game $\mathcal{G}$ is $(p^{NE},\tau^{NE},\gamma^{NE})=(P,\tau^{NE},\Gamma)$, where $\tau^{NE}\in\{0,\tau^*\}$ with $\tau^*\in(0,1)$ the unique solution of

$$\frac{\partial C^{EH}(P,\tau,\Gamma)}{\partial\tau}=0 \tag{9}$$

(which can be easily computed numerically), depending on the system parameters. Moreover, the NE always outperforms the NJ state.

Proof: The transmit power is maximum $p^{NE}=P$ since C^{EH} is increasing in the transmit power p. Then, we prove by *reductio ad absurdum* that at the NE we have: $p_{th}(\tau^{NE})\leq p^{NE}$, implying that $\gamma^{NE}=\Gamma$ from Proposition 1. Finding τ^{NE} reduces to solving the optimization problem:

$$\tau^{NE}=\arg\max_{\tau\in[0,1]}\ C^{EH}(P,\tau,\Gamma). \tag{10}$$

Based on the first and second order derivatives, $C^{EH}(P,\tau,\Gamma)$ is concave and either decreases ($\tau^{NE}=0$) or it has a unique critical point ($\tau^{NE}=\tau^*$), depending on the parameters. To prove that the NE always outperforms the NJ state, we use two ingredients. From Proposition 1, whenever $p=p_{th}(\tau)$, the Shannon capacity is constant w.r.t. γ and, hence, $C^{EH}(p^{NJ},\tau^{NJ},0)=C^{EH}(p^{NJ},\tau^{NJ},\Gamma)$. From the NE definition and knowing that $p^{NE}=P$ and $\gamma^{NE}=\Gamma$:

$$(P,\tau^{NE})=\arg\max_{p,\tau}\ C^{EH}(p,\tau,\Gamma). \tag{11}$$

These two facts yield that the NE outperforms the NJ state: $C^{EH}(p^{NJ},\tau^{NJ},0)\leq C^{EH}(P,\tau^{NE},\Gamma)$.

In Fig. 3, we compare the capacity at the NE, $C^{EH}(P,\tau^{NE},\Gamma)$, with the capacity when neutralizing the jammer, $C^{EH}(p^{NJ},\tau^{NJ},0)$, normalized to the former by illustrating

$$F^{NJ}\triangleq\frac{C^{EH}(P,\tau^{NE},\Gamma)-C^{EH}(p^{NJ},\tau^{NJ},0)}{C^{EH}(P,\tau^{NE},\Gamma)}. \tag{12}$$

The simulation setting is identical to Fig. 2. We remark that the NE always outperforms the NJ policy, which is consistent to our analysis. The intuition is that, when neutralizing the jammer, Alice does not necessarily transmit at maximum power P and has to spend a minimum proportion of time $\tau^{NJ}>0$

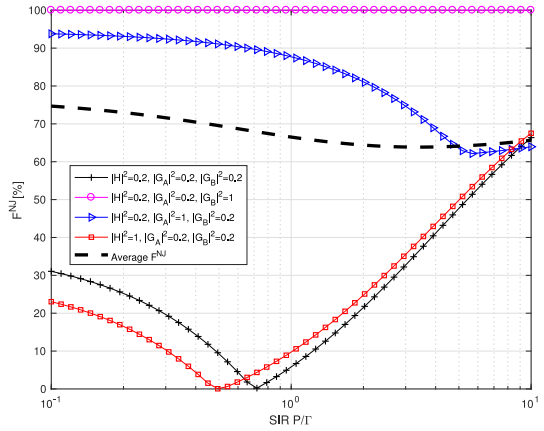


Fig. 3. NE vs. NJ efficiency: F^{NJ} in (12) as a function of SIR. The NE always outperforms the NJ state. At low SIR, exploiting the dominant jamming interference is more beneficial than silencing the jammer. At high SIR, neutralizing the jammer via EH is inefficient.

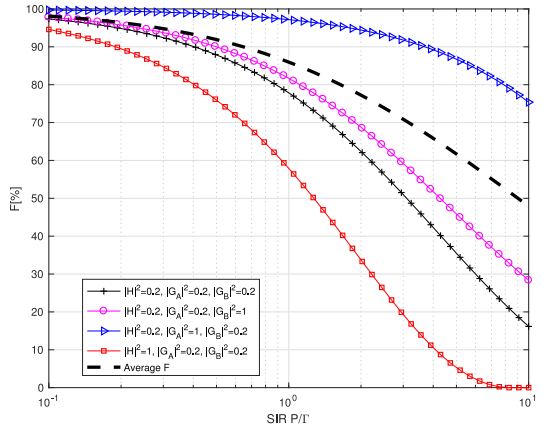


Fig. 4. EH efficiency: F in (13) as a function of SIR. EH is particularly beneficial at low SIR when the dominant jamming interference can be exploited. At high SIR, the interference becomes negligible and EH is less useful.

for EH, as a threat to force the jammer to remain silent, even though no energy can actually be harvested during this time. Only in two of the four specific channel settings, the capacity at the NJ state equals the NE capacity and only at specific SIR values (the 0% minimum points). Surprisingly, using EH to neutralize the jammer is not beneficial in most cases. Instead, the legitimate users should harvest the jamming interference and use it for information transmission. The 100% relative gain curve corresponds to the case in which the jammer cannot be neutralized because of the poor quality of the harvesting link. The dashed black curve represents the average efficiency of the NE w.r.t. the NJ state and shows a relative gain between 64%–75%.

In the same setting, in Fig. 4 we evaluate the efficiency of EH as a measure against a strategic jammer and compare the capacity at the NE, $C^{EH}(P,\tau^{NE},\Gamma)$, with $C^{EH}(P,0,\Gamma)$, the capacity in absence of EH capability normalized to the former, by analyzing

$$F \triangleq \frac{C^{EH}(P,\tau^{NE},\Gamma) - C^{EH}(P,0,\Gamma)}{C^{EH}(P,\tau^{NE},\Gamma)}. \qquad (13)$$

At low SIR, the relative EH gain in terms of capacity is very high, approaching 100% when the gain of the harvesting link is high. The jamming interference is dominant and exploiting it for useful communication is very beneficial. At high SIR, the gain from EH decreases progressively towards zero since the jamming interference that can be harvested becomes negligible. The potential of using EH as an anti-jamming measure is demonstrated by the substantial relative gains in terms of Shannon capacity that can on average reach 95% in the low SIR regime.

V. Conclusion

In this letter, we showed that EH can be exploited to efficiently mitigate jamming attacks. We proved that a jammer can be completely neutralized by appropriately tuning the transmit power and the EH duration. However, this restricts the transmit power of the legitimate user and requires a minimum harvesting duration, during which little energy can actually be harvested (as the jammer is forced to remain silent). Therefore, neutralizing the jammer is not necessarily optimal. Employing a zero-sum game formulation, we showed that at the NE both players should transmit at full power and the optimal EH duration depends on the system parameters. Our simulation results show that EH can offer substantial gains in terms of capacity. At low SIR, the average gains can reach 95%, showcasing the high potential of EH as an efficient counter-jamming measure.

References

[1] X. Zhou, R. Zhang, and C. K. Ho, "Wireless information and power transfer: Architecture design and rate-energy tradeoff," *IEEE Trans. Commun.*, vol. 61, no. 11, pp. 4754–4767, Nov. 2013.

[2] Y. Gu and S. Aissa, "RF-based energy harvesting in decode-and-forward relaying systems: Ergodic and outage capacities," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6425–6434, Nov. 2015.

[3] R. Rajesh, V. Sharma, and P. Viswanath, "Capacity of Gaussian channels with energy harvesting and processing cost," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2563–2575, May 2014.

[4] M.-L. Ku, W. Li, Y. Chen, and K. J. R. Liu, "Advances in energy harvesting communications: Past, present, and future challenges," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1384–1412, 2nd Quart., 2016.

[5] S. Wei, R. Kannan, V. Chakravarthy, and M. Rangaswamy, "CSI usage over parallel fading channels under jamming attacks: A game theory study," *IEEE Trans. Wireless Commun.*, vol. 60, no. 4, pp. 1167–1176, Apr. 2012.

[6] L. Xiao, T. Chen, J. Liu, and H. Dai, "Anti-jamming transmission Stackelberg game with observation errors," *IEEE Commun. Lett.*, vol. 19, no. 6, pp. 949–952, Jun. 2015.

[7] K. Wang, L. Yuan, T. Miyazaki, Y. Chen, and Y. Zhang, "Jamming and eavesdropping defense in green cyber-physical transportation systems using Stackelberg game," *IEEE Trans Ind. Informat.*, vol. 14, no. 9, pp. 4232–4242, Sep. 2018.

[8] Z. Fang, T. Song, and T. Li, "Energy harvesting for two-way OFDM communications under hostile jamming," *IEEE Signal Process. Lett.*, vol. 22, no. 4, pp. 413–416, Apr. 2015.

[9] H. Xing, K.-K. Wong, Z. Chu, and A. Nallanathan, "To harvest and jam: A paradigm of self-sustaining friendly jammers for secure AF relaying," *IEEE Trans. Signal Process.*, vol. 63, no. 24, pp. 6616–6631, Dec. 2015.

[10] E. V. Belmega and A. Chorti, "Energy harvesting in secret key generation systems under jamming attacks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.

[11] E. V. Belmega and A. Chorti, "Protecting secret key generation systems against jamming: Energy harvesting and channel hopping approaches," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 11, pp. 2611–2626, Nov. 2017.

[12] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Somerset, U.K.: Wiley, 2012.

[13] D. Fudenberg and J. Tirole, *Game Theory*. Cambridge, MA, USA: MIT Press, 1991.

Effective Secrecy Rate for a Downlink NOMA Network

Wenjuan Yu, *Member, IEEE*, Arsenia Chorti, *Member, IEEE*, Leila Musavian, *Member, IEEE*, H. Vincent Poor, *Fellow, IEEE*, and Qiang Ni, *Senior Member, IEEE*

***Abstract*— In this paper, a novel approach is introduced to study the achievable delay-guaranteed secrecy rate, by introducing the concept of the effective secrecy rate (ESR). This study focuses on the downlink of a non-orthogonal multiple access (NOMA) network with one base station, multiple single-antenna NOMA users and an eavesdropper. Two possible eavesdropping scenarios are considered: 1) an internal, unknown, eavesdropper in a purely antagonistic network; and 2) an external eavesdropper in a network with trustworthy peers. For a purely antagonistic network with an internal eavesdropper, the only receiver with a guaranteed positive ESR is the one with the highest channel gain. A closed-form expression is obtained for the ESR at high signal-to-noise ratio (SNR) values, showing that the strongest user's ESR in the high SNR regime approaches a constant value irrespective of the power coefficients. Furthermore, it is shown the strongest user can achieve higher ESR if it has a distinctive advantage in terms of channel gain with respect to the second strongest user. For a trustworthy NOMA network with an external eavesdropper, a lower bound and an upper bound on the ESR are proposed and investigated for an arbitrary legitimate user. For the lower bound, a closed-form expression is derived in the high SNR regime. For the upper bound, the analysis shows that if the external eavesdropper cannot attain any channel state information (CSI), the legitimate NOMA user at high SNRs can always achieve positive ESR, and the value of it depends on the power coefficients. Simulation results numerically validate the accuracy of the derived closed-form expressions and verify the analytical results given in the theorems and lemmas.**

***Index Terms*— Effective capacity, secrecy rate, NOMA, delay-outage probability.**

Manuscript received June 13, 2018; revised March 29, 2019; accepted August 14, 2019. Date of publication September 6, 2019; date of current version December 10, 2019. This work was supported in part by the U.K. Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/N032268/1 and Grant EP/K011693/1, in part by the EU Seventh Framework Programme (FP7) under Grant PIRSES-GA-2013-610524, in part by the Royal Society Project under Grant IEC170324, and in part by the U.S. National Science Foundation under Grant ECCS-1647198 and Grant CNS-1702808. The associate editor coordinating the review of this article and approving it for publication was Prof. S. Yang. *(Corresponding author: Wenjuan Yu.)*

W. Yu is with the 5G Innovation Centre, Institute for Communication Systems, University of Surrey, Guildford GU2 7XH, U.K. (e-mail: w.yu@surrey.ac.uk).

A. Chorti is with ETIS, UMR 8051, Université Paris Seine, Université Cergy-Pontoise, ENSEA, CNRS, 95000 Cergy, France (e-mail: arsenia.chorti@ensea.fr).

L. Musavian is with the School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, U.K. (e-mail: leila.musavian@essex.ac.uk).

H. V. Poor is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: poor@princeton.edu).

Q. Ni is with the InfoLab21, School of Computing and Communications, Lancaster University, Lancaster LA1 4WA, U.K. (e-mail: q.ni@lancaster.ac.uk).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2019.2938515

I. Introduction

NON-ORTHOGONAL multiple access (NOMA) is considered to be a promising multiple access (MA) technique for fifth generation (5G) and beyond (B5G) networks, because of its advantages over conventional orthogonal multiple access (OMA) schemes, in terms of spectral efficiency [1], cell-edge throughput [2], and energy efficiency [3]. Power-domain NOMA[1] allows multiple users to transmit with different transmission power levels, but using the same radio resources, such as subcarrier channels, codes and time slots [4], [5]. Specifically, superposition coding is applied at the transmitter to enable user-multiplexing, while multiuser separation techniques such as successive interference cancellation (SIC) are applied at the receiver to eliminate the co-channel interference and decode the superimposed messages [6], [7]. The users with higher channel gains can obtain the prior information of weaker users in accordance with the NOMA principle. On one hand, the obtained prior information can be utilized to help the weaker users to decode their messages [2], but as mentioned in [8], this can also cause security issues.

Providing secure communication has always been an important issue in wireless networks. Traditionally, security is carried out at upper layers of the protocol stack, relying on encryption algorithms which are agnostic to the wireless channels' physical properties [9], [10]. However, in 5G and Internet of things (IoT) networks, with the explosive growth in the number of low-complexity, power and computationally constrained devices, the concept of physical layer security (PLS) is attracting considerable attention. PLS exploits the randomness of wireless channels to ensure that the transmitted information cannot be decoded by a malicious eavesdropper [10], [11]. Based on the concept of perfect secrecy proposed by Shannon [12], Wyner introduced the wiretap channel model, in which two legitimate users can communicate reliably through a main channel while keeping the exchanged messages confidential from an eavesdropper. Considering Gaussian wiretap channels, the secrecy capacity, i.e., the maximum achievable rate which guarantees reliable communication while the eavesdropper cannot decode any confidential message, is equal to the difference between the main channel's Shannon capacity and the adversary channel's Shannon capacity [9], [13]. Consequently, confidential transmission in Gaussian wiretap

[1]In the following sections, the power-domain NOMA is simply referred to as NOMA.

channels requires that the legitimate user's channel has a higher signal-to-noise ratio (SNR) than the wiretap channel [11]. On the other hand, the ergodic secrecy capacity for wireless fading channels can be positive even when the adversary has a higher average SNR than the legitimate user's channel, which indicates that fading can be beneficial for secrecy [11]. This is because whenever the legitimate user experiences a higher channel gain than the eavesdropper, this fading realization can be exploited for secure transmission [9].

Focusing on large-scale networks utilizing the NOMA protocol, PLS was investigated in [14] and [15] by invoking stochastic geometry. By adopting a user pairing technique which allows two mobile users to share one orthogonal radio resource, the exact analytical expressions for the secrecy outage probability were derived and analyzed for single-antenna and multiple-antenna scenarios [15]. In [16], the feasible transmit power region was first identified to maximize the sum of secrecy rates for a single-input single-output NOMA system, which satisfies all users' required quality-of-service (QoS) values. Then, a closed-form expression was derived for the proposed optimal power allocation strategy. In [17], a network in which the source and an untrusted relay simultaneously transmit signals through non-orthogonal channels was considered. It was concluded that the proposed non-orthogonal relaying scheme provides an improved ergodic secrecy rate, compared to the conventional orthogonal relaying schemes.

Although investigating the secrecy capacity in different wireless networks with the NOMA protocol applied, the aforementioned literature adopts the physical layer channel model, i.e., Shannon theory, without placing emphasis on the legitimate users' delay requirements. On the other hand, for the emerging delay-sensitive wireless communication networks and applications [18], such as vehicular communications, e-health communication and Tactile Internet, delay QoS guarantees will play a critical role in 5G and beyond 5G networks. Furthermore, in future wireless networks, users are expected to necessitate flexible delay guarantees for achieving different service requirements. Henceforth, in order to satisfy diverse delay requirements, a simple and flexible delay QoS model is imperative to be applied and investigated. In this respect, the effective capacity (EC) theory was proposed in [19], with EC denoting the maximum constant arrival rate which can be served by a given service process, while guaranteeing the required statistical delay provisioning. By considering the secrecy rate as the given service rate, we propose in turn the novel concept of the effective secrecy rate (ESR); ESR represents the maximum constant arrival rate that can be securely served, on the condition that the required delay constraint can be statistically satisfied.

In this paper, focusing on a downlink NOMA network with one base station (BS), multiple single-antenna mobile users and an eavesdropper, we aim to propose and thoroughly analyze the ESR for delay-sensitive NOMA users. Different from the user paring design in [14] and [15], we assume that all legitimate NOMA users can transmit using the same resource slot. To provide a comprehensive study, two different eavesdropping scenarios are considered in this paper. Firstly, a purely antagonistic network is studied, in which every NOMA user can act as a passive eavesdropper intercepting confidential messages intended for other users. A practical scenario for this case would be ad-hoc networks with confidential information broadcasted with existing untrusted peers [10]. Secondly, when all NOMA users are trustworthy, there may exist an external eavesdropper that has an interest in compromising the network's security. Hence, with an external eavesdropper intending to decode the NOMA users' messages, the ESR of an arbitrary legitimate user is proposed and investigated, while satisfying their corresponding statistical delay requirement.

Considering the above eavesdropping scenarios, this paper has the following main contributions:

- After proposing the concept of ESR, we theoretically analyze the impact of delay exponent θ on ESR and provide Theorem 1. It is proved that the ESR monotonically decreases with θ. Specifically, we prove that when $\theta \to 0$, the ESR converges to the traditional ergodic secrecy rate, while when $\theta \to \infty$, it represents the delay-limited case and simulation results show that the value of ESR reduces to zero, for Rayleigh fading channels.
- In the presence of an unknown internal eavesdropper, the only legitimate receiver which can achieve a non-zero secrecy rate is the user with the highest channel gains.[2] For the strongest user, we derive the closed-form expressions for the ESR and the traditional ergodic secrecy rate, at high SNRs. Furthermore, we show that the strongest user can achieve higher ESR if it has a distinctive advantage in terms of channel gain with respect to the second strongest user. Also, it is proved that the ESR in the high SNR regime approaches a constant value irrespective of the power coefficients.
- In the presence of a malicious external eavesdropper, a lower bound and an upper bound are respectively analyzed for the ESR and the traditional ergodic secrecy rate.[3] For the lower bound, the closed-form expressions are derived for the high SNR regime. For the upper bound, the analysis shows that if the external eavesdropper cannot attain any channel state information (CSI), the legitimate NOMA user in the high SNR regime can always achieve positive delay-guaranteed secrecy rate, and the value of it depends on the power coefficients.
- Simulation results verify the accuracy of the derived closed-form expressions and confirm the tightness of the proposed bounds. The impact of the delay requirements, the size of the NOMA set and the power coefficient settings is also investigated. Specifically, it is shown that when there is an external eavesdropper, a legitimate NOMA user with stronger channel gains will be more impacted in terms of its achievable ESR, in order to guarantee a required statistical delay QoS. Further, numerical results also reveal that a larger user set leads to smaller ESR values for the legitimate users in both eavesdropping scenarios.

[2]Hereafter, the user with the highest channel gain is referred to as the strongest user.

[3]We note that in this case, these results are not limited to the strongest user.

The rest of the paper is organised as following: the system model is discussed in Section II. The theory of EC is briefly reviewed in Section III, followed by analytical expressions of the ESR for the scenarios of an internal and an external eavesdropper, respectively. Section IV includes simulation results and discussions, followed by conclusions summarized in Section V.

II. System Model

A classical cellular downlink transmission is considered, where one BS transmits public and confidential messages to M single-antenna NOMA users in the presence of a malicious eavesdropper. The wireless channels from the BS to legitimate NOMA users and the eavesdropper are all assumed to be block fading, i.e., the channel gains remain constant within each fading-block, but independently change from one block to the next. Each fading-block duration $T_{\rm f}$ is equal to the frame size, which is an integer multiple of the symbol period.

The channel gains from the BS to the $m^{\rm th}$ user and the eavesdropper are assumed to be Rayleigh distributed and denoted by $h_m, m \in \{1\ldots,,M\}$[4] and h_e,[5] respectively. Without loss of generality, all NOMA users and the malicious adversary's channel gains are assumed to be ordered as $0 < |h_1|^2 \leq |h_2|^2 \ldots \leq |h_{M_E}|^2 \leq |h_e|^2 \leq |h_{M_{E+1}}|^2 \ldots \leq |h_M|^2$, in which M_E indicates the number of NOMA users that have smaller or equal channel gains with respect to the eavesdropper. The BS transmits the signal $\sum_{i=1}^{M}\sqrt{\gamma_i}Ps_i$ to all legitimate users in accordance with NOMA principle. Here, γ_i is the $i^{\rm th}$ user's power coefficient, P is the total transmission power, and s_i is the message for the $i^{\rm th}$ user with $\mathbb{E}\left[|s_i|^2\right]=1$. By following the NOMA protocol [20], the power coefficients[6] are ordered as $\gamma_1 \geq \cdots \geq \gamma_M$, and $\sum_{i=1}^{M}\gamma_i = 1$.

The received signals y_m at the $m^{\rm th}$ legitimate user, $1 \leq m \leq M$, and y_e at the eavesdropper are respectively given as [21]:

$$y_m = h_m \sum_{i=1}^{M}\sqrt{\gamma_i}Ps_i + n_m, \tag{1}$$

$$y_e = h_e \sum_{i=1}^{M}\sqrt{\gamma_i}Ps_i + n_e, \tag{2}$$

where n_m, n_e denote zero-mean additive white Gaussian noise (AWGN) at the $m^{\rm th}$ user and at the eavesdropper, respectively, i.e., $n_m, n_e \sim \mathcal{N}(0,\sigma^2)$.[7]

Based on the NOMA principle, the $m^{\rm th}$ user applies the SIC technique to detect its own messages, by successively decoding the weaker users' messages, i.e., the $i^{\rm th}$ user with $|h_i|^2 < |h_m|^2$, and then eliminating the message from the SNR received signals [22]. On the other hand, the messages for the user with stronger channel gains, i.e., the $i^{\rm th}$ user with $|h_i|^2 > |h_m|^2$, will be considered as noise at the $m^{\rm th}$ user. To ensure that SIC is successfully applied at the $m^{\rm th}$ user, it is assumed that $R_{i\to m} \geq \tilde{R}_i$ [23], where $R_{i\to m}$ denotes the $m^{\rm th}$ user's data rate to decode the $i^{\rm th}$ user's message and $\tilde{R}_i$ is the target data rate for the $i^{\rm th}$ user. Therefore, when it decodes its own message, the $m^{\rm th}$ legitimate NOMA user's achievable rate, in b/s/Hz, is given by [14]

$$R_m = \log_2\left(1+\frac{\rho|h_m|^2\gamma_m}{\rho|h_m|^2\sum_{i=m+1}^{M}\gamma_i+1}\right), \quad 1 \leq m \leq M, \tag{3}$$

where ρ is the transmit SNR, i.e., $\rho = \dfrac{P}{\sigma^2}$.

Regarding the eavesdropper, it employs SIC to detect the $m^{\rm th}$ legitimate user's messages with an achievable decoding rate denoted by $R_e^{(m)}$. Considering that the eavesdropper can be within the set of NOMA users or distinct from them, the corresponding mathematical expressions of $R_e^{(m)}$ can be different and we will study them respectively in the following Section. The $m^{\rm th}$ NOMA user's secrecy rate is achievable when an encoding scheme exists that simultaneously ensures reliable communication and perfect secrecy with respect to the eavesdropper. In the following, the m-th user's achievable secrecy rate is denoted by R_s^m and expressed as [9]

$$R_s^m = \left[R_m - R_e^{(m)}\right]^+, \quad 1 \leq m \leq M, \tag{4}$$

where R_m is given in (3) and $[x]^+ = \max\{0,x\}$.

III. Effective Secrecy Rate

In order to support the emerging delay-sensitive wireless communication services and applications, in the following, we first introduce the theory of EC. Then, we introduce the concept of the ESR as an achievable arrival rate that can be securely served, while statistically satisfying the required delay QoS constraints. Let us take the $m^{\rm th}$ user as an example. Assume a first-in-first-out (FIFO) buffer for the $m^{\rm th}$ user at the BS.[8] Define $D_m(t)$ as the delay experienced by a packet arriving at time t. From [19], the probability of the delay $D_m(t)$ exceeding a maximum delay limit $D_{\max}^m$ can be estimated as

$$P_{\rm delay}^{\rm out} = \Pr\{D_m(t) > D_{\max}^m\} \approx \Pr\{\mathcal{Q}(t)>0\}{\rm e}^{-\theta_m\mu D_{\max}^m}, \tag{5}$$

where $P_{\rm delay}^{\rm out}$ denotes the delay violation probability limit for the $m^{\rm th}$ user, $\Pr\{\mathcal{Q}(t)>0\}$ is the probability of a non-empty buffer at time t, $D_{\max}^m$ is the given delay bound in the unit of symbol duration, and θ_m ($\theta_m > 0$) represents the exponential decay rate. The authors in [19] proved that the constant arrival rate needs to be limited to the value of μ, which equals to EC, so that a target delay violation probability limit can be met. Let $\{R_s^m(t), t = 1,2,\ldots\}$ be a series of non-negative random variables, representing the service process of the $m^{\rm th}$ user. Assume that the service process satisfies Gärtner-Ellis theorem [24]. Then, the EC for the $m^{\rm th}$ user on a block-fading channel is defined as

$$E_{\rm s}^m = -\frac{1}{\theta_m T_{\rm f} B}\ln\left(\mathbb{E}\left[e^{-\theta_m T_{\rm f} B R_s^m}\right]\right), \quad \text{(b/s/Hz)}, \tag{6}$$

[4]The time index t is omitted hereafter.

[5]The instantaneous channel gain h_e is unknown, if the eavesdropper is an external adversary.

[6]Adaptive power allocation can influence the exact values of ESR, but this is beyond the scope of this paper and optimal power allocation to maximize the network's sum ESR is considered to be a future research topic.

[7]For simplicity, the noise variances at all users and the eavesdropper are assumed to be identical and equal to σ^2.

[8]It is assumed that for every served user, there is one virtual buffer at the BS.

where $\mathbb{E}[\cdot]$ is the expectation over its channel gains. When the focus is on the rate that can be securely transmitted, we can obtain the ESR for the m^{th} user, by inserting the achievable secrecy rate R_s^m, given in (4), into (6).

From (5), it can be noted that θ_m denotes the exponential decay rate of delay violation probability, for the m^{th} user. A smaller value of θ_m indicates that the user has a relatively loose delay QoS requirement, while a larger value of θ_m means that a more stringent delay QoS is required. In particular, when $\theta_m \to 0$, the probability of the experienced delay exceeding a given bound approaches one. When $\theta_m \to \infty$, it indicates that the user cannot tolerate any delay outage. To clarify, we summarize in the following theorem.

Theorem 1: The ESR for the m^{th} user, i.e., E_{s}^m in (6), is a monotonically decreasing function in θ_m. When $\theta_m \to 0$, E_{s}^m converges to the ergodic secrecy rate, i.e., $\mathbb{E}[R_s^m]$. When $\theta_m \to \infty$, this represents the delay-limited scenario and the value of E_{s}^m reduces to zero.

Proof: See Appendix A. ∎

Theorem 1 shows that the proposed ESR, describing the delay-guaranteed secrecy rate,[9] contains a delay exponent θ_m indicating the stringency of delay requirement. Specifically, this theorem reveals that the ESR is a more general performance metric, which includes the traditional ergodic secrecy rate at an extreme case. For delay-limited scenarios, i.e., when $\theta_m \to \infty$, the value of E_{s}^m reduces to zero for Rayliegh fading channels, which will be shown in Section IV.

A. Effective Secrecy Rate With an Internal Eavesdropper

In this section, we first consider a purely antagonistic network in which every user can be a potential eavesdropper intercepting the confidential messages of the other users. Assume that the knowledge of CSI for all legitimate users is perfectly known at the BS, which implies that the internal eavesdropper's CSI is available. Note that by applying SIC, the user with the strongest channel gains can successfully decode the information of other NOMA users which have weaker channel gains. Hence, when there is an untrusted internal adversary, the only legitimate receiver which can achieve a non-zero secrecy rate is the M^{th} user which has the strongest channel gains. Specifically, the worst case scenario is that the $(M-1)^{\text{th}}$ user acts as the eavesdropper and intends to detect the M^{th} user's messages. Then, the secrecy rate for all legitimate users can be expressed as

$$R_s^m = \begin{cases} \log_2\left(1+\rho|h_M|^2\gamma_M\right) \\ -\log_2\left(1+\rho|h_{M-1}|^2\gamma_M\right), & m=M, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

For ease and compactness, in the following we introduce the notation: $q_m = \rho\gamma_m$, $Q_m = \rho\sum_{i=m}^{M}\gamma_i$, and $\beta_m = -\frac{\theta_m T_f B}{\ln 2}$, where $m \in \{1, 2\ldots, M\}$.

[9]Here, we are talking about the amount of arrival rate that can be securely served and delay statistically guaranteed.

Then, the M^{th} user's ESR can be provided by inserting (7) into (6), which yields

$$E_{\text{s}}^M = \frac{1}{\beta_M}\log_2\left(\mathbb{E}\left[\left(\frac{1+q_M|h_M|^2}{1+q_M|h_{M-1}|^2}\right)^{\beta_M}\right]\right). \tag{8}$$

By setting $y = |h_M|^2$, and $x = |h_{M-1}|^2$, (8) can be expanded as

$$E_{\text{s}}^M = \frac{1}{\beta_M}\log_2\Bigg(\iint\limits_{\substack{0<x<\infty \\ y\geq x}}\left(\frac{1+q_M y}{1+q_M x}\right)^{\beta_M} \times f_{(M-1,M)}(x,y)\,\mathrm{d}x\,\mathrm{d}y\Bigg), \tag{9}$$

where $f_{(M-1,M)}(x,y)$ denotes the joint probability density function (PDF) of the ordered channel gains $|h_{M-1}|^2$ and $|h_M|^2$, with $|h_{M-1}|^2 \leq |h_M|^2$. For M unordered independent channel gains which are Rayleigh distributed with a unit-variance, we define the PDFs of the unordered $|h_{M-1}|^2$ and $|h_M|^2$ as $f(x)$ and $f(y)$, respectively. Then, the cumulative distribution functions (CDF) of the unordered channel gains are given as $F(x)$ and $F(y)$. When all users' channel gains are ordered, the statistical features follow the theory of order statistics [25]. Hence, the joint PDF of the ordered $|h_{M-1}|^2$ and $|h_M|^2$, with $|h_{M-1}|^2 \leq |h_M|^2$, is given by [25]

$$f_{(M-1,M)}(x,y) = M(M-1)f(x)\left(F(x)\right)^{M-2}f(y). \tag{10}$$

Finally, by inserting the joint PDF $f_{(M-1,M)}(x,y)$ into (9), we provide the following theorem.

Theorem 2: Suppose that there is an internal eavesdropper among all NOMA users. Considering the worst case scenario, the M^{th} user's achievable ESR can be written as

$$E_{sc}^M = B_M + \frac{1}{\beta_M}\log_2\Bigg(\sum_{\nu=0}^{M-2}\binom{M-2}{\nu}(-1)^\nu e^{\frac{\nu}{q_M}} \times \sum_{k=0}^{\infty}\frac{(-1)^k}{k!}\frac{(\nu+1)^k}{k-\beta_M+1}\left[\Gamma\left(k+2,\frac{1}{q_M}\right) - \left(\frac{1}{q_M}\right)^{k-\beta_M+1}\times\ \Gamma\left(1+\beta_M,\frac{1}{q_M}\right)\right]\Bigg), \tag{11}$$

where $B_M = \frac{\ln(M(M-1))+2q_M^{-1}}{\beta_M \ln 2}$ and $\Gamma(\cdot,\cdot)$ is the incomplete Γ function.

Proof: See Appendix B. ∎

Note that it is difficult to directly analyze (11) and to obtain intuition of the impact of the various parameters on the ESR. To obtain a tractable expression, in the following, we derive the closed-form expression for E_{s}^M at high SNRs. Firstly, at high SNRs, i.e., $\rho \gg 1$, R_s^M can be simplified to

$$\lim_{\rho\to\infty} R_s^M = \log_2\left(\frac{|h_M|^2}{|h_{M-1}|^2}\right). \tag{12}$$

The ergodic secrecy rate at high SNRs, i.e., $\lim_{\rho\to\infty}\mathbb{E}[R_s^M]$, equals to $\mathbb{E}\left[\log_2\left(\frac{|h_M|^2}{|h_{M-1}|^2}\right)\right]$. This shows that at high SNRs, the M^{th} user's ergodic secrecy rate depends only on the (ratio of) channel gains between the M^{th} user and the internal eavesdropper.

Then, the ESR at high SNRs for the M^{th} user, denoted as $\lim_{\rho\to\infty} E_{\text{s}}^{M}$, can be expressed as

$$\lim_{\rho\to\infty} E_{\text{s}}^{M} = \frac{1}{\beta_M}\log_2\left(\mathbb{E}\left[\left(\frac{|h_M|^2}{|h_{M-1}|^2}\right)^{\beta_M}\right]\right). \quad (13a)$$

Comparing to the case of the ergodic capacity, we can clearly see the impact of the exponential delay decay exponent captuted in β_M. From (13a), we can note that for a fixed value of delay factor β_M, the ESR value at high SNRs increases with the ratio of channel gains. This demonstrates that the M^{th} user can achieve higher delay-guaranteed secrecy rate if it has a distinctive advantage in terms of channel gain with respect to second in rank user. By setting $y = |h_M|^2$, and $x = |h_{M-1}|^2$, (13a) can be expanded as

$$\lim_{\rho\to\infty} E_{\text{s}}^{M} = \frac{1}{\beta_M}\log_2\left(\iint\limits_{\substack{0<x<\infty\\ y\geq x}} \left(\frac{y}{x}\right)^{\beta_M} f_{(M-1,M)}(x,y)\,\mathrm{d}x\,\mathrm{d}y\right). \quad (14)$$

After applying the joint PDF $f_{(M-1,M)}(x,y)$, we derive the closed-form expressions for $\lim_{\rho\to\infty} E_{\text{s}}^{M}$ and $\lim_{\rho\to\infty}\mathbb{E}\left[R_s^M\right]$ in the following theorem.

Theorem 3: Suppose that there is an unknown internal eavesdropper among all NOMA users. Considering the worst case scenario, the closed-form expression for the M^{th} user's ESR at high SNRs, i.e., $\lim_{\rho\to\infty} E_{\text{s}}^{M}$, is given by

$$\lim_{\rho\to\infty} E_{\text{s}}^{M} = \frac{1}{\beta_M}\log_2\Bigg(M(M-1)\Gamma(1-\beta_M) \times \sum_{s=0}^{M-2}\binom{M-2}{s}(-1)^s {}_2\mathbf{F}_1\left[\begin{matrix}1-\beta_M,\ 2\\ 2-\beta_M\end{matrix}; -1-s\right]\Bigg), \quad (15)$$

where $\Gamma(\cdot)$ is gamma function and ${}_2\mathbf{F}_1\left[\begin{smallmatrix}a,\ b\\ c\end{smallmatrix}; z\right]$ is the generalized hypergeometric function [26]. Furthermore, for comparison purposes, the ergodic secrecy rate for the M^{th} user at high SNRs, i.e., $\lim_{\rho\to\infty}\mathbb{E}\left[R_s^M\right]$, can be expressed in closed-form, given in (16).

$$\lim_{\rho\to\infty}\mathbb{E}[R_s^M] = M(M-1)\sum_{s=0}^{M-2}\binom{M-2}{s} \times(-1)^s\frac{1}{s+1}\log_2(s+2). \quad (16)$$

Proof: See Appendix C. ∎

From (15), we can notice that when the transmit SNR asymptotically approaches infinity, the M^{th} user's ESR approaches a constant value, irrespective of the transmit SNR and the power coefficients. Furthermore, from (15) and (16), one can note that when the M^{th} user's delay requirement changes (when β_M varies), $\lim_{\rho\to\infty}\mathbb{E}\left[R_s^M\right]$ will not change but the value of ESR, i.e., $\lim_{\rho\to\infty} E_{\text{s}}^{M}$, will be influenced. This is due to the fact that the delay violation probability is not taken into account in traditional secrecy rate, but is considered in the proposed ESR. This demonstrates the gains in considering the delay-guaranteed ESR in low-latency communications.

The validity of the above derived closed-form expressions, given in (15) and (16), will be verified in Section IV, by comparing with Monte Carlo results. Moreover, simulation results will also show that $\lim_{\rho\to\infty} E_{\text{s}}^{M}$ converges to $\lim_{\rho\to\infty}\mathbb{E}\left[R_s^M\right]$, when $\theta_M \to 0$ (or $\beta_M \to 0$). In other words, we can get that $\lim_{\substack{\rho\to\infty\\ \theta_M\to 0}} E_{\text{s}}^{M} = \lim_{\rho\to\infty}\mathbb{E}\left[R_s^M\right]$. This verifies Theorem 1 and confirms that the proposed ESR is a more flexible metric, with the traditional ergodic secrecy rate emerging as a special case.

B. Effective Secrecy Rate With an External Eavesdropper

Here, we assume that all NOMA users are trustworthy and there exists an external eavesdropper which is distinct from the set of legitimate users and intends to decode as many NOMA users' confidential messages as possible. Then, by employing SIC, the adversary's achievable rate for detecting the m^{th} user's message, namely $R_e^{(m)}$, can be given in (17) [21], shown at the top of the next page.

By inserting (3) and (17) into (4) and applying the defined notations $q_m = \rho\gamma_m$, $Q_m = \rho\sum_{i=m}^{M}\gamma_i$, where $m = \{1, 2, \ldots, M\}$, the secrecy rate for the m^{th} user can be then given in (18), shown at the top of the next page. Firstly, when $1 \leq m \leq M_E$, i.e., $|h_m|^2 \leq |h_e|^2$, we have that $\log_2\left(1+\frac{q_m|h_m|^2}{Q_{m+1}|h_m|^2+1}\right)$ is smaller than or equal to $\log_2\left(1+\frac{q_m|h_e|^2}{Q_{m+1}|h_e|^2+1}\right)$, which means that $R_s^m = 0$. On the other hand, when $M_E+1 \leq m \leq M$, we have $\frac{1}{|h_m|^2} \leq \frac{1}{|h_e|^2}$ and $Q_{m+1} \leq Q_{M_E+1} - q_m$. This means that $R_s^m \geq 0$, when $M_E+1 \leq m \leq M$. Hence, the secrecy rate R_s^m can be simplified to (19), shown at the top of the next page.

Assume that the relative order of all NOMA users and the eavesdropper's channel gains is known, i.e., $0 < |h_1|^2 \leq |h_2|^2 \ldots \leq |h_{M_E}|^2 \leq |h_e|^2 \leq |h_{M_E+1}|^2 \ldots \leq |h_M|^2$. Then, by inserting (19) into (6), we have the conditional ESR, namely E_{cs}^{m}, given below.

Case 1: for the m^{th} user with $1 \leq m \leq M_E$

In this case, it is known that the m^{th} user has weaker channel gains compared to the eavesdropper, i.e., $|h_m|^2 \leq |h_e|^2$. Under this condition, by inserting (19) into (6), we have that $E_{\text{cs}}^{m} = 0$.

Case 2: for the m^{th} user with $M_E+1 \leq m \leq M$

In this case, it is known that the m^{th} user has stronger channel gains compared to the eavesdropper, i.e., $|h_m|^2 \geq |h_e|^2$. Under this condition, we can get that

$$E_{\text{cs}}^{m} = \frac{1}{\beta_m}\log_2\left(\mathbb{E}\left[\left(\frac{Q_m|h_m|^2+1}{Q_{m+1}|h_m|^2+1} \times\frac{(Q_{M_E+1}-q_m)|h_e|^2+1}{Q_{M_E+1}|h_e|^2+1}\right)^{\beta_m}\right]\right). \quad (20)$$

At this point, a short note on the circumstances under which the ESR can be evaluated is in place. The design of secrecy encoders utilizes so the called (double) binning techniques and relies on full CSI knowledge, i.e., both the legitimate

$$R_e^{(m)} = \begin{cases} \log_2\left(1+\dfrac{\rho|h_e|^2\gamma_m}{\rho|h_e|^2\sum_{i=m+1}^{M}\gamma_i+1}\right), & 1\le m\le M_e, \\ \log_2\left(1+\dfrac{\rho|h_e|^2\gamma_m}{\rho|h_e|^2\sum_{i=M_E+1,i\neq m}^{M}\gamma_i+1}\right), & M_E+1\le m\le M, \end{cases} \tag{17}$$

$$R_s^m = \begin{cases} \left[\log_2\left(1+\dfrac{q_m|h_m|^2}{Q_{m+1}|h_m|^2+1}\right)-\log_2\left(1+\dfrac{q_m|h_e|^2}{Q_{m+1}|h_e|^2+1}\right)\right]^+, & 1\le m\le M_E, \\ \left[\log_2\left(1+\dfrac{q_m|h_m|^2}{Q_{m+1}|h_m|^2+1}\right)-\log_2\left(1+\dfrac{q_m|h_e|^2}{(Q_{M_E+1}-q_m)|h_e|^2+1}\right)\right]^+, & M_E+1\le m\le M, \end{cases} \tag{18}$$

$$R_s^m = \begin{cases} 0, & 1\le m\le M_E, \\ \log_2\left(1+\dfrac{q_m|h_m|^2}{Q_{m+1}|h_m|^2+1}\right)-\log_2\left(1+\dfrac{q_m|h_e|^2}{(Q_{M_E+1}-q_m)|h_e|^2+1}\right), & M_E+1\le m\le M, \end{cases} \tag{19}$$

and the eavesdropper's CSI need to be readily available. This assumption is reasonable in the internal eavesdropper scenario, as the in the NOMA network the source (BS) needs the full CSI to perform the power allocation among the users; indeed, the scenario of a NOMA network with internal eavesdroppers provides an excellent example of how an eavesdropper's CSI can be known to the legitimate transmitter.

On the other hand, in the external eavesdropper case, this assumption is no longer viable; an external passive attacker would indeed have every incentive to conceal themselves and not leak information regarding their actual CSI. However, there is a fundamental difference between a network's secrecy rate and effective secrecy rate. Inspecting the expression in (19) for the ESR, it is clear that it involves an expectation over the distribution of the attacker's channel gains when the legitimate receiver is stronger than the attacker. What is notably different with respect to the evaluation of the secrecy rate, is the fact that in essence only the order – in terms of received SNR – of the eavesdropper among the set of M NOMA users comes into play, as opposed to the case of the secrecy rate in which the exact eavesdropper's SNR needs to be known for the evaluation.

This in turn, is consistent with the way the CSI is fedback to the BS in the uplink of actual systems, such as LTE and NB-IoT, in which instead of the exact CSI and SNR values, an SNR range is determined in the form of a "channel quality indicator" (CQI) [27]. In a realistic setting, it is therefore conceivable that with the aid of artificial noise techniques [28] it is possible to control the range of SNRs that are attainable by the attacker and provide the legitimate users the opportunity to feedback to the BS relevant information regarding the CQI of a potential eavesdropper, therefore removing ambiguities in the evaluation of the ESR.

In terms of the actual design of the secrecy encoders, although it is beyond the scope of the present work, it can be argued that in dense NOMA networks with multiple CQI levels, this information can be taken into account in the design, accounting for the worst case scenario in which the SNR of the eavesdropper is assumed to be in the upper limit of the respective CQI range. Such an approach would of course need to be taken into account in the evaluation of the ESR, but at this point is left as future work.

To calculate E_{cs}^m, we define $z_1 = |h_m|^2$, $z_2 = |h_e|^2$, and note that the joint PDF $f(z_1,z_2) = f_{(m)}(z_1)f(z_2)$.[10] Here, $f_{(m)}(z_1)$ is the PDF of the ordered m^{th} user's channel gains following order statistics and $f(z_2)$ is the PDF of the external adversary's channel gains, which is Rayleigh distributed with unit variance. From the theory of order statistics, we have that

$$f_{(m)}(z_1) = \psi_m f(z_1)\left(1-F(z_1)\right)^{M-m}F(z_1)^{m-1}. \tag{21}$$

Here, $\psi_m = \frac{1}{B(m,M-m+1)}$ and $B(\mu,w)$ is the beta function, i.e., $B(\mu,w) = \Gamma(\mu)\Gamma(w)\big(\Gamma(\mu+w)\big)^{-1}$, where $\Gamma(\mu)=\mu!$, when μ is a positive integer. By inserting the joint PDF $f(z_1,z_2)$ into (20), we provide the following theorem.

Theorem 4: Suppose that there is an external eavesdropper. Assume that the order of the eavesdropper's channel gains among the set of NOMA users is known. For the m^{th} user with $m \ge M_E+1$, its conditional ESR, i.e., E_{cs}^m, can be simplified to (22), while assuming $\gamma_m \le \sum_{i=m+1}^{M}\gamma_i$, where $m\neq\{M-1,M\}$, and $a = M-m+1+s$. At high SNRs, its conditional ESR, i.e., $\lim_{\rho\to\infty} E_{\text{cs}}^m$, equals to $\log_2\left(\frac{Q_m}{Q_{m+1}}\frac{Q_{M_E+1}-q_m}{Q_{M_E+1}}\right)$.

$$\begin{aligned} E_{\text{cs}}^m \approx \frac{1}{\beta_m}\log_2\Bigg(&\psi_m\left(\frac{Q_{m+1}}{Q_m}\right)^{-\beta_m}\Bigg(\sum_{s=0}^{m-1}\binom{m-1}{s}(-1)^s\frac{1}{a} \\ &+\frac{\beta_m q_m}{Q_{m+1}Q_m}\sum_{s=0}^{m-1}\binom{m-1}{s}(-1)^s e^{\frac{a}{Q_m}}E_i\left(-\frac{a}{Q_m}\right)\Bigg) \\ &\times\left(\frac{Q_{M_E+1}-q_m}{Q_{M_E+1}}\right)^{\beta_m}\Bigg(1-\frac{\beta_m q_m e^{\frac{1}{Q_{M_E+1}}}}{(Q_{M_E+1}-q_m)Q_{M_E+1}} \\ &\times E_i\left(-\frac{1}{Q_{M_E+1}}\right)\Bigg)\Bigg). \end{aligned} \tag{22}$$

Proof: See Appendix D. ■

[10]The legitimate NOMA users and the external adversary are independent, so the joint PDF is the product of two marginal PDFs.

$$\breve{R}_s^m = \begin{cases} \log_2\left(1+\dfrac{q_m|h_m|^2}{Q_{m+1}|h_m|^2+1}\right) - \log_2\left(1+\dfrac{q_m|h_e|^2}{Q_{m+1}|h_e|^2+1}\right), & |h_m|^2 \geq |h_e|^2, \\ 0, & \text{otherwise.} \end{cases} \tag{25}$$

Note that when there exists an external eavesdropper, obtaining the conditional ESR, i.e., E_{cs}^m, requires the relative order of the eavesdropper, in terms of the received SNR, among the set of M NOMA users. If this information is not available, the exact value of E_{cs}^m cannot be obtained. Hence, in the following sections, we study and analyze a lower bound and an upper bound for the ESR, which do not require any prior information of the adversary's relative order.

1) Lower Bound on the ESR With an External Eavesdropper: From [21, ch. 15], it is noted that before the adversary detects the m^{th} user's message, if we assume that the first $m-1$ NOMA users' information has already been successfully decoded, then, we overestimate the malicious adversary's decoding capability. This can happen if the external eavesdropper attains the prior information of the first $m-1$ users' CSI. Therefore, an upper bound on $R_e^{(m)}$ can be given by

$$\hat{R}_e^{(m)} = \log_2\left(1+\frac{q_m|h_e|^2}{Q_{m+1}|h_e|^2+1}\right),\ 1 \leq m \leq M. \tag{23}$$

The lower bound on the m^{th} user's achievable secrecy rate can then be expressed as

$$\breve{R}_s^m = \left[R_m - \hat{R}_e^{(m)}\right]^+, \tag{24}$$

which can be extended to (25), shown at the top of this page.

In practice, the external eavesdropper is independent from NOMA users, which means its channel gains can be higher or lower than the m^{th} user. Hence, we aim to provide a lower bound on the ESR for the m^{th} user, which represents an average delay-guaranteed rate that can be at least obtained, no matter whether the eavesdropper has a better channel condition or not. By inserting $\breve{R}_s^m$ into (6), the lower bound on the ESR for the m^{th} user, i.e., $\breve{E}_{\text{s}}^m$, in the presence of an external eavesdropper, can be expressed as

$$\breve{E}_{\text{s}}^m = -\frac{1}{\theta_m T_{\text{f}} B}\ln\left(\iint\limits_{D_1}\left(\frac{Q_m z_1+1}{Q_{m+1}z_1+1}\frac{Q_{m+1}z_2+1}{Q_m z_2+1}\right)^{\beta_m} \times f(z_1,z_2)\,\mathrm{d}z_1\,\mathrm{d}z_2 + \iint\limits_{D_2} f(z_1,z_2)\,\mathrm{d}z_1\,\mathrm{d}z_2\right), \tag{26}$$

where $D_1 = \{(z_1,z_2), z_1 \geq z_2\}$, and $D_2 = \{(z_1,z_2), z_1 < z_2\}$. Then, after applying the joint PDF $f(z_1,z_2)$, $\breve{E}_{\text{s}}^m$ in (26) can be expanded as

$$\breve{E}_{\text{s}}^m = -\frac{1}{\theta_m T_{\text{f}} B}\ln\left(\psi_m \iint\limits_{D_1}\left(\frac{Q_m z_1+1}{Q_{m+1}z_1+1}\frac{Q_{m+1}z_2+1}{Q_m z_2+1}\right)^{\beta_m} \times \frac{e^{-(M-m+1)z_1-z_2}}{\left(1-e^{-z_1}\right)^{1-m}}\,\mathrm{d}z_1\,\mathrm{d}z_2 + \psi_m \iint\limits_{D_2} e^{-(M-m+1)z_1} \times \left(1-e^{-z_1}\right)^{m-1} e^{-z_2}\,\mathrm{d}z_1\,\mathrm{d}z_2\right). \tag{27}$$

To bring more insights, we approximate $\breve{E}_{\text{s}}^m$ at high SNRs and provide the following theorem.

Theorem 5: Suppose that there is an external eavesdropper. The lower bound on the ESR for the m^{th} user, i.e., $\breve{E}_{\text{s}}^m$, can be approximated at high SNRs and given in (28), based on the condition that $\gamma_m \leq \sum_{i=m+1}^{M}\gamma_i$, where $m \neq \{M-1, M\}$.

$$\breve{E}_{\text{s}}^m \approx \frac{1}{\beta_m}\log_2\left(\psi_m\left(\frac{Q_{m+1}}{Q_m}\right)^{-\beta_m}\left(A_1+\frac{\beta_m q_m}{Q_m Q_{m+1}}A_2\right) + \psi_m\sum_{s=0}^{m-1}\binom{m-1}{s}(-1)^s\frac{1}{a+1}\right), \tag{28}$$

where A_1 and A_2 are given by

$$A_1 \approx \left(\frac{Q_{m+1}}{Q_m}\right)^{\beta_m}\sum_{s=0}^{m-1}\binom{m-1}{s}\frac{(-1)^s}{a}\left(\frac{1}{a+1} - \frac{\beta_m q_m}{Q_m Q_{m+1}}e^{\frac{a+1}{Q_m}}E_i\left(-\frac{a+1}{Q_m}\right)\right), \tag{29}$$

$$A_2 \approx \left(\frac{Q_{m+1}}{Q_m}\right)^{\beta_m}\sum_{s=0}^{m-1}\binom{m-1}{s}(-1)^s e^{\frac{a}{Q_m}}\Bigg(e^{\frac{1}{Q_m}} \times\left(E_1\left(\frac{a+1}{Q_m}\right) - e^{-\frac{1}{Q_m}}E_1\left(\frac{a}{Q_m}\right)\right) - \frac{\beta_m q_m}{Q_m Q_{m+1}}e^{\frac{1}{Q_m}} \times\Bigg(\left(r-\ln\left(Q_m\right)+E_1\left(\frac{1}{Q_m}\right)\right)E_1\left(\frac{a}{Q_m}\right) + \frac{1}{2}\left(\zeta(2)+\left(r+\ln\left(\frac{a}{Q_m}\right)\right)^2\right) + e^{-\frac{a}{Q_m}}\sum_{\delta=0}^{\Delta}\frac{e_\delta\left(\frac{a}{Q_m}\right)}{(\delta+1)^2} \times\left(-\frac{1}{a}\right)^{\delta+1} - \frac{a}{Q_m}{}_3\mathbf{F}_3\begin{bmatrix}1,\ 1,\ 1\\ 2,\ 2,\ 2\end{bmatrix}; -\frac{a}{Q_m}\Bigg)\Bigg). \tag{30}$$

and $a = M-m+s+1$. Furthermore, $\zeta(\cdot)$ is the Riemann zeta function, r is the Euler's constant, $e_m(x) = \sum_{s=0}^{m}\frac{x^s}{s!}$, $\Delta \geq 50$,[11] and $E_1(\cdot)$ is the exponential integral function [26]. For comparison purposes, the closed-form expression for the lower bound on ergodic secrecy rate, i.e., $\mathbb{E}[\breve{R}_s^m]$, is given in (31).

$$\mathbb{E}[\breve{R}_s^m] = \frac{\psi_m}{\ln 2}\Bigg(\sum_{s=0}^{m}\binom{m}{s}(-1)^s\frac{1}{a}\left(-e^{\frac{a}{Q_m}}E_i\left(-\frac{a}{Q_m}\right) + e^{\frac{a}{Q_{m+1}}}E_i\left(-\frac{a}{Q_{m+1}}\right)\right) + \sum_{s=0}^{m-1}\binom{m-1}{s}(-1)^s\frac{1}{a(a+1)} \times\left(-e^{\frac{a+1}{Q_{m+1}}}E_i\left(-\frac{a+1}{Q_{m+1}}\right) + e^{\frac{a+1}{Q_m}}E_i\left(-\frac{a+1}{Q_m}\right)\right)\Bigg). \tag{31}$$

Proof: See Appendix E. ∎

[11] Here, Δ is used in a finite sum, which approximates an infinite sum. The complete information is given in Appendix E.

We will demonstrate the validity of the derived analytical closed-forms in Section IV. Furthermore, in the following lemma, we explore the impact of ρ on the proposed lower bounds on the ergodic secrecy rate and the ESR, i.e., $\mathbb{E}[\breve{R}_s^m]$ and $\breve{E}_s^m$, given in (24) and (28).

Lemma 1: When $\rho \to 0$, $\lim_{\rho\to 0} \mathbb{E}[\breve{R}_s^m] = 0$, $\lim_{\rho\to 0} \breve{E}_s^m = 0$. When $\rho \to \infty$, $\lim_{\rho\to \infty} \mathbb{E}[\breve{R}_s^m] = 0$, and $\lim_{\rho\to \infty} \breve{E}_s^m = 0$.

Proof: See Appendix F. ∎

Lemma 1 reveals that if the external eavesdropper has the prior information of the first $m-1$ users' CSI, the m^{th} user's achievable secrecy rate at high SNRs, no matter delay-guaranteed or delay-unguaranteed, becomes zero. Note that the above analysis is for a constant delay exponent. In simulation results, we will show more results for various delay requirements.

2) Upper Bound on the ESR With an External Eavesdropper: If none of the first $m-1$ users' information can be decoded when the eavesdropper intends to decode the m^{th} user's message, then we underestimate the decoding ability of the malicious adversary with SIC employed. This may happen if the eavesdropper cannot attain any prior information of the first $m-1$ users' CSI. Therefore, a lower bound on $R_e^{(m)}$ is given by

$$\breve{R}_e^{(m)} = \log_2\left(1 + \frac{q_m|h_e|^2}{Q_1|h_e|^2+1}\right), \quad 1 \le m \le M. \tag{32}$$

Hence, an upper bound on the secrecy rate, i.e., $\hat{R}_s^m$, can be written as

$$\begin{aligned}
&\hat{R}_s^m \\
&= \left[R_m - \breve{R}_e^{(m)}\right]^+ \\
&= \left[\log_2\left(1 + \frac{q_m|h_m|^2}{Q_{m+1}|h_m|^2+1}\right) - \log_2\left(1+\frac{q_m|h_e|^2}{Q_1|h_e|^2+1}\right)\right]^+ \\
&= \left[\log_2\left(1+\frac{q_m}{Q_{m+1}+\frac{1}{|h_m|^2}}\right) - \log_2\left(1+\frac{q_m}{Q_1+\frac{1}{|h_e|^2}}\right)\right]^+.
\end{aligned} \tag{33}$$

From (33), we can note that $\log_2\left(1+\frac{q_m}{Q_{m+1}+\frac{1}{|h_m|^2}}\right) \ge \log_2\left(1+\frac{q_m}{Q_1+\frac{1}{|h_e|^2}}\right)$, when $|h_e|^2 \le |h_m|^2$. However, when $|h_e|^2 \ge |h_m|^2$, the sign cannot be distinguished. Hence, the upper bound on the ESR for the m^{th} user with an external eavesdropper, i.e., $\hat{E}_s^m$, can only be obtained numerically, by inserting $\hat{R}_s^m$ into (6).

Although the exact analytical closed-form for the proposed upper bound on the ESR is not available, the following lemma is provided to explore the impact of ρ on the upper bounds on the ergodic secrecy rate and the ESR, i.e., $\mathbb{E}[\hat{R}_s^m]$ and $\hat{E}_s^m$.

Lemma 2: When $\rho \to 0$, $\lim_{\rho\to 0} \mathbb{E}[\hat{R}_s^m] = 0$, $\lim_{\rho\to 0} \hat{E}_s^m = 0$. When $\rho \to \infty$, $\lim_{\rho\to \infty} \mathbb{E}[\hat{R}_s^m] = \lim_{\rho\to \infty} \hat{E}_s^m = \log_2\left(\frac{Q_m}{Q_{m+1}}\frac{Q_1}{Q_1+q_m}\right)$.

Proof: See Appendix G. ∎

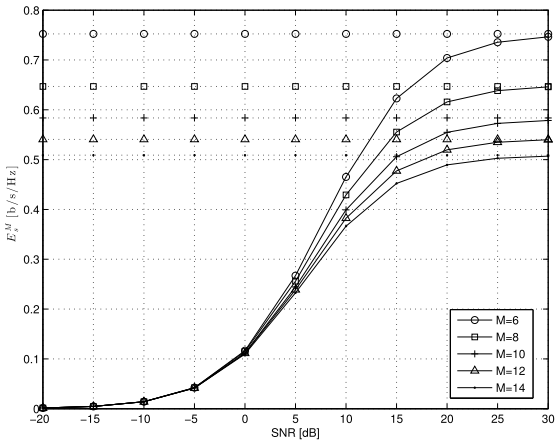


Fig. 1. E_s^M, vs. the transmit SNR ρ, with an internal eavesdropper.

Lemma 2 reveals that if the malicious adversary cannot decode any of the first $m-1$ users' information, the m^{th} user can always achieve constant positive secrecy rates at high SNRs, for both delay-guaranteed and delay-unguaranteed, and the values depend on power coefficients. This result is in agreement with previous analyses in [10] which demonstrated that the secrecy rate in wireless multiuser networks reduces to finite asymptotic value at high SNRs.

IV. Numerical Results

The accuracy of the derived analytical closed-forms and the theoretical analysis given in Section III will be numerically validated in this section. Further, the impact of the delay requirements, the size of the NOMA set and the transmit SNR on the secrecy rate and the ESR will be examined as well, by assuming that a passive internal eavesdropper or an external eavesdropper exists. It is assumed that the bandwidth $B = 100$ kHz, the fading-block length $T_{\text{f}} = 0.01$ ms, the power coefficients are given as $\gamma_i = \frac{M-i+1}{\mu}$ and μ is to ensure $\sum_{i=1}^M \gamma_i = 1$ [23], unless otherwise indicated. Note that a fixed power coefficient setting is adopted in this paper. This is because the main aim of this paper is to provide analytical results and reveal some insights about the delay-guaranteed secrecy rate. In future work, we will consider applying optimal power allocation to improve the system performance through optimally allocating available resources.

Suppose that there is an internal eavesdropper among all NOMA users. To validate the correctness of the analytical closed-form for $\lim_{\rho\to\infty} E_s^M$, given in (15), we depict in Fig. 1 E_s^M versus the transmit SNR ρ, for different values of M. To plot this figure, it is assumed that the power coefficient is set to $\gamma_M = 0.1$ and the delay QoS exponent to $\theta_M = 0.01$. Specifically, the solid lines in Fig. 1 are obtained using Monte Carlo simulations, and the dashed lines are plotted using the closed-form expression, given in (15). From this figure, one can first notice that the proposed closed-form expression is accurate, because the Monte Carlo results at high SNRs converge to the analytical closed-form. Furthermore, at high SNRs, the value of E_s^M achieved with a larger M is smaller

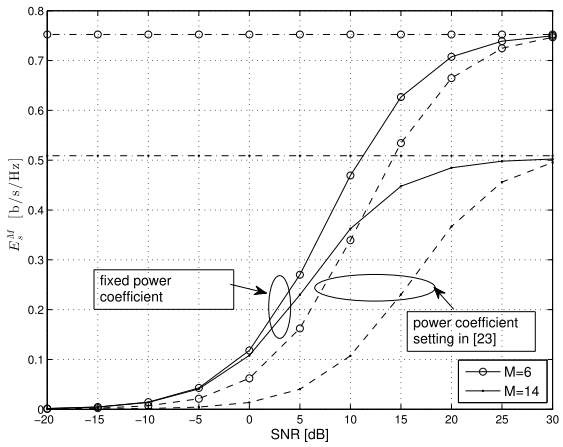


Fig. 2. E_{s}^{M} vs. the transmit SNR ρ, for two different power coefficient settings, with an internal eavesdropper.

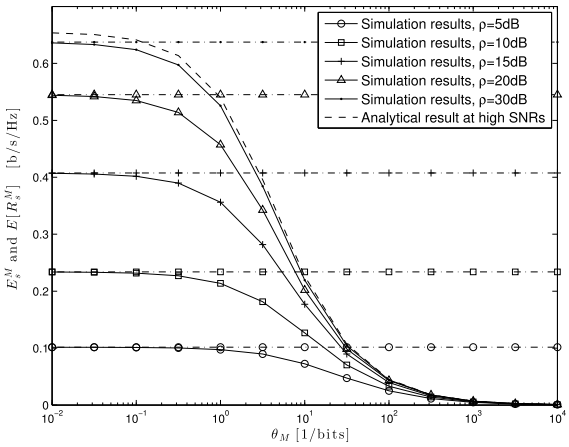


Fig. 3. E_{s}^{M} and $\mathbb{E}[R_s^M]$ vs. θ_M, for different values of ρ, with an internal eavesdropper.

than those obtained with smaller values of M. This indicates that for a larger set of NOMA users, the delay-guaranteed maximum arrival rate that can be served by the strongest user decreases, in the presence of an internal eavesdropper. The reason is that in this case, the second best NOMA user, i.e., the internal eavesdropper, has a high probability of having similar channel conditions with the M^{th} user. Since our analysis in Section III-A show that the M^{th} user's ESR at high SNRs depends on the ratio of channel gains between the M^{th} user and the internal eavesdropper. Hence, we can expect that when the number of NOMA users increases, the M^{th} user will achieve smaller E_{s}^{M} values in the high SNR regime.

Note that Fig. 1 is plotted by setting a fixed power coefficient for the strongest user, i.e., $\gamma_M = 0.1$. What if the power coefficient γ_M is a value which depends on M? To explore the influence of power coefficients, Fig. 2 is depicted which include the curves of E_{s}^{M} versus the transmit SNR, with two different power coefficient settings considered. The solid lines show the curves by applying a fixed power coefficient setting, while the dashed lines are plotted for the varied power coefficient setting given in [23]. This figure first indicates that for fixed values of ρ and M, the E_{s}^{M} obtained with $\gamma_M = 0.1$ is larger than the one obtained with a varied power setting. Further, for a larger value of M, the gap between the solid line and the dashed line is larger. This is due to the fact that by adopting the varied power coefficient setting in [23], γ_M reduces with M, which results in a smaller E_{s}^{M}. Fig. 2 also indicates that for a fixed M, both of the two E_{s}^{M} curves, obtained with different power coefficient settings, converge to the same maximum limit at high SNRs. This numerically validates Theorem 3 in Section III-A, which proves that $\lim_{\rho\to\infty} E_{\mathrm{s}}^{M}$ approaches a constant value, irrespective of the power coefficients.

Recall that for the adopted link-layer channel model, i.e., the theory of EC, the delay exponent θ_M represents the exponential decay rate of the M^{th} user's delay violation probability. With a smaller θ_M, it indicates a slower decay rate, which allows a looser delay guarantee. Meanwhile, a more stringent delay provisioning can be represented by a larger θ_M [6]. Hence, we depict Fig. 3 which plots E_{s}^{M} and $\mathbb{E}[R_s^M]$ versus θ_M, for different values of ρ, so that the impact of θ_M can be investigated. The solid lines are plotted for E_{s}^{M}, while the dash-dotted lines are plotted for $\mathbb{E}[R_s^M]$. All solid lines and the dash-dotted lines are simulated using Monte Carlo method. Furthermore, the dashed line in this figure is plotted using the analytical closed-form for $\lim_{\rho\to\infty} E_{\mathrm{s}}^{M}$, given in (15). Firstly, Fig. 3 shows that for a fixed ρ, the value of E_{s}^{M} decreases with θ_M, and approaches 0 when the value of θ_M becomes very large. This confirms the monotonicity proof given in Theorem 1, which indicates that, a user with a stringent delay requirement will have to settle for a smaller delay-guaranteed secrecy rate, compared to one with a loose delay constraint. To the best of our knowledge, it is the first time that a delay-constrained secrecy rate analysis has been performed and the trade-off between them is discussed for a NOMA network. Furthermore, Fig. 3 also shows that when $\theta_M \to 0$, the E_{s}^{M} value matches with the ergodic secrecy rate $\mathbb{E}[R_s^M]$. This validates the theoretical conclusion proposed in Theorem 1, which proves that the ESR converges to the ergodic secrecy rate, when there is no delay constraint. Finally, from Fig. 3, we can also notice that when the transmit SNR ρ becomes larger, E_{s}^{M} gradually increases and approaches the analytical limit, i.e., the dashed line.

Suppose that there is an external eavesdropper distinct from the set of NOMA users. Fig. 4 plots the ESR for the m^{th} user, i.e., E_{s}^{m}, versus ρ, for different values of M. This figure aims to investigate the influence of the size of the NOMA user set on the m^{th} user's ESR. To plot this figure, we assume that the eavesdropper intends to decode the 4^{th} user's messages, i.e., $m = 4$. From Fig. 4, it is noted that when ρ increases, the value of E_{s}^{m} first increases, then becomes stable at high SNRs. Further, when M becomes larger, E_{s}^{m} reduces, which shows the same trend with Fig. 1. This indicates that when the size of the NOMA user set increases, the delay-guaranteed maximum arrival rate that can be securely served decreases, when there exists an external eavesdropper. Contrary to previous work [10] with multiple users in which only the best user can be served by the BS, in a NOMA network with power settings

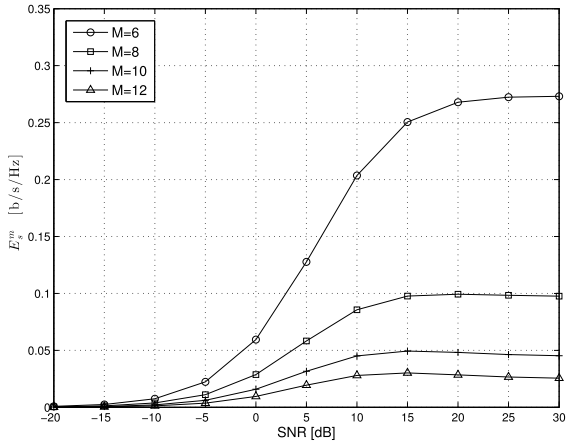


Fig. 4. $E_{\rm s}^m$ vs. the transmit SNR ρ, in the presence of an external eavesdropper.

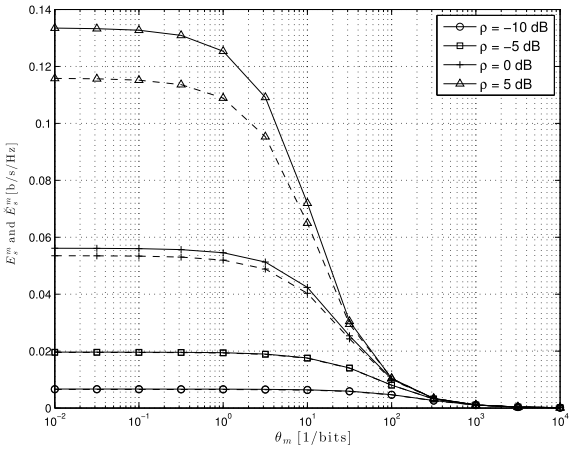


Fig. 6. $E_{\rm s}^m$ and $\breve{E}_{\rm s}^m$ vs. the delay exponent θ_m, for different values of ρ, with an external eavesdropper.

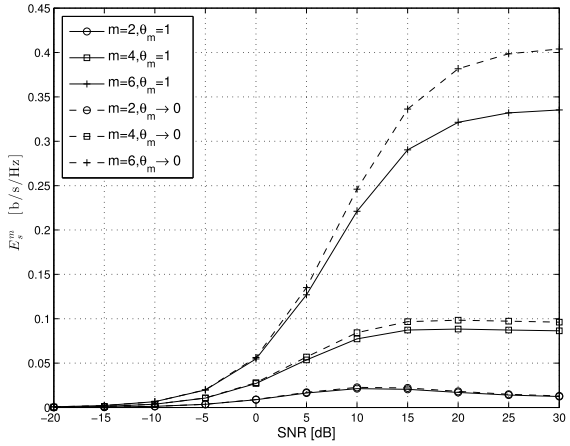


Fig. 5. $E_{\rm s}^m$ vs. the transmit SNR ρ, for different values of m and θ_m, with an external eavesdropper.

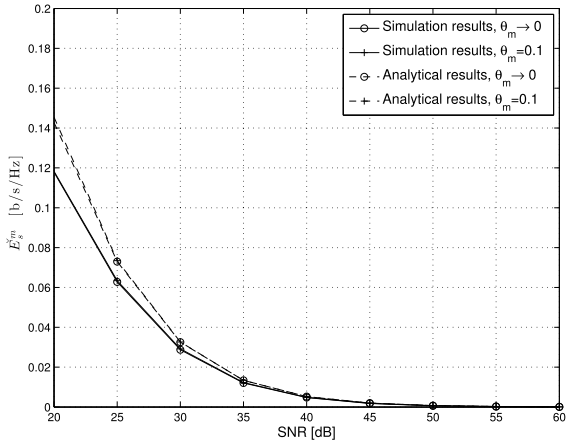


Fig. 7. $\breve{E}_{\rm s}^m$ vs. the transmit SNR ρ, for different θ_m values, with an external eavesdropper.

given in [23], increasing M will reduce the power available to each NOMA user, thus causing a decrease on the secrecy rate and the ESR.

To investigate the ESR for different users with various delay requirements, Fig. 5 plots the curves of $E_{\rm s}^m$ versus the transmit SNR, for various settings of m and θ_m. This figure first shows that for a larger value of m, the $E_{\rm s}^m$ value is larger. This indicates that when there is an external eavesdropper, the user with stronger channel conditions can achieve a higher delay-guaranteed rate. Furthermore, Fig. 5 also shows that for a specific user, the $E_{\rm s}^m$ obtained with $\theta_m \to 0$ is larger than the one achieved with $\theta_m = 1$. This is because the scenario of $\theta_m \to 0$ represents a no-delay-guaranteed situation, in which the delay violation probability approaches 1. Fig. 5 further shows that the gap of the $E_{\rm s}^m$ values between $\theta_m = 1$ and $\theta_m \to 0$ is larger for a larger value of m. This implies that a user with higher channel gains will have to make more sacrifices on its ESR value, so that the required statistical delay constraint can be satisfied.

In Section III-B.1, we proposed and analyzed a lower bound on the ESR for the $m^{\rm th}$ user, denoted as $\breve{E}_{\rm s}^m$, by overestimating the decoding capability of the external eavesdropper. Here, we include Fig. 6 which plots the curves of $E_{\rm s}^m$ (in solid lines) and the lower bound $\breve{E}_{\rm s}^m$ (in dashed lines) versus θ_m, for different values of ρ. To plot this figure, it is assumed that there are 8 NOMA users in total, i.e., $M = 8$, and the external eavesdropper intends to decode the $6^{\rm th}$ user's messages, i.e., $m = 6$. All curves shown in this figure are obtained using Monte Carlo simulation results. From Fig. 6, we first notice that both $E_{\rm s}^m$ and $\breve{E}_{\rm s}^m$ decrease with θ_m, which indicates that with an external eavesdropper existing, the achievable delay-guaranteed secrecy rate becomes smaller, when the user's delay requirement becomes more stringent. This numerically confirms the theoretical analysis given in Theorem 1. Furthermore, Fig. 6 shows that the proposed lower bound $\breve{E}_{\rm s}^m$ serves as a good lower bound for small SNR regime, and with the decrease of ρ, the gap between $E_{\rm s}^m$ and $\breve{E}_{\rm s}^m$ reduces.

To validate the accuracy of the closed-form expression for $\breve{E}_{\rm s}^m$, we include Fig. 7 which shows $\breve{E}_{\rm s}^m$ versus ρ, for various values of θ_m. The solid lines are plotted using the Monte Carlo results, while the dashed lines are shown using the

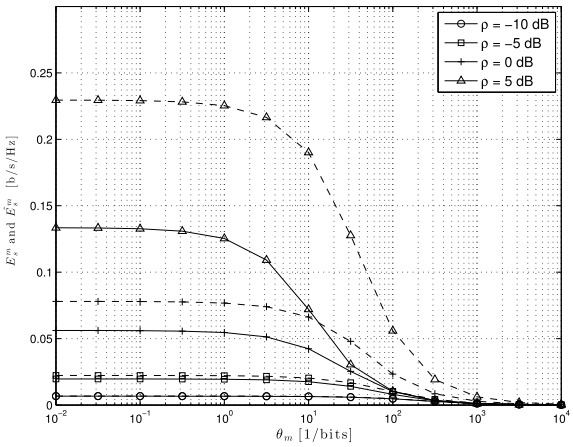


Fig. 8. $E_{\rm s}^m$ and $\hat{E_{\rm s}^m}$ vs. the delay exponent θ_m, for different values of ρ, with an external eavesdropper.

analytical results given in Theorem 5. Note that the given closed-form expression is valid only when $\gamma_m \leq \sum_{i=m+1}^{M} \gamma_i$, with $m \neq \{M-1, M\}$. Hence, to plot Fig. 7, we apply the power coefficient settings given in [23], i.e., $\gamma_i = \frac{M-i+1}{\mu}$, where μ is to ensure $\sum_{i=1}^{M} \gamma_i = 1$. By setting $M = 8$, $m = 6$, we can calculate the power coefficient values and find that $\gamma_m \leq \sum_{i=m+1}^{M} \gamma_i$ is satisfied. Fig. 7 first shows that when the transmit SNR gradually increases, the analytical closed-form results match with the Monte Carlo results, which confirms the validity of the derived closed-form at high SNRs. Furthermore, both the analytical and simulation results approach 0, at high SNRs. This confirms Lemma 1 in Section III-B.1.

In Section III-B.2, by underestimating the external eavesdropper's decoding capability, we proposed an upper bound on the ESR for the $m^{\rm th}$ user, denoted as $\hat{E_{\rm s}^m}$. To confirm the validity of the upper bound and to further investigate, Fig. 8 is plotted which includes the simulated $E_{\rm s}^m$ values (in solid lines) and the upper bound $\hat{E_{\rm s}^m}$ (in dashed lines) versus θ_m, for different values of ρ. Similar to Fig. 6, Fig. 8 also shows that both the $E_{\rm s}^m$ and the upper bound $\hat{E_{\rm s}^m}$ decrease with θ_m. This is due to the fact that with a larger θ_m indicating a more stringent statistical delay guarantee, the maximum achievable arrival rate which can be securely supported becomes smaller [6]. Furthermore, Fig. 8 also shows that the proposed $\hat{E_{\rm s}^m}$ can serve as a good upper bound, for relatively small transmit SNR values.

V. Conclusion

The delay-guaranteed secrecy rate, namely, ESR, has been introduced and investigated for a downlink NOMA network; ESR represents the maximum constant arrival rate which can be securely served by a legitimate user, while guaranteeing the required statistical delay constraints. Two eavesdropping scenarios have been considered: a purely antagonistic network with an unknown internal eavesdropper and a trustworthy NOMA network with an external eavesdropper. Assuming an internal eavesdropper exists, a closed-form expression for the ESR at high SNRs has been derived for the strongest user, which is the only user guaranteed to have a positive ESR in this case. Assuming that an external eavesdropper exists and has interest in jeopardizing the security of the network, a lower bound and an upper bound on the ESR have been proposed, respectively, and have been shown to be tight in the low SNR regime. Simulation results have shown that for both eavesdropping scenarios, a user with a stringent delay requirement serves a smaller amount of ESR, comparing to those with relatively loose delay constraints. Further, it has been shown that a legitimate NOMA user with higher channel gains will make greater sacrifices on its ESR value, so that a required statistical delay guarantee can be satisfied.

Appendix A
Proof for Theorem 1

Recall that the ESR for the $m^{\rm th}$ user, i.e., $E_{\rm s}^m$, is calculated by inserting the achievable secrecy rate R_s^m into (6), which is then given by

$$E_{\rm s}^m = -\frac{1}{\theta_m T_{\rm f} B} \ln\left(\mathbb{E}\left[e^{-\theta_m T_{\rm f} B R_s^m}\right]\right). \tag{34}$$

Note that $e^{-\theta_m T_{\rm f} B R_s^m}$ is a log-convex function in θ_m because $\ln\left(e^{-\theta_m T_{\rm f} B R_s^m}\right) = -\theta_m T_{\rm f} B R_s^m$ is a convex function in θ_m [29]. Since the log-convexity still holds under summations, therefore we could conclude that $\mathbb{E}\left[e^{-\theta_m T_{\rm f} B R_s^m}\right]$ is also a log-convex function in θ_m. Hence, it is clear that $\ln\left(\mathbb{E}\left[e^{-\theta_m T_{\rm f} B R_s^m}\right]\right)$ is convex, which means $-\ln\left(\mathbb{E}\left[e^{-\theta_m T_{\rm f} B R_s^m}\right]\right)$ is a concave function in θ_m.

We rewrite $E_{\rm s}^m$ as $\frac{f(\theta_m)}{g(\theta_m)}$, where $f(\theta_m) = -\ln\left(\mathbb{E}\left[e^{-\theta_m T_{\rm f} B R_s^m}\right]\right)$ and $g(\theta_m) = \theta_m T_{\rm f} B$. In order to prove that $E_{\rm s}^m$ is a monotonically decreasing function in θ_m, we take the first derivative of $E_{\rm s}^m$, which gives

$$\frac{\partial E_{\rm s}^m}{\partial \theta_m} = \left(\frac{f(\theta_m)}{g(\theta_m)}\right)' = \frac{f'(\theta_m)g(\theta_m) - g'(\theta_m)f(\theta_m)}{\left(g(\theta_m)\right)^2}. \tag{35}$$

Apparently, the denominator is a non-negative value. Let us consider the numerator only. We take the first derivative of the denominator and it gives

$$\begin{aligned}&\left(f'(\theta_m)g(\theta_m) - g'(\theta_m)f(\theta_m)\right)' \\ &= f''(\theta_m)g(\theta_m) - g''(\theta_m)f(\theta_m),\end{aligned} \tag{36}$$

which can be simplified as $f''(\theta_m)g(\theta_m)$ because $g''(\theta_m) = 0$. Since we have proved that $f(\theta_m)$ is a concave function, i.e., $f''(\theta_m) \leq 0$, and also $g(\theta_m) \geq 0$, therefore it can be concluded that the numerator in (35) is a non-increasing function. Furthermore, it is easy to see that the numerator in (35) equals to 0 when $\theta_m = 0$, i.e., $f'(\theta_m)g(\theta_m) - g'(\theta_m)f(\theta_m)|_{\theta_m=0} = 0$. Finally, we can conclude that $\frac{\partial E_{\rm s}^m}{\partial \theta_m} \leq 0$, which implies that $E_{\rm s}^m$ monotonically decreases with θ_m.[12]

[12]The conclusion of monotonicity is obtained by excluding the possibility of $E_{\rm s}^m$ being a constant value, with respect to θ_m.

When $\theta_m \to 0$, we can obtain that

$$\lim_{\theta_m \to 0} E_{\mathrm{s}}^m = \lim_{\theta_m \to 0} -\frac{\left(\mathbb{E}\left[e^{-\theta_m T_{\mathrm{f}} B R_s^m}\right]\right)'}{T_{\mathrm{f}} B \mathbb{E}\left[e^{-\theta_m T_{\mathrm{f}} B R_s^m}\right]} \tag{37}$$

$$= \lim_{\theta_m \to 0} -\frac{\mathbb{E}\left[e^{-\theta_m T_{\mathrm{f}} B R_s^m}\left(-T_{\mathrm{f}} B R_s^m\right)\right]}{T_{\mathrm{f}} B \mathbb{E}\left[e^{-\theta_m T_{\mathrm{f}} B R_s^m}\right]} \tag{38}$$

$$= \mathbb{E}\left[R_s^m\right]. \tag{39}$$

Hence, this proves that when $\theta_m \to 0$, E_{s}^m converges to the ergodic secrecy rate $\mathbb{E}\left[R_s^m\right]$.

When $\theta_m \to \infty$, from (5), one can note that the probability of the delay exceeding a given delay bound approaches zero. It means that the user cannot tolerate any delay outage, which refers to the delay-limited scenario. According to [30], it shows that Rayleigh channel cannot support very stringent delay QoS requirement (when θ is extremely large), even using the optimal power policy. Our simulation results also confirm that the ESR becomes zero in this case.

APPENDIX B
PROOF FOR THEOREM 2

According to the theory of order statistics and by inserting (10) into (9), E_{s}^M can be written as

$$E_{\mathrm{s}}^M = \frac{1}{\beta_M}\log_2\left(M(M-1)\int_0^\infty\int_x^\infty\left(\frac{1+q_M y}{1+q_M x}\right)^{\beta_M} \times f(x)\left(F(x)\right)^{M-2} f(y)\,\mathrm{d}y\,\mathrm{d}x\right)$$

$$= \frac{1}{\beta_M}\log_2\left(M(M-1)\int_0^\infty\left(\frac{1}{1+q_M x}\right)^{\beta_M} e^{-x} \times\left(1-e^{-x}\right)^{M-2}\int_x^\infty\left(1+q_M y\right)^{\beta_M} e^{-y}\,\mathrm{d}y\,\mathrm{d}x\right). \tag{40}$$

Consider $E_{sc}^1 = \int_x^\infty \left(1+q_M y\right)^{\beta_M} e^{-y}\,\mathrm{d}y$ first. By setting $z_1 = \frac{1}{q_M} + y$, E_{sc}^1 can be transformed into $E_{sc}^1 = q_M^{\beta_M} e^{\frac{1}{q_M}} \int_{\frac{1}{q_M}+x}^\infty z_1^{\beta_M} e^{-z_1}\,\mathrm{d}z_1$. Then, from (3.381.6) in [31], we note that

$$\int_u^\infty \frac{e^{-x}}{x^v}\,\mathrm{d}x = u^{-\frac{v}{2}} e^{-\frac{u}{2}} W_{-\frac{v}{2},\frac{1-v}{2}}(u) \qquad [u>0], \tag{41}$$

where $W_{k,\mu}(z)$ is the Whittaker W function [26]. By applying (41), E_{sc}^1 can be given as

$$E_{sc}^1 = q_M^{\beta_M} e^{\frac{1}{q_M}}\left(\frac{1}{q_M}+x\right)^{\frac{\beta_M}{2}} e^{-\frac{1}{2q_M}-\frac{x}{2}} W_{\frac{\beta_M}{2},\frac{1+\beta_M}{2}}\left(\frac{1}{q_M}+x\right). \tag{42}$$

Finally, by inserting E_{sc}^1 into (40), E_{s}^M can be given as

$$E_{\mathrm{s}}^M = \frac{1}{\beta_M}\log_2\left(M(M-1)q_M^{\frac{\beta_M}{2}} e^{\frac{1}{2q_M}}\int_0^\infty\left(1+q_M x\right)^{-\frac{\beta_M}{2}} \times e^{-\frac{3}{2}x}\left(1-e^{-x}\right)^{M-2} W_{\frac{\beta_M}{2},\frac{1+\beta_M}{2}}\left(\frac{1}{q_M}+x\right)\mathrm{d}x\right)$$

$$= A_M + \frac{1}{\beta_M}\log_2\left(\int_0^\infty\left(1+q_M x\right)^{-\beta_M} e^{-x} \times\left(1-e^{-x}\right)^{M-2}\Gamma\left(1+\beta_M, \frac{1}{q_M}+x\right)dx\right), \tag{43}$$

where $A_M = \frac{1}{\beta_M}\log_2\left(M(M-1)q_M^{\beta_M} e^{\frac{1}{q_M}}\right)$, and $\Gamma(\cdot,\cdot)$ is the incomplete Γ function. Making use of the Binomial theorem we have that

$$\left(1-e^{-x}\right)^{M-2} = \sum_{\nu=0}^{M-2}\binom{M-2}{\nu}(-1)^\nu e^{-\nu x}, \tag{44}$$

so that the following integral appears

$$\int_0^\infty\left(1+q_M x\right)^{-\beta_M} e^{-(\nu+1)x}\Gamma\left(1+\beta_M, \frac{1}{q_M}+x\right)\mathrm{d}x$$

$$= \int_{1/q_M}^\infty\left(q_M z\right)^{-\beta_M} e^{-(\nu+1)(z-\frac{1}{q_M})}\Gamma(1+\beta_M, z)\,\mathrm{d}z \tag{45a}$$

$$= q_M^{-\beta_M} e^{\frac{\nu+1}{q_M}}\int_{1/q_M}^\infty z^{-\beta_M} e^{-(\nu+1)z}\Gamma(1+\beta_M, z)\,\mathrm{d}z, \tag{45b}$$

by change of variable $z = x + \frac{1}{q_M}$. We set $I_\nu = \int_{1/q_M}^\infty z^{-\beta_M} e^{-(\nu+1)z}\Gamma(1+\beta_M, z)\,\mathrm{d}z$, so that

$$E_{sc}^M = B_M + \frac{1}{\beta_M}\log_2\left(\sum_{\nu=0}^{M-2}\binom{M-2}{\nu}(-1)^\nu e^{\frac{\nu}{q_M}} I_\nu\right), \tag{46}$$

where $B_M = \frac{\log_2(M(M-1))+2q_M^{-1}}{\beta_M}$. To evaluate I_ν we will use the following property

$$\int x^b\Gamma(s,x)\,\mathrm{d}x = \frac{1}{b+1}\left(x^{b+1}\Gamma(s,x) - \Gamma(s+b+1,x)\right), \tag{47}$$

and note that the limit of the right-hand side (RHS) for $x \to \infty$ is 0. To have only powers of z in I_ν, we resort in using the Taylor series expansion for the exponential function $e^{-(\nu+1)z} = \sum_{k=0}^\infty \frac{(-1)^k(\nu+1)^k z^k}{k!}$. Hence, I_ν becomes

$$I_\nu = \sum_{k=0}^\infty\frac{(-1)^k(\nu+1)^k}{k!}\int_{\frac{1}{q_M}}^\infty z^{k-\beta_M}\Gamma\left(1+\beta_M, z\right)\mathrm{d}x$$

$$= -\sum_{k=0}^\infty\frac{(-1)^k}{k!}\frac{(\nu+1)^k}{k-\beta_M+1}\left[\left(\frac{1}{q_M}\right)^{k-\beta_M+1} \times\Gamma\left(1+\beta_M, \frac{1}{q_M}\right) - \Gamma\left(1+\beta_M+k-\beta_M+1, \frac{1}{q_M}\right)\right]$$

$$= \sum_{k=0}^\infty\frac{(-1)^k}{k!}\frac{(\nu+1)^k}{k-\beta_M+1}\left[\Gamma\left(k+2, \frac{1}{q_M}\right) - \left(\frac{1}{q_M}\right)^{k-\beta_M+1}\Gamma\left(1+\beta_M, \frac{1}{q_M}\right)\right], \tag{48}$$

and finally

$$E_{sc}^M = B_M + \frac{1}{\beta_M}\log_2\left(\sum_{\nu=0}^{M-2}\binom{M-2}{\nu}(-1)^\nu e^{\frac{\nu}{q_M}}\right.$$

$$\times \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \frac{(\nu+1)^k}{k-\beta_M+1} \left[\Gamma\left(k+2, \frac{1}{q_M}\right) - \left(\frac{1}{q_M}\right)^{k-\beta_M+1} \Gamma\left(1+\beta_M, \frac{1}{q_M}\right) \right] \right). \quad (49)$$

APPENDIX C
PROOF FOR THEOREM 3

By applying the theory of order statistics and inserting (10) into (14), the M^{th} user's ESR at high SNRs can be given by

$$\lim_{\rho\to\infty} E_{\text{s}}^M = \frac{1}{\beta_M} \log_2 \left(M(M-1) \int_0^\infty \int_x^\infty \left(\frac{y}{x}\right)^{\beta_M} \times f(x) \left(F(x)\right)^{M-2} f(y)\, \mathrm{d}y\, \mathrm{d}x \right) \quad (50a)$$

$$= \frac{1}{\beta_M} \log_2 \left(M(M-1) \int_0^\infty \left(\frac{1}{x}\right)^{\beta_M} e^{-x} \left(1-e^{-x}\right)^{M-2} \times \int_x^\infty y^{\beta_M} e^{-y}\, \mathrm{d}y\, \mathrm{d}x \right). \quad (50b)$$

Then, by applying (41) to (50b), one can get that

$$\lim_{\rho\to\infty} E_{\text{s}}^M = \frac{1}{\beta_M} \log_2 \left(M(M-1) \int_0^\infty \left(\frac{1}{x}\right)^{\beta_M} e^{-x} \times \left(1-e^{-x}\right)^{M-2} x^{\frac{\beta_M}{2}} e^{-\frac{x}{2}} W_{\frac{\beta_M}{2}, \frac{1+\beta_M}{2}}(x)\, \mathrm{d}x \right). \quad (51)$$

Further, by using the binomial expansion and expanding $(1-e^{-x})^{M-2}$ as $\sum_{s=0}^{M-2} \binom{M-2}{s} (-1)^s e^{-xs}$, (51) can be transformed into

$$\lim_{\rho\to\infty} E_{\text{s}}^M = \frac{1}{\beta_M} \log_2 \left(M(M-1) \sum_{s=0}^{M-2} \binom{M-2}{s} (-1)^s \times \int_0^\infty x^{-\frac{\beta_M}{2}} e^{-\left(\frac{3}{2}+s\right)x} W_{\frac{\beta_M}{2}, \frac{1+\beta_M}{2}}(x)\, \mathrm{d}x \right). \quad (52)$$

From (13.23.4) in [26], we note that

$$\int_0^\infty e^{-zt} t^{w-1} W_{k,\mu}(t)\, \mathrm{d}t = \Gamma\left(\frac{1}{2}+\mu+w\right) \Gamma\left(\frac{1}{2}-\mu+w\right) \times {}_2\mathbf{F}_1 \left[\begin{matrix} \frac{1}{2}-\mu+w,\ \frac{1}{2}+\mu+w \\ w-k+1 \end{matrix} ; \frac{1}{2}-z \right] \times \left[\mathrm{Re}\left(w+\frac{1}{2}\right) > |\mathrm{Re}(\mu)|, \mathrm{Re}(z) > -\frac{1}{2} \right], \quad (53)$$

where ${}_2\mathbf{F}_1 \left[\begin{smallmatrix} a,\ b \\ c \end{smallmatrix} ; z \right]$ is the generalized hypergeometric function, and $\Gamma(\cdot)$ is the gamma function. By applying (53) to (52), the closed-form expression for $\lim_{\rho\to\infty} E_{\text{s}}^M$ can be finally expressed as

$$\lim_{\rho\to\infty} E_{\text{s}}^M = \frac{1}{\beta_M} \log_2 \left(M(M-1)\Gamma(1-\beta_M) \times \sum_{s=0}^{M-2} \binom{M-2}{s} (-1)^s {}_2\mathbf{F}_1 \left[\begin{matrix} 1-\beta_M,\ 2 \\ 2-\beta_M \end{matrix} ; -1-s \right] \right).$$

Then, for comparison purposes, here we derive the closed-form expression for the ergodic secrecy rate for the M^{th} user at high SNRs, i.e., $\lim_{\rho\to\infty} \mathbb{E}[R_s^M]$. Firstly, we note that $\lim_{\rho\to\infty} \mathbb{E}[R_s^M] = \mathbb{E}\left[\log_2\left(\frac{|h_M|^2}{|h_{M-1}|^2}\right)\right]$, which can be expanded as follows, after inserting the joint PDF (10).

$$\lim_{\rho\to\infty} \mathbb{E}[R_s^M] = M(M-1) \int_0^\infty \int_x^\infty \log_2\left(\frac{y}{x}\right) e^{-x} \times (1-e^{-x})^{M-2} e^{-y}\, \mathrm{d}y\, \mathrm{d}x. \quad (54)$$

By defining $z_0 = \frac{y}{x}$ and $1 \le z_0 \le \infty$, (54) can be rewritten as

$$\lim_{\rho\to\infty} \mathbb{E}[R_s^M] = M(M-1) \int_0^\infty x e^{-x} (1-e^{-x})^{M-2} \times \int_1^\infty \log_2(z_0) e^{-xz_0}\, \mathrm{d}z_0\, \mathrm{d}x \quad (55)$$

$$= \frac{M(M-1)}{\ln 2} \int_0^\infty x e^{-x} (1-e^{-x})^{M-2} \times \int_1^\infty \ln(z_0) e^{-xz_0}\, \mathrm{d}z_0\, \mathrm{d}x. \quad (56)$$

From (4.331.2) in [31], we have that

$$\int_1^\infty e^{-\mu x} \ln x\, \mathrm{d}x = -\frac{1}{\mu} E_i(-\mu), \quad \mathrm{Re}\ \mu > 0. \quad (57)$$

By applying (57), we get that

$$\lim_{\rho\to\infty} \mathbb{E}[R_s^M] = \frac{M(M-1)}{\ln 2} \sum_{s=0}^{M-2} \binom{M-2}{s} (-1)^s \times \int_0^\infty e^{-(s+1)x} E_1(x)\, \mathrm{d}x, \quad (58)$$

obtained after using $(1-e^{-x})^{M-2} = \sum_{s=0}^{M-2} \binom{M-2}{s} (-1)^s e^{-xs}$ and $E_i(-x) = -E_1(x)$, for $x > 0$. From (4.2.3) in [32], we have that

$$\int_0^\infty e^{-ax} E_1(bx)\, \mathrm{d}x = \frac{1}{a} \ln\left(1+\frac{a}{b}\right). \quad (59)$$

Hence, by applying (59), we finally obtain the close-form expression for the ergodic secrecy rate at high SNRs, given as

$$\lim_{\rho\to\infty} \mathbb{E}[R_s^M] = M(M-1) \times \sum_{s=0}^{M-2} \binom{M-2}{s} (-1)^s \frac{1}{s+1} \log_2(s+2). \quad (60)$$

APPENDIX D
PROOF FOR THEOREM 4

By inserting the joint PDF $f(z_1, z_2)$ into (20), E_{cs}^m is given by

$$E_{\text{cs}}^m = \frac{1}{\beta_m}\log_2\left(\int_0^\infty\int_0^\infty\left(\frac{Q_m z_1+1}{Q_{m+1}z_1+1}\times\frac{(Q_{M_E+1}-q_m)\,z_2+1}{Q_{M_E+1}z_2+1}\right)^{\beta_m} f(z_1,z_2)\,\mathrm{d}z_1\,\mathrm{d}z_2\right). \tag{61}$$

Here, $f(z_1,z_2) = f_{(m)}(z_1)f(z_2)$, where $f_{(m)}(z_1) = \psi_m f(z_1)F(z_1)^{m-1}\left(1-F(z_1)\right)^{M-m}$, $f(z_1) = e^{-z_1}$, $F(z_1) = 1-e^{-z_1}$, and $f(z_2) = e^{-z_2}$. Then, E_{cs}^m can be extended as

$$E_{\text{cs}}^m = \frac{1}{\beta_m}\log_2\left(\psi_m\int_0^\infty\left(\frac{(Q_{M_E+1}-q_m)\,z_2+1}{Q_{M_E+1}z_2+1}\right)^{\beta_m} e^{-z_2}\times\int_0^\infty\left(\frac{Q_m z_1+1}{Q_{m+1}z_1+1}\right)^{\beta_m} e^{-(M-m+1)z_1}\times\left(1-e^{-z_1}\right)^{m-1}\mathrm{d}z_1\,\mathrm{d}z_2\right). \tag{62}$$

By replacing $\left(1-e^{-z_1}\right)^{m-1}$ with binominal expansion $\sum_{s=0}^{m-1}\binom{m-1}{s}(-1)^s e^{-z_1 s}$ and defining $a = M-m+s+1$, we can get that

$$E_{\text{cs}}^m = \frac{1}{\beta_m}\log_2\left(\psi_m\int_0^\infty\left(\frac{(Q_{M_E+1}-q_m)\,z_2+1}{Q_{M_E+1}z_2+1}\right)^{\beta_m} e^{-z_2}\times\sum_{s=0}^{m-1}\binom{m-1}{s}(-1)^s\int_0^\infty\left(\frac{Q_m z_1+1}{Q_{m+1}z_1+1}\right)^{\beta_m}\times e^{-az_1}\,\mathrm{d}z_1\,\mathrm{d}z_2\right). \tag{63}$$

To simplify the above equation, we define A_{D_1} and A_{D_2} as follows and E_{cs}^m can be written as

$$E_{\text{cs}}^m = \frac{1}{\beta_m}\log_2\left(\psi_m A_{D_1}A_{D_2}\right), \tag{64a}$$

$$A_{D_1} = \sum_{s=0}^{m-1}\binom{m-1}{s}(-1)^s\int_0^\infty\left(\frac{Q_m z_1+1}{Q_{m+1}z_1+1}\right)^{\beta_m} e^{-az_1}\,\mathrm{d}z_1, \tag{64b}$$

$$A_{D_2} = \int_0^\infty\left(\frac{(Q_{M_E+1}-q_m)\,z_2+1}{Q_{M_E+1}z_2+1}\right)^{\beta_m} e^{-z_2}\,\mathrm{d}z_2. \tag{64c}$$

Let us focus on A_{D_1} first. It can be further expressed as

$$A_{D_1} = \left(\frac{Q_{m+1}}{Q_m}\right)^{-\beta_m}\int_0^\infty\sum_{s=0}^{m-1}\binom{m-1}{s}(-1)^s\times\left(1+\frac{\frac{q_m}{Q_{m+1}}}{Q_m z_1+1}\right)^{-\beta_m} e^{-az_1}\,\mathrm{d}z_1. \tag{65a}$$

By assuming that $q_m \le Q_{m+1}$, where $m \neq \{M-1, M\}$, $\left(1+\frac{\frac{q_m}{Q_{m+1}}}{Q_m z_1+1}\right)^{-\beta_m}$ can be approximated using the first two terms of generalized binomial expansion. Then, (65a) can be approximated as:

$$A_{D_1} \approx \left(\frac{Q_{m+1}}{Q_m}\right)^{-\beta_m}\left(\int_0^\infty\sum_{s=0}^{m-1}\binom{m-1}{s}(-1)^s e^{-az_1}\,\mathrm{d}z_1 - \frac{\beta_m q_m}{Q_{m+1}}\sum_{s=0}^{m-1}\binom{m-1}{s}(-1)^s\int_0^\infty\frac{e^{-az_1}}{Q_m z_1+1}\,\mathrm{d}z_1\right). \tag{66}$$

From (3.352.4) in [31], we have that

$$\int_0^\infty\frac{e^{-\mu x}}{x+\beta}\,\mathrm{d}x = -e^{\beta\mu}E_i(-\mu\beta), |\arg\beta|<\pi, \quad \text{Re }\mu>0. \tag{67}$$

Then, by applying (67), A_{D_1} becomes

$$A_{D_1} \approx \left(\frac{Q_{m+1}}{Q_m}\right)^{-\beta_m}\left(\sum_{s=0}^{m-1}\binom{m-1}{s}(-1)^s\frac{1}{a}+\frac{\beta_m q_m}{Q_{m+1}Q_m}\times\sum_{s=0}^{m-1}\binom{m-1}{s}(-1)^s e^{\frac{a}{Q_m}}E_i\left(-\frac{a}{Q_m}\right)\right). \tag{68}$$

Then, we can start to consider A_{D_2}, which can be expressed as

$$A_{D_2} = \left(\frac{Q_{M_E+1}-q_m}{Q_{M_E+1}}\right)^{\beta_m}\int_0^\infty\left(1+\frac{\frac{q_m}{Q_{M_E+1}-q_m}}{Q_{M_E+1}z_2+1}\right)^{\beta_m}\times e^{-z_2}\,\mathrm{d}z_2 \tag{69}$$

$$\approx \left(\frac{Q_{M_E+1}-q_m}{Q_{M_E+1}}\right)^{\beta_m}\left(\int_0^\infty e^{-z_2}\,\mathrm{d}z_2+\frac{\beta_m q_m}{Q_{M_E+1}-q_m}\times\int_0^\infty\frac{e^{-z_2}}{Q_{M_E+1}z_2+1}\,\mathrm{d}z_2\right) \tag{70}$$

$$\approx \left(\frac{Q_{M_E+1}-q_m}{Q_{M_E+1}}\right)^{\beta_m}\left(1-\frac{\beta_m q_m}{(Q_{M_E+1}-q_m)\,Q_{M_E+1}}\times e^{\frac{1}{Q_{M_E+1}}}E_i\left(-\frac{1}{Q_{M_E+1}}\right)\right). \tag{71}$$

Finally, by inserting A_{D_1} and A_{D_2} into (64a), the E_{cs}^m for the m^{th} user is given by (22).

At high SNRs, for the m^{th} user with $m \ge M_E+1$, $\lim_{\rho\to\infty} E_{\text{cs}}^m$ is given by

$$\lim_{\rho\to\infty}E_{\text{cs}}^m = \frac{1}{\beta_m}\log_2\left(\mathbb{E}\left[\left(\frac{Q_m}{Q_{m+1}}\frac{Q_{M_E+1}-q_m}{Q_{M_E+1}}\right)^{\beta_m}\right]\right) = \log_2\left(\frac{Q_m}{Q_{m+1}}\frac{Q_{M_E+1}-q_m}{Q_{M_E+1}}\right). \tag{72}$$

APPENDIX E
PROOF FOR THEOREM 5

Recall that $\breve{E}_{\text{s}}^m$ can be expressed as

$$\breve{E}_{\text{s}}^m = \frac{1}{\beta_m}\log_2\left(B_{D_1}+B_{D_2}\right), \tag{73}$$

where

$$B_{D_1} = \psi_m \iint\limits_{D_1} \left(\frac{Q_m z_1 + 1}{Q_{m+1} z_1 + 1} \frac{Q_{m+1} z_2 + 1}{Q_m z_2 + 1}\right)^{\beta_m} \times \frac{e^{-(M-m+1)z_1 - z_2}}{\left(1 - e^{-z_1}\right)^{1-m}} \, \mathrm{d}z_1 \, \mathrm{d}z_2, \tag{74a}$$

$$B_{D_2} = \psi_m \iint\limits_{D_2} e^{-(M-m+1)z_1} \left(1 - e^{-z_1}\right)^{m-1} e^{-z_2} \, \mathrm{d}z_1 \, \mathrm{d}z_2. \tag{74b}$$

First, let us consider B_{D_2}. By replacing $\left(1 - e^{-z_1}\right)^{m-1}$ with $\sum_{s=0}^{m-1} \binom{m-1}{s} (-1)^s e^{-z_1 s}$, we get that

$$B_{D_2} = \psi_m \int_0^\infty e^{-z_2} \int_0^{z_2} \sum_{s=0}^{m-1} \binom{m-1}{s} (-1)^s e^{-a z_1} \, \mathrm{d}z_1 \, \mathrm{d}z_2 \tag{75a}$$

$$= \psi_m \sum_{s=0}^{m-1} \binom{m-1}{s} (-1)^s \left(-\frac{1}{a}\right) \int_0^\infty e^{-z_2} \left(e^{-a z_2} - 1\right) \, \mathrm{d}z_2 \tag{75b}$$

$$= \psi_m \sum_{s=0}^{m-1} \binom{m-1}{s} (-1)^s \frac{1}{a+1}. \tag{75c}$$

Then, we can consider B_{D_1}. Let us define $BB_{D_1}(z_2)$ and rewrite B_{D1} as follows:

$$B_{D_1} = \psi_m \int_0^\infty \left(\frac{Q_{m+1} z_2 + 1}{Q_m z_2 + 1}\right)^{\beta_m} e^{-z_2} BB_{D_1}(z_2) \, \mathrm{d}z_2, \tag{76a}$$

$$BB_{D_1}(z_2) = \int_{z_2}^\infty \left(\frac{Q_m z_1 + 1}{Q_{m+1} z_1 + 1}\right)^{\beta_m} e^{-(M-m+1)z_1} \times \left(1 - e^{-z_1}\right)^{m-1} \, \mathrm{d}z_1, \tag{76b}$$

where $BB_{D_1}(z_2)$ can be written as

$$BB_{D_1}(z_2) = \left(\frac{Q_{m+1}}{Q_m}\right)^{-\beta_m} \sum_{s=0}^{m-1} \binom{m-1}{s} (-1)^s \times \int_{z_2}^\infty \left(1 + \frac{\frac{q_m}{Q_{m+1}}}{Q_m z_1 + 1}\right)^{-\beta_m} e^{-a z_1} \, \mathrm{d}z_1. \tag{77}$$

By assuming that $q_m \leq Q_{m+1}$, where $m \neq \{M-1, M\}$, $\left(1 + \frac{\frac{q_m}{Q_{m+1}}}{Q_m z_1 + 1}\right)^{-\beta_m}$ can be approximated using the first two terms of generalized binomial expansion. Then, (77) can be approximated as

$$BB_{D_1}(z_2) \approx \left(\frac{Q_{m+1}}{Q_m}\right)^{-\beta_m} \sum_{s=0}^{m-1} \binom{m-1}{s} (-1)^s \times \left(\int_{z_2}^\infty e^{-a z_1} \, \mathrm{d}z_1 - \frac{\beta_m q_m}{Q_{m+1}} \int_{z_2}^\infty \frac{e^{-a z_1}}{Q_m z_1 + 1} \, \mathrm{d}z_1\right). \tag{78}$$

From (3.352.2) in [31], we have that

$$\int_u^\infty \frac{e^{-\mu x}}{x + \beta} \, \mathrm{d}x = -e^{\beta\mu} E_i(-\mu u - \mu\beta), \quad u \geq 0, |\arg(u + \beta)| < \pi, \mathrm{Re}\ \mu > 0. \tag{79}$$

Then, $BB_{D_1}(z_2)$ can be finally approximated as

$$BB_{D_1}(z_2) \approx \left(\frac{Q_{m+1}}{Q_m}\right)^{-\beta_m} \sum_{s=0}^{m-1} \binom{m-1}{s} (-1)^s \times \left(\frac{e^{-a z_2}}{a} + \frac{\beta_m q_m}{Q_m Q_{m+1}} e^{\frac{a}{Q_m}} E_i\left(-a z_2 - \frac{a}{Q_m}\right)\right). \tag{80}$$

By inserting $BB_{D_1}(z_2)$ back into B_{D_1}, we get (81), shown at the top of the next page. To simplify, (81) can be rewritten as

$$B_{D_1} = \psi_m \left(\frac{Q_{m+1}}{Q_m}\right)^{-\beta_m} \left(A_1 + \frac{\beta_m q_m}{Q_m Q_{m+1}} A_2\right). \tag{82}$$

Let us consider A_1 first.

$$A_1 = \left(\frac{Q_{m+1}}{Q_m}\right)^{\beta_m} \sum_{s=0}^{m-1} \binom{m-1}{s} \frac{(-1)^s}{a} \times \int_0^\infty \left(1 + \frac{\frac{q_m}{Q_{m+1}}}{Q_m z_2 + 1}\right)^{\beta_m} e^{-(a+1)z_2} \, \mathrm{d}z_2, \approx \left(\frac{Q_{m+1}}{Q_m}\right)^{\beta_m} \sum_{s=0}^{m-1} \binom{m-1}{s} \frac{(-1)^s}{a} \times \left(\int_0^\infty e^{-(a+1)z_2} \, \mathrm{d}z_2 + \frac{\beta_m q_m}{Q_{m+1}} \int_0^\infty \frac{e^{-(a+1)z_2}}{Q_m z_2 + 1} \, \mathrm{d}z_2\right), \tag{83}$$

which is approximated by applying the first two terms of the generalized binomial expansion. By applying (3.352.4) in [31], given in (67), we can get that

$$A_1 \approx \left(\frac{Q_{m+1}}{Q_m}\right)^{\beta_m} \sum_{s=0}^{m-1} \binom{m-1}{s} \frac{(-1)^s}{a} \times \left(\frac{1}{a+1} - \frac{\beta_m q_m}{Q_m Q_{m+1}} e^{\frac{a+1}{Q_m}} E_i\left(-\frac{a+1}{Q_m}\right)\right). \tag{84}$$

Now we can start to work on A_2. Recall that

$$A_2 = \left(\frac{Q_{m+1}}{Q_m}\right)^{\beta_m} \sum_{s=0}^{m-1} \binom{m-1}{s} (-1)^s e^{\frac{a}{Q_m}} \times \int_0^\infty \left(1 + \frac{\frac{q_m}{Q_{m+1}}}{Q_m z_2 + 1}\right)^{\beta_m} e^{-z_2} E_i\left(-a z_2 - \frac{a}{Q_m}\right) \, \mathrm{d}z_2. \tag{85}$$

in which $\left(1 + \frac{\frac{q_m}{Q_{m+1}}}{Q_m z_2 + 1}\right)^{\beta_m}$ can be approximated using the first two terms of the generalized binomial expansion. Then, A_2 can be transformed into

$$A_2 \approx \left(\frac{Q_{m+1}}{Q_m}\right)^{\beta_m} \sum_{s=0}^{m-1} \binom{m-1}{s} (-1)^s e^{\frac{a}{Q_m}} \times \left(\int_0^\infty e^{-z_2} E_i\left(-a z_2 - \frac{a}{Q_m}\right) \, \mathrm{d}z_2 + \frac{\beta_m q_m}{Q_{m+1}} \int_0^\infty \frac{1}{Q_m z_2 + 1} e^{-z_2} E_i\left(-a z_2 - \frac{a}{Q_m}\right) \, \mathrm{d}z_2\right). \tag{86}$$

$$B_{D_1}=\psi_m\left(\frac{Q_{m+1}}{Q_m}\right)^{-\beta_m}\left(\underbrace{\sum_{s=0}^{m-1}\binom{m-1}{s}\frac{(-1)^s}{a}\int_0^{\infty}\left(\frac{Q_{m+1}z_2+1}{Q_mz_2+1}\right)^{\beta_m}e^{-(a+1)z_2}\,\mathrm{d}z_2}_{A_1}+\frac{\beta_m q_m}{Q_mQ_{m+1}}\right.$$
$$\left.\times\underbrace{\sum_{s=0}^{m-1}\binom{m-1}{s}(-1)^s e^{\frac{a}{Q_m}}\int_0^{\infty}\left(\frac{Q_{m+1}z_2+1}{Q_mz_2+1}\right)^{\beta_m}e^{-z_2}E_i\left(-az_2-\frac{a}{Q_m}\right)\,\mathrm{d}z_2}_{A_2}\right), \tag{81}$$

By setting $y=az_2+\frac{a}{Q_m}$ and $\frac{a}{Q_m}\le y\le\infty$, A_2 can be rewritten as

$$A_2\approx\left(\frac{Q_{m+1}}{Q_m}\right)^{\beta_m}\sum_{s=0}^{m-1}\binom{m-1}{s}(-1)^s e^{\frac{a}{Q_m}}\times\left(-\frac{1}{a}e^{\frac{1}{Q_m}}\int_{\frac{a}{Q_m}}^{\infty}e^{-\frac{1}{a}y}E_1(y)\,\mathrm{d}y-\frac{\beta_m q_m}{Q_mQ_{m+1}}e^{\frac{1}{Q_m}}\int_{\frac{a}{Q_m}}^{\infty}e^{-\frac{1}{a}y}E_1(y)\frac{1}{y}\,\mathrm{d}y\right). \tag{87}$$

From (4.2.1) in [32], we have that

$$\int e^{-ux}E_1(vx)\,\mathrm{d}x=\frac{1}{u}\big(E_1((u+v)x)-e^{-ux}E_1(vx)\big). \tag{88}$$

By applying (88), we get that

$$\int_{\frac{a}{Q_m}}^{\infty}e^{-\frac{1}{a}y}E_1(y)\,\mathrm{d}y=-a\left(E_1\left(\frac{a+1}{Q_m}\right)-e^{-\frac{1}{Q_m}}E_1\left(\frac{a}{Q_m}\right)\right). \tag{89}$$

Further, from (4.2.29) in [32], we have that

$$\int_c^{\infty}e^{-ux}E_1(vx)\frac{1}{x}\,\mathrm{d}x=\left(r+\ln(uc)+E_1(uc)\right)E_1(vc)+\frac{1}{2}\big(\zeta(2)+(r+\ln(vc))^2\big)+e^{-vc}\sum_{\delta=0}^{\infty}\frac{e_\delta(vc)}{(\delta+1)^2}\left(-\frac{u}{v}\right)^{\delta+1}+\sum_{\delta=1}^{\infty}\frac{(-vc)^\delta}{\delta!\delta^2}. \tag{90}$$

By applying (90), we get that

$$\int_{\frac{a}{Q_m}}^{\infty}e^{-\frac{1}{a}y}E_1(y)\frac{1}{y}\,\mathrm{d}y=\left(r-\ln(Q_m)+E_1\left(\frac{1}{Q_m}\right)\right)\times E_1\left(\frac{a}{Q_m}\right)+\frac{1}{2}\left(\zeta(2)+\left(r+\ln\left(\frac{a}{Q_m}\right)\right)^2\right)+e^{-\frac{a}{Q_m}}\sum_{\delta=0}^{\infty}\frac{e_\delta\left(\frac{a}{Q_m}\right)}{(\delta+1)^2}\left(-\frac{1}{a}\right)^{\delta+1}+\sum_{\delta=1}^{\infty}\frac{\left(-\frac{a}{Q_m}\right)^\delta}{\delta!\delta^2}. \tag{91}$$

For simplicity, we define two notations, i.e., $\Psi_1=\sum_{\delta=0}^{\infty}\frac{e_\delta\left(\frac{a}{Q_m}\right)}{(\delta+1)^2}\left(-\frac{1}{a}\right)^{\delta+1}$ and $\Psi_2=\sum_{\delta=1}^{\infty}\frac{\left(-\frac{a}{Q_m}\right)^\delta}{\delta!\delta^2}$. In the following, we will show that both infinite summations, i.e., Ψ_1 and Ψ_2, can be calculated easily. Let us rewrite Ψ_1 as $\lim_{\Delta\to\infty}\sum_{\delta=0}^{\Delta}\frac{e_\delta\left(\frac{a}{Q_m}\right)}{(\delta+1)^2}\left(-\frac{1}{a}\right)^{\delta+1}$. One can easily show that Ψ_1 can be approximated using a finite summation, which converges for any values as long as $\Delta\ge 50$. Furthermore, by applying the definition of generalized hypergeometric function [26], we can replace Ψ_2 with $-\frac{a}{Q_m}{}_3\mathbf{F}_3\left[\begin{smallmatrix}1,\,1,\,1\\2,\,2,\,2\end{smallmatrix};-\frac{a}{Q_m}\right]$. Henceforth, A_2 can be finally expressed as

$$A_2\approx\left(\frac{Q_{m+1}}{Q_m}\right)^{\beta_m}\sum_{s=0}^{m-1}\binom{m-1}{s}(-1)^s e^{\frac{a}{Q_m}}\left(e^{\frac{1}{Q_m}}\times\left(E_1\left(\frac{a+1}{Q_m}\right)-e^{-\frac{1}{Q_m}}E_1\left(\frac{a}{Q_m}\right)\right)-\frac{\beta_m q_m}{Q_mQ_{m+1}}e^{\frac{1}{Q_m}}\times\left(\left(r-\ln(Q_m)+E_1\left(\frac{1}{Q_m}\right)\right)E_1\left(\frac{a}{Q_m}\right)+\frac{1}{2}\left(\zeta(2)+\left(r+\ln\left(\frac{a}{Q_m}\right)\right)^2\right)+e^{-\frac{a}{Q_m}}\sum_{\delta=0}^{\infty}\frac{e_\delta\left(\frac{a}{Q_m}\right)}{(\delta+1)^2}\times\left(-\frac{1}{a}\right)^{\delta+1}-\frac{a}{Q_m}{}_3\mathbf{F}_3\left[\begin{smallmatrix}1,\,1,\,1\\2,\,2,\,2\end{smallmatrix};-\frac{a}{Q_m}\right]\right)\right). \tag{92}$$

By inserting A_1, A_2, and B_{D_2} into (73), $\breve{E}_{\mathrm{s}}^m$ can be finally given in (28)-(30).

Then, we can start to derive the closed-form expression for $\mathbb{E}[\breve{R}_s^m]$, given in (93), shown at the top of the next page. Since B_1-B_4 have similar structures, here we only show the steps of deriving B_1 for simplicity.

$$B_1=\psi_m\int_0^{\infty}\log_2\left(Q_mz_1+1\right)e^{-(M-m+1)z_1}\left(1-e^{-z_1}\right)^{m-1}\times\int_0^{z_1}e^{-z_2}\,\mathrm{d}z_2\,\mathrm{d}z_1 \tag{94a}$$
$$=\psi_m\int_0^{\infty}\log_2(Q_mz_1+1)\,e^{-(M-m+1)z_1}\big(1-e^{-z_1}\big)^m\,\mathrm{d}z_1 \tag{94b}$$
$$=\psi_m\sum_{s=0}^{m}\binom{m}{s}(-1)^s\int_0^{\infty}\log_2\left(Q_mz_1+1\right)e^{-az_1}\,\mathrm{d}z_1. \tag{94c}$$

$$\begin{aligned}\mathbb{E}[\breve{R}_s^m] &= \iint\limits_{D_1}\left(\log_2\left(1+\frac{q_m z_1}{Q_{m+1}z_1+1}\right)-\log_2\left(1+\frac{q_m z_2}{Q_{m+1}z_2+1}\right)\right) f(z_1,z_2)\,\mathrm{d}z_1\,\mathrm{d}z_2\\ &= \underbrace{\iint\limits_{D_1}\log_2\left(Q_m z_1+1\right) f(z_1,z_2)\,\mathrm{d}z_1\,\mathrm{d}z_2}_{B_1} - \underbrace{\iint\limits_{D_1}\log_2(Q_{m+1}z_1+1) f(z_1,z_2)\,\mathrm{d}z_1\,\mathrm{d}z_2}_{B_2}\\ &\quad - \underbrace{\iint\limits_{D_1}\log_2(Q_m z_2+1) f(z_1,z_2)\,\mathrm{d}z_1\,\mathrm{d}z_2}_{B_3} + \underbrace{\iint\limits_{D_1}\log_2(Q_{m+1}z_2+1) f(z_1,z_2)\,\mathrm{d}z_1\,\mathrm{d}z_2}_{B_4}. \end{aligned} \tag{93}$$

By defining $Q_m z_1 = x$, B_1 can be transformed to

$$B_1 = \psi_m \frac{1}{Q_m}\sum_{s=0}^{m}\binom{m}{s}(-1)^s\int_0^\infty \log_2\left(x+1\right)e^{-\frac{a}{Q_m}x}\,\mathrm{d}x. \tag{95}$$

From (4.337.2) in [31], we have that

$$\int_0^\infty e^{-ux}\ln(1+vx)\,\mathrm{d}x = -\frac{1}{u}e^{\frac{u}{v}}E_i\left(-\frac{u}{v}\right), \quad |\arg\, v|<\pi, \mathrm{Re}\; u>0. \tag{96}$$

By applying (96), B_1 can be finally written as

$$B_1 = -\frac{\psi_m}{\ln 2}\sum_{s=0}^{m}\binom{m}{s}(-1)^s\frac{1}{a}e^{\frac{a}{Q_m}}E_i(-\frac{a}{Q_m}). \tag{97}$$

By following similar methods, B_2-B_4 can also be expressed in closed-form and finally, $\mathbb{E}[\breve{R}_s^m]$ is given in (31).

APPENDIX F
PROOF FOR LEMMA 1

Note that the lower bound on the ergodic secrecy rate, i.e., $\mathbb{E}[\breve{R}_s^m]$, is given by

$$\mathbb{E}[\breve{R}_s^m] = \mathbb{E}\left[\log_2\left(1+\frac{q_m|h_m|^2}{Q_{m+1}|h_m|^2+1}\right) - \log_2\left(1+\frac{q_m|h_e|^2}{Q_{m+1}|h_e|^2+1}\right)\right]. \tag{98}$$

By inserting $\rho\to 0$ (which means $q_m=0$ and $Q_{m+1}=0$) into (98), one can get that $\mathbb{E}[\breve{R}_s^m]=0$. Also, by inserting $\rho\to 0$ into (26), we can get that $\lim_{\rho\to 0}\breve{E}_s^m = 0$.

When $\rho\to\infty$, $\mathbb{E}[\breve{R}_s^m]$ can be approximated as

$$\mathbb{E}[\breve{R}_s^m] = \mathbb{E}\left[\log_2\left(1+\frac{q_m|h_m|^2}{Q_{m+1}|h_m|^2}\right) - \log_2\left(1+\frac{q_m|h_e|^2}{Q_{m+1}|h_e|^2}\right)\right], \tag{99}$$

which equals to 0. Also, by inserting $\rho\to\infty$ into (26), we can get that $\lim_{\rho\to\infty}\breve{E}_s^m = 0$.

APPENDIX G
PROOF FOR LEMMA 2

By inserting $\rho\to 0$ into $\hat{R}_s^m$, one can easily get that $\hat{R}_s^m = 0$ and $\mathbb{E}[\hat{R}_s^m]=0$. Then, by inserting $\lim_{\rho\to 0}\hat{R}_s^m = 0$ into (6), it is clear that $\lim_{\rho\to 0}\hat{E}_s^m = 0$.

On the other hand, when $\rho\to\infty$, $\lim_{\rho\to\infty}\hat{R}_s^m$ can be written as

$$\lim_{\rho\to\infty}\hat{R}_s^m = \left[\log_2\left(1+\frac{q_m}{Q_{m+1}}\right)-\log_2\left(1+\frac{q_m}{Q_1}\right)\right]^+, \tag{100}$$

which is a positive value. Then, we can get that $\lim_{\rho\to\infty}\mathbb{E}[\hat{R}_s^m] = \log_2\left(\frac{Q_m}{Q_{m+1}}\frac{Q_1}{Q_1+q_m}\right)$. By inserting (100) into (6), we can notice that $\lim_{\rho\to\infty}\hat{E}_s^m = \lim_{\rho\to\infty}\mathbb{E}[\hat{R}_s^m]$, which completes the proof.

REFERENCES

[1] Z. Ding *et al.*, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[2] Z. Ding, M. Peng, and H. V. Poor, "Cooperative non-orthogonal multiple access in 5G systems," *IEEE Commun. Lett.*, vol. 19, no. 8, pp. 1462–1465, Aug. 2015.

[3] F. Fang, H. Zhang, J. Cheng, and V. C. M. Leung, "Energy efficiency of resource scheduling for non-orthogonal multiple access (NOMA) wireless network," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–5.

[4] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo, "Non-orthogonal multiple access for 5G and beyond," *Proc. IEEE*, vol. 105, no. 12, pp. 2347–2381, Dec. 2017.

[5] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 721–742, 2nd Quart., 2017.

[6] W. Yu, L. Musavian, and Q. Ni, "Link-layer capacity of NOMA under statistical delay QoS guarantees," *IEEE Trans. Commun.*, vol. 66, no. 10, pp. 4907–4922, Oct. 2018.

[7] Y. Liu, Z. Qin, M. Elkashlan, A. Nallanathan, and J. A. McCann, "Non-orthogonal multiple access in large-scale heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2667–2680, Dec. 2017.

[8] S. M. R. Islam, M. Zeng, O. A. Dobre, and K.-S. Kwak, "Resource allocation for downlink NOMA systems: Key techniques and open issues," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 40–47, Apr. 2018.

[9] H. V. Poor and R. F. Schaefer, "Wireless physical layer security," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 1, pp. 19–26, Jan. 2017.

[10] A. Chorti, S. M. Perlaza, Z. Han, and H. V. Poor, "On the resilience of wireless multiuser networks to passive and active eavesdroppers," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 9, pp. 1850–1863, Sep. 2013.

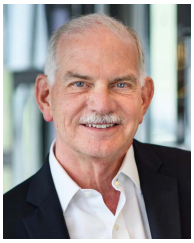
[11] J. Barros and M. R. D. Rodrigues, "Secrecy capacity of wireless channels," in *Proc. IEEE Int. Symp. Inf. Theory*, Seattle, WA, USA, Jul. 2006, pp. 356–360.

[12] C. E. Shannon, "Communication theory of secrecy systems," *Bell Labs Tech. J.*, vol. 28, no. 4, pp. 656–715, Oct. 1949.
[13] S. Leung-Yan-Cheong and M. Hellman, "The Gaussian wiretap channel," *IEEE Trans. Inf. Theory*, vol. IT-24, no. 4, pp. 451–456, Jul. 1978.
[14] Z. Qin, Y. Liu, Z. Ding, Y. Gao, and M. Elkashlan, "Physical layer security for 5G non-orthogonal multiple access in large-scale networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.
[15] Y. Liu, Z. Qin, M. Elkashlan, Y. Gao, and L. Hanzo, "Enhancing the physical layer security of non-orthogonal multiple access in large-scale networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1656–1672, Mar. 2017.
[16] Y. Zhang, H.-M. Wang, T.-X. Zheng, and Q. Yang, "Energy-efficient transmission design in non-orthogonal multiple access," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2852–2857, Mar. 2017.
[17] L. Lv, J. Chen, Q. Ni, and Z. Ding, "Design of cooperative non-orthogonal multicast cognitive multiple access for 5G systems: User scheduling and performance analysis," *IEEE Trans. Commun.*, vol. 65, no. 6, pp. 2641–2656, Jun. 2017.
[18] GSMA Intelligence. (Dec. 2014). *Understanding 5G: Perspectives on Future Technological Advancements in Mobile*. [Online]. Available: https://gsmaintelligence.com/research/2014/12/understanding-5g/451/
[19] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality-of-service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, Jul. 2003.
[20] Z. Ding *et al.*, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.
[21] T. Q. Duong, X. Zhou, and H. V. Poor, *Trusted Communications with Physical Layer Security for 5G and Beyond*. Edison, NJ, USA: IET, Nov. 2017.
[22] Y. Liu, Z. Ding, M. Elkashlan, and H. V. Poor, "Cooperative non-orthogonal multiple access with simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 938–953, Apr. 2016.
[23] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.
[24] M. Ozmen and M. C. Gursoy, "Secure transmission of delay-sensitive data over wireless fading channels," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 9, pp. 2036–2051, Sep. 2017.
[25] H. A. David and H. N. Nagaraja, *Order Statistics*, 3rd ed. New York, NY, USA: Wiley, 2003.
[26] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. New York, NY, USA: Dover, 1965.
[27] *Evolved Universal Terrestrial Radio Access (E-UTRA); LTE Physical Layer; General Description, Release 8*, document TS 36.201, 3GPP, 2009.
[28] W. Wang, K. C. Teh, and K. Li, "Artificial noise aided physical layer security in multi-antenna small-cell networks," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 6, pp. 1470–1482, Jun. 2017.
[29] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[30] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation over wireless links," *IEEE Trans. Wireless Commun.*, vol. 6, no. 8, pp. 3058–3068, Aug. 2007.
[31] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 6th ed. New York, NY, USA: Academic, 2000.
[32] M. Geller and E. W. Ng, "A table of integrals of the exponential integral," *J. Res. Nat. Bureau Standards*, vol. 73B, no. 3, pp. 191–210, Sep. 1969.

Wenjuan Yu received the Ph.D. degree in communication systems from the School of Computing and Communications, Lancaster University, Lancaster, U.K., in 2018. She is currently a Research Fellow with the 5G Innovation Centre, Institute for Communication Systems, University of Surrey, Guildford, U.K. Her research interests include radio resource management, uRLLC, 5G and beyond wireless networks. She is an Executive Editor of the *Transactions on Emerging Telecommunications Technologies*.

Arsenia Chorti (S'00–M'05) received the Ph.D. degree in electrical and electronic engineering from Imperial College London, U.K., in 2005. She undertook post-doctoral positions at the Universities of Southampton, U.K.; TCU, Greece; and UCL, U.K. She was a Marie Curie IOF with Princeton University, USA and ICS-FORTH, Greece. She has served as a Senior Lecturer of communications and networks with Middlesex University and as a Lecturer with the University of Essex, U.K. Since September 2017, she has been an Associate Professor with ETIS, UMR 8051, Université Paris Seine, Université Cergy-Pontoise, ENSEA, CNRS, France. Her research interests include physical layer security, resource allocation for B5G, and stochastic signal processing, wireless communications, and information theory in general. She has been a member of the IEEE Teaching Awards Committee since 2017.

Leila Musavian (S'05–M'07) received the Ph.D. degree in telecommunications from Kings College London, U.K. She was a Post-Doctoral Fellow with INRS-EMT, Canada, from 2006 to 2008, a Research Associate with McGill University, from 2011 to 2012, and a Lecturer with InfoLab21, Lancaster University, from 2012 to 2016. She is currently a Reader with the School of Computer Science and Electronic Engineering, University of Essex. Her research interests lie in 5G/B5G, uRLLC, radio resource management for next generation wireless networks, and energy harvesting communication systems. She is an Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and an Associate Editor of Wiley's *Internet Technology Letters*. She has served as an Executive Editor for the *Transactions on Emerging Telecommunications Technologies* from 2016 to 2019.



H. Vincent Poor (S'72–M'77–SM'82–F'87) received the Ph.D. degree in EECS from Princeton University in 1977. From 1977 until 1990, he was on the faculty of the University of Illinois at Urbana–Champaign. Since 1990, he has been on the faculty at Princeton, where he is currently the Michael Henry Strater University Professor of Electrical Engineering. From 2006 to 2016, he served as the Dean of Princeton's School of Engineering and Applied Science. He has also held visiting appointments at several other universities, including most recently at Berkeley and Cambridge. His research interests are in the areas of information theory and signal processing, and their applications in wireless networks, energy systems, and related fields. Among his publications in these areas is the recent book *Multiple Access Techniques for 5G Wireless Networks and Beyond* (Springer, 2019).

Dr. Poor is a member of the National Academy of Engineering and the National Academy of Sciences, and is a Foreign Member of the Chinese Academy of Sciences, the Royal Society, and other national and international academies. Recent recognition of his work includes the 2017 IEEE Alexander Graham Bell Medal, the 2019 ASEE Benjamin Garver Lamme Award, a D.Sc. *honoris causa* from Syracuse University in 2017, and a D.Eng. *honoris causa* from the University of Waterloo in 2019.

Qiang Ni (M'04–SM'08) received the B.Sc., M.Sc., and Ph.D. degrees from the Huazhong University of Science and Technology, China, all in engineering. He is currently a Professor and the Head of the Communication Systems Group, School of Computing and Communications, Lancaster University, Lancaster, U.K. His research interests include the area of future generation communications and networking, including green communications and networking, millimeter-wave wireless communications, cognitive radio network systems, non-orthogonal multiple access (NOMA), heterogeneous networks, 5G and 6G, SDN, cloud networks, energy harvesting, wireless information and power transfer, the IoTs, cyber physical systems, machine learning, big data analytics, and vehicular networks. He has authored or coauthored more than 200 articles in these areas. He was an IEEE 802.11 Wireless Standard Working Group Voting Member and a contributor to the IEEE Wireless Standards.

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 6, 2019, accepted January 15, 2020, date of publication February 10, 2020, date of current version February 18, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2972640

Real-Time Algorithms for the Detection of Changes in the Variance of Video Content Popularity

SOTIRIS SKAPERAS[1], (Student Member, IEEE), LEFTERIS MAMATAS[1], (Member, IEEE), AND ARSENIA CHORTI[2], (Member, IEEE)

[1]Department of Applied Informatics, University of Macedonia, 546 36 Thessaloniki, Greece
[2]ETIS/Université Paris Seine, Université Cergy-Pointoise, ENSEA, CNRS, 95000 Cergy, France

Corresponding author: Sotiris Skaperas (sotskap@uom.edu.gr)

This work was supported in part by the EU's Horizon 2020 Research and Innovation Programme through the 4th open call scheme of the FED4FIRE+ under Grant 732638, in part by the EU-BRA Horizon 2020 NECOS Project under Grant 777067, in part by the Initiative d'Excellence Paris//Seine project eNiGMA, and in part by the ELIOT project under Grant ANR-18-CE40-0030/FAPESP 2018/12579-7.

ABSTRACT As video content is responsible for more than 70% of the global IP traffic, related resource allocation approaches, e.g., using content caching, become increasingly important. In this context, to avoid under-provisioning, it is important to rapidly detect and respond to changes in content popularity dynamics, including volatility, i.e., changes in the second order moment of the underlying process. In this paper, we focus on the early identification of changes in the variance of video content popularity, which we address as a statistical change point (CP) detection problem. Unlike changes in the mean that can be well captured by non-parametric statistical approaches, to address this more demanding problem, we construct a hypothesis test that uses in the test statistic both parametric and non-parametric approaches. In the context of parametric models, we consider linear, in the form of autoregressive moving average (ARMA), and, nonlinear, in the form of generalized autoregressive conditional heteroskedasticity (GARCH) processes. We propose an integrated algorithm that combines off-line and on-line CP schemes, with the off-line scheme used as a training (learning) phase. The algorithm is first assessed over synthetic data; our analysis demonstrates that non parametric and GARCH model based approaches can better generalize and are better suited for content views time series with unknown statistics. Finally, the non-parametric and the GARCH based variations of our proposed integrated algorithm are applied on real YouTube video content views time series, to illustrate the performance of the proposed approach of volatility change detection.

INDEX TERMS Content popularity dynamics detection, change point analysis, variance change detection, volatility detection.

I. INTRODUCTION

Understanding the popularity characteristics of online content and predicting the future popularity of individual videos are of great importance. They have direct implications in various contexts [1], such as service design, advertisement planning, network management [2], and so on. As an example, an efficient content caching scheme should be popularity-driven [3], meaning that it should incorporate the future popularity of content into the caching decision making. In this framework, novel cache replacement methods that are ''popularity-driven'' have recently appeared, e.g., the algorithms proposed in [4], based on learning the popularity of content and using it to determine which content should be retained and which should be evicted from the cache. Other important applications include content delivery networks (CDNs) in which ''analytics-as-a-service'' approaches are employed and information centric networks (ICNs) with emphasis on the Internet of things (IoT) [5].

Higher order moments of the underlying random process are unarguably important for the efficient statistical characterization of content popularity; in particular, ''volatility'' plays a central role in capturing the underlying dynamics of content views. As an example, in caching applications, it has been established in [6] that a major factor greatly impacting

The associate editor coordinating the review of this manuscript and approving it for publication was Jose Saldana.

efficiency is related to demand volatility; this reflects the fact that files might not be constantly requested following a stationary model, but rather, only be requested once or twice and subsequently exhibit vanishing demand in time (e.g., volatility in YouTube content). Based on these findings, an efficient strategy for resource provisioning should in principle consider not only conditional mean demands but also demand fluctuations, thus avoiding under-provisioning or over-provisioning.

To analyze the underlying statistics of content views data, the latter are typically represented as a time series. Time series data are sequences of measurements over time, describing the behavior of systems. The behavior can change over time due to external events and/or internal systematic changes in dynamics/distribution. Success in revealing such patterns can be translated to the ability to respond rapidly to these changes. In this direction, there has recently been a surge of research in the area of content popularity prediction using artificial intelligence (AI) [7]. In this context, machine learning based methods (e.g., deep learning) need effective feature mining and a huge mass of labeled examples to provide successful performance [8], [9]. In applications in which *real time* content popularity monitoring is required this might become a challenge. As an example, in [10] the authors propose an *off-line* deep learning approach to detect popularity that is subsequently integrated into the on-line caching policy in fog radio applications; however, whenever there is an important change in the underlying dynamics of content popularity, it follows that a new off-line training might be required to run the algorithm properly.

In this work, we alternatively turn our attention to lightweight *statistical* procedures that fall in the general context of AI (instead of deep learning specifically), in order to operate in an on-line manner (real-time) and to keep the size of the required set of historical data as small as possible. Our proposed algorithm is autonomous, in the sense that all its parameters are determined without manual intervention during a training period; furthermore, the training period is limited to only a few hundred data points (instead of thousands or millions as is typical in deep learning).

Importantly, instead of attempting to *predict* the evolution of content popularity, in this work we rather focus on *detecting* changes in its underlying statistics, and doing so in real-time. To this end, we propose the use of on-line change point (CP) analysis; to complement our work [11], [12] that focused on the identification of changes in the mean of a time series, here, we alternatively investigate the performance of corresponding on-line algorithms to identify changes in the variance of a time series using CP analysis.

In general, CP methods are either off-line or on-line. Off-line algorithms operate retrospectively and identify CPs in a historical dataset, a thorough study can be found in [13]. On-line algorithms [14] monitor in real time a data sequence and aim to detect CPs as soon as they occur. In this work, we propose an efficient combination of an off-line and various on-line procedures for the detection of changes in the second order statistics of video content popularity, as soon as they occur (real-time). The proposed detector is built upon our earlier proposal for a real-time CP detector of mean changes in data series, that we applied to monitor the average number of video content [11], [12]. Albeit, the monitoring of changes in the variance of a time series is much more challenging.

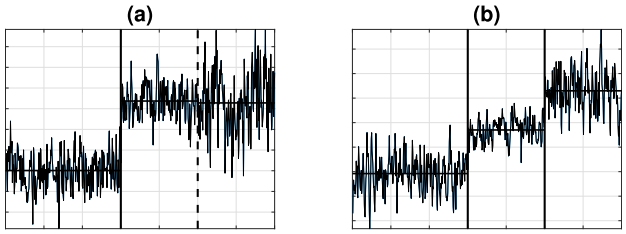


FIGURE 1. Simulated time series with CPs in the mean (solid line) and the variance (dashed line) for (a) separated and (b) simultaneous changes in the mean/variance. Horizontal lines illustrate the mean value.

To further illustrate our motivation behind this work, we note that an overall approach considering both mean and variance changes allows for a more efficient handling of content popularity changes as highlighted in Fig. 1. For example, Fig. 1(a) depicts that a crucial popularity change may affect only the variance parameter, in the specific example at the third segment of the time series. On the other hand, Fig. 1(b), depicts that in the case of a simultaneous change in the mean and the variance, e.g., in the second segment of the time series, the latter is critical to estimate the actual impact of this change. Monitoring the variance may also be used as a measure of uncertainty, determining the degree of fluctuation of popularity around its expectation; for instance, compare the behaviour of the time series in Fig. 1(b) after the first and the second CP (second and third segments of the data series, respectively).

To identify changes in the variance, a more elaborate test statistic is employed in the present study. With respect to [11], [12], we further introduce novel on-line tracking mechanisms based on autoregressive moving average (ARMA) and generalized autoregressive conditional heteroskedasticity (GARCH) models. The most important novel aspects of this paper are listed below:

- We show that variance CP detection is important in the context of content popularity.
- We introduce a relevant on-line detection algorithm, enhanced by the following two mechanisms: (a) an offline CP detection over training data for the estimation of the on-line test parameters; and (b) identification of the change magnitude in the pro- and post-change variance structure.
- Our algorithm supports three alternative on-line tests for content popularity detection – based on ARMA and GARCH models as well as a non-parametric approach – covering a wide-range of time series characteristics.
- We performed experiments both on synthetic and real time series datasets. Our results show that:

(i) the GARCH and the non-parametric approaches perform better when the time series does not follow a linear model; (ii) overall, these approaches can generalize better with respect to the true alarm rates; and (iii) the non parametric approach can identify CPs more rapidly.

In future work we intend to expand the algorithm to include additional dimensions that can be volatility indicators, such as the number of likes, viewer comments, content size, as well as network parameters such as the utilization of servers, in order to enhance the agility of the volatility estimation of the so called ''content workload'' as a whole. We will also investigate the algorithm's scalability properties, theoretically and experimentally, i.e., identify the number of videos that can be analyzed in parallel.

The rest of this contribution is structured as follows: In Section II, background concepts and high level properties of the proposed integrated algorithm are discussed. In Section III, the offline training is presented in detail, while Section IV presents three different approaches for the construction of the online test statistic. The integrated algorithms are assessed on synthetic data in Section V and applied to real YouTube content view data in Section VI. Conclusions and discussion on future enhancements are included in Section VII.

II. BACKGROUND CONCEPTS AND INTEGRATED ALGORITHM

A. CHANGE POINT ANALYSIS

Change point (CP) detection refers to the problem of identifying data structures that do not correspond to the anticipated ''normal'' behavior. We note that, to the best of our knowledge, this is the first work in the literature proposing an automated mechanism for the detection of volatility changes in a time series in the context of content popularity detection.

The theory of CP analysis is typically pertinent to anomaly detection. In the domain of networking in particular, the theory of CP detection has played an instrumental role in the modelling of network traffic monitoring represented through time series [15] and network anomaly/intrusion detection [16]; for a comprehensive review the interested reader may refer to [17]. In this framework, CP detection techniques [18] are used for the identification of: (i) point anomalies and outliers, i.e., data points deviating distinctively from the bulk of collected data; (ii) pattern anomalies, i.e., groups of data points that are collectively anomalous with respect to historical data; and, (iii) CP anomalies due to changes in the time series's statistical structure (in the mean/variance and in general in the underlying distribution). In this work, we focus on the detection of CP anomalies and consider the other two categories as disturbances. The reasoning behind this choice is that, on one hand, a resource allocation scheduler should be insensitive to instantaneous/very short-term changes in resource demand (e.g., represented as outliers in the content demand), but, on the other hand, should be highly responsive to changes in the underlying statistics of the demand.

B. PARAMETRIC AND NON PARAMETRIC CP DETECTION ALGORITHMS

Statistical based approaches are categorized as parametric [19] and non-parametric [14]. Non-parametric methods do not make use of a particular time series model fit and apply directly the observed data to the monitoring procedures. In this context, CUSUM based methods are non-parametric by design. For example, the authors in [20] provide a CUSUM stopping rule with application in computer vision problems. A CUSUM approach for CP detection on observations with an unknown distribution before and after a change, has been recently developed in [21]. Furthermore, an algorithm based on the Shiryaev-Roberts procedure was proposed in [22], to detect anomalies in computer network traffic.

On the other hand, parametric methods utilize as inputs values obtained from a specific model that has been fit to the original data (instead of using the original data set directly). As an example, Kalman filtering is combined with several CP methods in [23]. In [24], traffic flows are modeled using Markov chains and an anomaly detection mechanism based on the generalized likelihood ratio test (LRT) algorithm. Further examples assuming specific distribution for the data include [25], in which a bivariate sequential generalized LRT algorithm was proposed, assuming that the packet rate and the packet size follow a Poisson and a normal distribution, respectively. Other, non residual methods, include estimates' detectors based on the differences between the estimated model parameters (see [13], [26]), or based on the quasi-likelihood scores estimators of the parameters of a GARCH process [27].

C. VIDEO CONTENT POPULARITY PREDICTION VS DETECTION

The prediction of video content popularity characteristics and dynamics [28], as well as models to predict popularity evolution, e.g., [29] and [30], is a well studied topic in the literature. Among others, in [31], the authors perform a detailed analysis to characterize the YouTube traffic within a campus network and conclude that in this scenario the content popularity can be well approximated by the Zipf distribution. A comprehensive survey on video traffic models can be found in [32]. Overall, several methods have been proposed in this context, including time series models, regression models [33]–[35] and machine learning (deep neural networks) techniques [36], [37].

Focusing on time series modelling in particular, linear, non linear and hybrid models have invariably been proposed. In early works, linear time series models have been used, e.g., the authors in [38] introduce an ARMA(7, 7) model to describe and predict the daily views of individual videos. Alternatively, in [39], by taking into consideration seasonality, an autoregressive integrated moving average (ARIMA) model is used to forecast the popularity of online content. Other approaches include fractional ARIMA (FARIMA)

models, that capture both short-range dependence (SRD) and long-range dependence (LRD) statistical properties [40].

Recently, non linear models have further been proposed to take into account the conditional heteroskedasticity and the conditional volatility of the data series (seen as a stochastic process). In these cases, GARCH models are involved. For example, in the comparative study [41], the authors showed that a hybrid ARIMA/GARCH model was superior to FARIMA and wavelet neural network models, while in [42], a similar hybrid FARIMA/GARCH approach was also introduced. In essence, the existing hybrid models consider the second order characteristics of a time series as a supplementary element to further improve the forecasting or estimation of the content popularity. More precisely, these solutions assume conditional heteroskedasticity for the errors of the ARMA or FARIMA model. An exception can be found in [43], where a video demand predictor forecasts the volatility and correlation of the streaming traffic associated with different videos, based on multivariate GARCH models.

On the other hand, the problem of detecting (i.e., estimating), non-parametrically and in real time, CPs on content popularity sequences, has not been adequately investigated yet. Among of only a handful of related studies, in our previous works [11], [12], [44] we proposed and implemented a real-time, non-parametric and low-complexity video content popularity CP detector (as opposed to predictor) for changes in the mean value of video content popularity. In the present contribution, in contrast to [11], [12], we introduce an innovative online algorithm for the detection of CPs in the second order statistics of content popularity data. We also present an enlarged statistical framework, that includes parametric as well as non-parametric detectors.

Our algorithm can be used as a "stand alone" mechanism, but may also be a helpful complementary tool for prediction approaches. With respect to the latter, it can be employed in validating whether assumptions made by a prediction model are still reasonably satisfied, or, whether the prediction model/procedure needs adjustment. Since, data are often influenced by a multitude of external factors, stationarity assumptions cannot be guaranteed over the whole monitoring period, especially for long time ranges.

D. OVERVIEW OF THE PROPOSED INTEGRATED ALGORITHM

We summarize in Fig. 2 the overall algorithm as a flow diagram that links an off-line (training) and an on-line phase, as well as their individual components. Without loss of generality we assume an arbitrary time instance m_s as the starting point of a monitoring period. Then, the off-line analysis is applied to the historical (training) data until $t = m_s$, resulting in the division of the data sequence in stable subsequences. The last subsequence is the training sample representing the initial sample of the on-line phase. During the training stage, if a parametric approach is chosen, we estimate the model parameters (e.g., ARMA or GARCH) and any other necessary statistical characteristics that describe the last stable subsequence's (time series) behavior. We note that without having first obtained a statistically robust division of the training sample into stable subsequences, the estimation of a model's parameters could be seriously impacted.

Next, an on-line detector is implemented for a monitoring period $t = m_{s+1}, \ldots, m_{s+l}$. If a CP is detected at cp^*_{on}, the CP magnitude on the data structure is evaluated. The new starting point for the subsequent monitoring window is then set to $m'_s = cp^*_{on} + d$, where d is a constant specifying a period assuming no change. Alternatively, if no change is detected after l instances, the procedure restarts automatically from the time point $m'_s = m_{s+l}$. The reasons behind this choice are twofold. First, to keep the algorithm running over a window of size at most l, in order to keep the computational complexity low (lightweight), as opposed to allowing increasing window sizes. Second, to facilitate the fast responsiveness of the algorithm, as will be demonstrated through numerical examples in Section V.

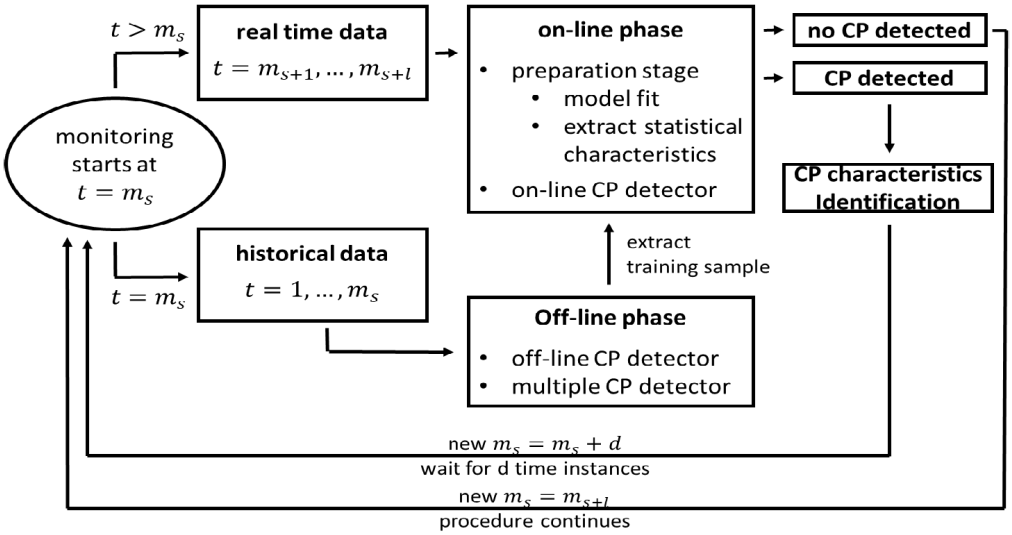


FIGURE 2. Flow diagram of the real-time variance CP detector for content views data.

III. OFF-LINE PHASE

In this Section, the training phase of the algorithm is discussed and the fundamental components of the off-line scheme are presented. We choose a retrospective CP scheme to ascertain that the on-line phase is indeed carried out on homogeneous data. We note that standard off-line CP schemes can only detect a single CP. To address the issue of detection of multiple CPs, we modify the basic scheme with a novel time series segmentation heuristic, that belongs to the family of binary segmentation algorithms, similarly to [11], [12].

Let $\{X_n : n \in \mathbb{N}\}$ be a time series representing the content views, for a specific video. Since we are interested only in the variance fluctuation of the underlying random value (r.v.), we assume a constant, over time, expected value $\mathrm{E}(X_i)$, where $\mathrm{E}(\cdot)$ denotes expectation. The stability of the mean value can be ensured by a data transformation, such as taking the first differences , $\Delta_n = X_n - X_{n-1}$, thus rendering $\mathrm{E}(X_i) = 0$.

Considering the training phase, we have to check if the variance structure remains stable over the whole training period N. Consequently we study the null hypothesis,

$$H_0 : \sigma_1^2 = \cdots = \sigma_N^2, \tag{1}$$

where $\sigma_n^2 = \mathrm{Var}(X_n) = \mathrm{E}(X_n^2)$, given that we have modified the time series so that $\mathrm{E}(X_n) = 0$. The (general) alternative hypothesis is designed to allow the existence of multiple changes $l_i \in \{1, \ldots, N\}$, $i = 1, \ldots, r$, where r is the multitude of changes,

$$\begin{aligned} H_1 : \ & \sigma_1^2 = \ldots = \sigma_{l_1}^2 \neq \sigma_{l_1+1}^2 = \ldots = \sigma_{l_2}^2 \neq \ldots \\ & \cdots \neq \sigma_{l_{\tau-1}+1}^2 = \ldots = \sigma_{l_\tau}^2 \neq \sigma_{l_\tau+1}^2 = \ldots = \sigma_N^2. \end{aligned} \tag{2}$$

We develop a CP detector that only requires very general sufficient assumptions to be satisfied by the time series of content views. More specifically, we followed the work in [45] in which the authors introduce a non-parametric test statistic that requires only that the time series $\{X_n : n \in \mathbb{N}\}$ can be approximated, with a distance measure, by an s-dependent r.v. This assumption assures that time series needs not be s-dependent itself. We also note that several popular weak dependent time series models for the description of video views satisfy the above assumption, e.g., ARMA or GARCH models. The exact form of the procedure is given in the quadratic scheme,

$$TS_N^{off} = \frac{1}{N} S_n^T \hat{\Omega}_N^{-1} S_n, \tag{3}$$

with $(\cdot)^T$ denoting transposition, and, it converges in distribution asymptotically to,

$$\int_0^1 B^2(n) dn, \quad (N \to \infty), \tag{4}$$

where $(B(n) : n \in [0, 1])$ are independent standard Brownian bridges. (4) can be used to derive the critical values (cv^{off}) of the test statistic TS_N^{off} by Monte Carlo simulations that approximate the paths of the Brownian bridge on a fine grid. As an example, using this approach, the crossing boundaries of (4) for alarm rates of 5% and 1% can be found to be 1.8 and 2.6, respectively.

The detector S_n is a variation of the squared CUSUM method,

$$S_n = \frac{1}{\sqrt{N}} \left(\sum_{i=1}^{n} \mathrm{vech}[\tilde{X}_i \tilde{X}_i^T] - \frac{n}{N} \sum_{i=1}^{N} \mathrm{vech}[\tilde{X}_i \tilde{X}_i^T] \right), \tag{5}$$

where the vech(.) operator denotes the half-vectorization of a matrix (as the covariance matrix is symmetric, half-vectorization contains all the strictly necessary information) and $\tilde{X}_i = X_i - \overline{X}_N$, with $\overline{X}_N = \frac{1}{N} \sum_{j=1}^{N} X_j$ the sample average.

Since the procedure (3) is non-parametric, the dependence between the observations enters only in the form of the long-run covariance Ω_N, expressed as

$$\Omega_N = \sum_{i=1}^{N} \mathrm{Cov}(\mathrm{vech}[X_0 X_0^T], \mathrm{vech}[X_i X_i^T]) \tag{6}$$

To build a consistent estimator of Ω_N, denoted by $\hat{\Omega}_N$, various different approaches exist. This estimation problem is well studied and we focus on the kernel based approach through the use of Newey-West estimator (see [46]),

$$\widehat{\Omega}_N = \widehat{\Sigma}_0 + \sum_{w=1}^{W} k_{BT} \left(\frac{w}{W+1} \right) \left(\widehat{\Sigma}_w + \widehat{\Sigma}_w^T \right), \tag{7}$$

where $k_{BT}(.)$ corresponds to the Bartlett weight,

$$k_{BT}(x) = \begin{cases} 1 - |x|, & \text{for } |x| \leqslant 1 \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

and $\widehat{\Sigma}_w$ denotes the empirical auto-covariance matrix for lag w,

$$\widehat{\Sigma}_w = \frac{1}{N} \sum_{n=w+1}^{N} \left(X_n - \overline{X} \right) \left(X_{n-w} - \overline{X} \right)^T. \tag{9}$$

Following common practice in literature we chose $W = \log_{10}(N)$. To summarize, the existence of a CP is announced if $TS_N^{off} > cv_V^{off}$ and the estimated time of change is,

$$cp_{off}^* = \frac{1}{N} \underset{1 \leqslant n \leqslant N}{\mathrm{argmax}} \, TS_N^{off}. \tag{10}$$

Finally, to face the potential of detecting multiple CPs on the historical data set, we have integrated an extended version of the binary segmentation (BS) algorithm, proposed in [11], to the original test TS_N^{off}. The algorithm combines the standard BS and the iterative cumulative sum of squares (ICSS) [47] methods and operates briefly as follows: First, a single CP is searched in the historical sample. In case of no change, the procedure stops and H_0 is accepted. Otherwise, the detected CP is used to divide the time series into two time

series in which new searches are performed. The procedure is iterated, until no more CPs are detected. In the last step, we consider the CPs estimated previously in pairs and check if H_0 is still rejected in the segment delimited by each pair. If not, the CPs that fall in the particular segment are eliminated.

IV. ON-LINE METHODS

In this Section we present three alternative on-line approaches and discuss jointly for each one the preparation stage and the corresponding on-line CP detector. The on-line phase is based on the assumption of an homogeneous data sequence of length $m \in \mathbb{N}^+$, determined by the off-line phase, for which,

$$\sigma_1^2 = \cdots = \sigma_m^2. \tag{11}$$

Our aim is to test if (11) holds as new observations become available in a time real framework. Hence, the statistical problem is formulated as the following hypothesis test,

$$\begin{aligned} &H_0 : \sigma_1^2 = \cdots = \sigma_m^2 = \sigma_{m+1}^2 = \cdots, \\ &H_1 : \sigma_{m+1}^2 = \cdots = \sigma_{m+l-1}^2 \neq \sigma_{m+l}^2 = \sigma_{m+l+1}^2 \cdots, \\ &\quad m, l \in \mathbb{N}^+. \end{aligned} \tag{12}$$

In general, any on-line CP method can be described as a stopping time procedure with stopping time $\tau(m)$,

$$\tau(m) = \min\{l \in \mathbb{N} : TS^{on}(m, l) \geq b\}. \tag{13}$$

The value of the test statistic $TS^{on}(m, l)$ is calculated online for every l in the monitoring period. The rule stops, and a change is announced, if the test statistic exceeds the boundary function $b = cv^{on} g$. The critical value cv^{on} is derived from the asymptotic behavior of the detector TS^{on}/g under the null hypothesis, for which $Pr\,(\tau(m) < \infty) = \alpha$, $\alpha \in (0, 1)$ the significance level. We note that γ, $\gamma \in (0, \frac{1}{2}]$ is a sensitivity parameter; the larger the value of γ, the smaller the value of b, which leads to a quicker detection of a potential CP, at the cost of an increase in the false alarm rate.

Below, we consider three on-line CP approaches, based on the general assumptions for the underlying process: i) a non-parametric approach based on [48], denoted by *NP*; ii) a linear time series (ARMA) approach as in [49], denoted by *L*; and, iii) a nonlinear time series (GARCH) approach like in [50], denoted by *NL*. The quantities $\{TS^{on}, b, cv^{on}, g\}$ will be indexed accordingly.

A. NON-PARAMETRIC (NP) APPROACH

Non-parametric approaches work directly with the observed data and are ideal for datasets with a high degree of model fitting ambiguity. In this framework, in the preparation phase we only compute a particular form of the long-run estimator, avoiding the difficulties related to the estimation of a parametric model.

The proposed procedure is applied under the assumption that the observations $\{X_n : n \in \mathbb{Z}\}$ satisfy the generalized dependence concept of L-2 near epoch dependence (see [51]). Since the test is model-independent, the dependence between observations is captured through the long-run function D_n, expressed as

$$D_n := \lim_{n\to\infty} \mathrm{E}\left(\frac{1}{n} A_i A_i^T\right), \tag{14}$$

where $A_i = \sum_{t=1}^{i} \left(X_t^2 - \mathrm{E}\left(X_t^2\right)\right)$. We also assume that D_n is finite under the H_0 hypothesis, which is necessary for the convergence of the asymptotic null behaviour.

As explained above, the long-run factor is computed in the preparation phase, considering the training sample. For its evaluation we choose the kernel estimation method, as in [52]. More specifically,

$$\hat{D}_m = \sum_{i=1}^{u} \sum_{j=1}^{u} k_{BT}\left(\frac{i-j}{r}\right) \hat{V}_i \hat{V}_j^T, \tag{15}$$

is an estimator of D_m, $\hat{V}_t = \frac{1}{\sqrt{m}}\left(X_t^2 - \frac{1}{m}\sum_{i=1}^{m} X_i^2\right)$ and $k_{BT}(\cdot)$ is the Bartlett kernel, already mentioned in (7).

The test statistic is expressed as

$$TS_{NP}^{on}(m, l) = \frac{l}{\sqrt{m}} \hat{D}_m^{-\frac{1}{2}} \left(\sum_{i=m}^{m+l} X_i^2 - \frac{1}{m}\sum_{i=1}^{m} X_i^2.\right) \tag{16}$$

The boundary function $b_{NP} = cv_{NP}^{on} g_{NP}$ is strictly aligned with the chosen size of the monitoring period l normalized to the length of the training period, denoted by $H = l/m$. Then the weight function is expressed as $g_{NP} = \left(1 + \frac{l}{m}\right)\left(\frac{l}{m+l}\right)^{\gamma}$, $\gamma \in [0, 1/2)$ and the critical value is derived from the asymptotic behavior of the stopping rule,

$$\begin{aligned} &\lim_{m\to\infty} Pr\{\tau(m) < \infty\} \\ &= \lim_{m\to\infty} Pr\left\{TS_{NP}^{on} \geq b_{NP}(\alpha)\right\} \\ &= \lim_{m\to\infty} Pr\left\{\frac{TS_{NP}^{on}}{g_{NP}} \geq c_{NP}^{on}(\alpha)\right\} \\ &= Pr\left(\sup_{n\in[0,1]} \left(\frac{H}{1+H}\right)^{\frac{1}{2}-\gamma} \frac{|W(n)|}{n^{\gamma}}\right) = \alpha. \end{aligned} \tag{17}$$

B. LINEAR (L) PARAMETRIC APPROACH USING AN AUTOREGRESSIVE MOVING AVERAGE (ARMA) MODEL

Parametric approaches, monitor the estimated values obtained from a specific model fit to the observed time-series. This is very efficient whenever a parametric model sufficiently describes the dependence structure of the real data. We present two residual based parametric schemes, constructed from the residuals of the model fit to the data, starting with an ARMA model. In the preparation stage, the model residuals are estimated, under the assumption of a homogeneous underlying process. Under H_0, the residuals before and after the beginning of the monitoring should

behave similarly. On the other hand, if a CP exists in the monitoring period, the residuals are expected to deviate from those in the training period.

ARMA processes provide linear and parsimonious descriptions of (weakly) stationary processes. A time series $\{X_n : n \in \mathbb{N}\}$ is called an ARMA(p, q) process of orders p and q, if it satisfies the stochastic equation,

$$\phi_n(B)(X_n - \mu_n) = \theta_n(B)\epsilon_n, \quad n \in \mathbb{Z}, \tag{18}$$

where μ_n are mean parameters (usually non stationary), $\phi_n(z) = 1-\phi_{1n}z-\cdots-\phi_{pn}z^p$ and $\theta_n(z) = 1-\theta_{1n}z-\cdots-\theta_{qn}z^q$ are the autoregressive and moving average polynomials of the model respectively, and B the backshift operator. It is also assumed that the ARMA process is causal and invertible, i.e.,

$$\phi_n(z) \neq 0 \text{ and } \theta_n(z) \neq 0, \quad \text{for all } |z| \leq 1. \tag{19}$$

The error terms $\{\epsilon_n : n \in \mathbb{Z}\}$ are a sequence of independent and identically distributed (i.i.d) r.v. with zero mean, $\mathrm{E}(\epsilon_1) = 0$ and constant variance, $\mathrm{E}(\epsilon_1^2) = \sigma^2$.

The ARMA model in (18) depends on $p+q+2$ parameters, represented by the vector $\beta_n = (\mu_n, \phi_n, \theta_n, \sigma_n^2)$, where $\phi_n = (\phi_{1n}, \cdots, \phi_{pn})$ and $\theta_n = (\theta_{1n}, \cdots, \theta_{qn})$. In the defined training period of size m the parameters of the ARMA model are not time dependent, i.e., they are the same for the observations $X_1, \cdots, X_m$, denoted by β_0 in the following,

$$\beta_0 = (\mu_0, \phi_0, \theta_0, \sigma_0^2). \tag{20}$$

The preparation stage is applied to the training sample for two reasons. Firstly, in order to specify the order (p, q) of the corresponding ARMA model, by selecting the combination that provides the lower value for the Bayes information criterion (BIC),

$$BIC = -2\ln(\hat{L}) + k\ln(n), \tag{21}$$

where $\hat{L}$ is the maximum value of the likelihood function of the model, k is the number of the estimated parameters and n is the sample size. Secondly, in order to estimate the parameters β_0 of the ARMA model through the estimators $\hat{\beta}_0 = (\hat{\mu}_0, \hat{\phi}_0, \hat{\theta}_0, \hat{\sigma}_0^2)$, computed, for example, by the method of maximum likelihood estimation or least squares.

Then, the model residuals are given by

$$\hat{\epsilon}_n = \hat{X}_n - \sum_{i=1}^{p} \hat{\phi}_{i0}\hat{X}_{n-i} - \sum_{i=1}^{q} \hat{\theta}_{i0}\hat{\epsilon}_{n-i}, \tag{22}$$

where $\hat{X}_n = X_n - \hat{\mu}_0$. The detector is built from the (squared) residuals $\hat{\epsilon}_n$, as:

$$\frac{1}{\sqrt{m}}TS_L^{on}(m, l) = \frac{1}{\sqrt{m}\hat{\eta}_m}\left|\sum_{n=m+1}^{m+l}\hat{\epsilon}_n^2 - \sum_{n=1}^{m}\hat{\epsilon}_n^2\right|, \tag{23}$$

where $\hat{\eta}_m^2$ is a weakly consistent estimator of the moment $\eta_m^2 = \mathrm{E}\left[\left(\epsilon_m^2 - \sigma_m^2\right)^2\right]$.

Finally, the boundary function is expressed as $b_L = cv_L^{on} g_L$, where $g_L = \left(1 + \frac{l}{m}\right)\left(\frac{l}{m+l}\right)^{\gamma}$, $\gamma \in [0, 1/2)$ and the critical value is obtained according to [49] as

$$\begin{aligned}\lim_{m\to\infty} Pr\{\tau(m) < \infty\} &= \lim_{m\to\infty} Pr\left\{\frac{TS_L^{on}}{g_L} \geq c_L^{on}(\alpha)\right\} \\ &= Pr\left(\sup_{n\in(0,1)} \frac{|W(n)|}{n^{\gamma}} \geq cv_L^{on}(\alpha)\right) = \alpha.\end{aligned} \tag{24}$$

C. NONLINEAR (NL) PARAMETRIC APPROACH USING A GENERALIZED AUTOREGRESSIVE CONDITIONAL HETEROSKEDASTICITY (GARCH) MODEL

A time series $\{X_n : n \in \mathbb{Z}\}$ follows the GARCH(p, q) process, if,

$$\begin{aligned}X_n &= \sigma_n\epsilon_n, \\ \sigma_n^2 &= \omega_n + \sum_{i=1}^{q}\alpha_{in}X_{n-i}^2 + \sum_{j=1}^{p}\beta_{jn}\sigma_{n-j}^2,\end{aligned}$$

where $\omega_n > 0$, $\alpha_{in}, \beta_{jn} \geqslant 0$ and $\{\epsilon_n : n \in \mathbb{Z}\}$ is a sequence of i.i.d r.v. with $\mathrm{E}(\epsilon_1) = 0$ and $\mathrm{E}(\epsilon_1^2) = 1$. We estimate the set of parameters θ_m during the initial training period, denoted in the following by $\theta_0 = (\omega_0, \alpha_{10,}, \cdots, \alpha_{q0}, \beta_{10}, \cdots, \beta_{p0})$; the estimation is performed by applying the Gaussian maximum-likelihood estimator (GMLE) $\hat{\theta}_0$ of θ_0 on the last m observations, as proposed in [53]. The GMLE function is given by

$$F_m(\theta; X_1, \cdots, X_m) = \prod_{n=1}^{m}\frac{1}{\sqrt{2\pi\hat{\sigma}_n^2}}\exp\left(-\frac{X_n^2}{2\hat{\sigma}_n^2}\right), \tag{25}$$

where $\hat{\sigma}_n^2$ are constructed recursively, as,

$$\hat{\sigma}_n^2 = \omega_n + \sum_{i=1}^{q}\alpha_{in}X_{n-i}^2 + \sum_{j=1}^{p}\beta_{jn}X_{n-j}^2. \tag{26}$$

Then, the GMLE of θ_m is,

$$\begin{aligned}\hat{\theta}_m &= \underset{\theta\in\Theta}{\operatorname{argmax}}\, F_m(\theta; X_1, \cdots, X_m) \\ &= \underset{\theta\in\Theta}{arg\,min}\, \frac{1}{m}\sum_{n=1}^{m}\left(\frac{X_n^2}{\hat{\sigma}_n^2} + \ln(\hat{\sigma}_n^2)\right).\end{aligned} \tag{27}$$

The residuals of the GARCH process are subsequently obtained from the GMLE as

$$\hat{\epsilon}_n = \frac{X_n}{\hat{\sigma}_n(\hat{\theta}_m)}. \tag{28}$$

Based on the (squared) residuals, the test statistic is described as in [54],

$$TS_{NL}^{on}(m, l) = \sqrt{\frac{m}{\mathrm{Var}(\hat{\epsilon}_m^2)}}\left|\frac{1}{l}\sum_{n=1}^{l}\hat{\epsilon}_n^2 - \frac{1}{m}\sum_{n=1}^{m}\hat{\epsilon}_n^2\right|, \tag{29}$$

where $\mathrm{Var}(\hat{\epsilon}_n^2)$ denotes the variance of the squared residuals of the training period, i.e., $\mathrm{Var}(\hat{\epsilon}_m^2) = \mathrm{E}(\hat{\epsilon}_m^4) - \left(\mathrm{E}\left(\hat{\epsilon}_m^2\right)\right)^2$.

Considering the boundary function $b_{NL} = cv_{NL}^{on} g_{NL}$, we choose to work with $g_{NL} = 1$ as in [50]; consequently, the critical value is given by

$$\lim_{m\to\infty} Pr\{\tau(m) < \infty\} = \lim_{m\to\infty} Pr\left\{TS_{NL}^{on} \geq cv_{NL}^{on}(\alpha)\right\} = Pr\left(\sup_{n\in(0,1)} |W(n)| \geq cv_{NL}^{on}(\alpha)\right) = \alpha. \tag{30}$$

D. EVALUATION OF THE CRITICAL VALUES FOR THE CPS TESTS

The on-line critical values for the three procedures are estimated using Monte Carlo simulations, similarly to the off-line case, considering that

$$cv_{NP}^{on}(\alpha) = \sup_{n\in[0,1]} \left(\frac{H}{1+H}\right)^{\frac{1}{2}-\gamma} \frac{|W(n)|}{n^{\gamma}}, \tag{31}$$

$$cv_{L}^{on}(\alpha) = \sup_{n\in(0,1)} \frac{|W(n)|}{n^{\gamma}}, \tag{32}$$

$$cv_{NL}^{on}(\alpha) = \sup_{n\in(0,1)} |W(n)|. \tag{33}$$

With respect to the estimation of the magnitude of a detected CP denoted by cp_{on}^{*}, in the *NP* scenario, we estimate the deviation of the variance in pre-CP and post-CP data by comparing the variance of a pre-determined historical subsample, $\mathrm{Var}(X_{m_s} : X_{cp_{on}^{*}-h})$ to the variance "in the range" of the detected CP as $\mathrm{Var}(X_{cp_{on}^{*}-h} : X_{cp_{on}^{*}+h})$, accounting for the fact that a time lag $\pm h$ is required to establish the presence of an actual change.

We finally propose an alternative scheme to predict the post CP behavior in the case of a parametric model. We apply the parametric model (ARMA or GARCH) on the time horizon $t_{cp_{on}^{*}-h}, \ldots, t_{cp_{on}^{*}}$, in which we assume that the actual change has already occurred. Thus, a well defined subsample is provided to fit the model parameters and predict the next values using this adaptive model.

V. PERFORMANCE EVALUATION OF THE VARIANCE CP DETECTION APPROACHES ON SYNTHETIC DATA

In this Section, we evaluate the performance of the integrated algorithm with the three aforementioned variations of the on-line phase – *NP*, *L* and *NL*, – on two sets of synthetic data. In further detail, we report the results of Monte Carlo simulations using either an ARMA(1,1) or a GARCH(1,1) process to generate the time series; as a reminder, both ARMA and GARCH are well known models that have been shown to fit well video content popularity dynamics (see section II).

The synthetic sample size under consideration is $N = 1000$ while we introduce a variance CP at $cp^{*} = 500$; this is achieved by transforming the initial parameters vector of the chosen model. Evaluations are conducted based on 1000 repetitions for a significance level $\alpha = 0.01$. In all tests we set the beginning of the monitoring period at $m_s = 200$, the monitoring window length at $l = 100$ and the minimum interval between two successive CPs at $d = 80$ (this latter choice is justified by experiments with real data that will be presented in Section V). We experiment with two values for the sensitivity parameter $\gamma \in \{0, 0.25\}$ (as a reminder, γ only affects cv_{NP}^{on} and cv_{L}^{on}, see (31) and (32)).

We first evaluate the performance of the three alternative on-line procedures in the integrated algorithm, for a wide range of ARMA(1,1) models. We recall that the variance of an ARMA(1,1) model depends on the model parameters ϕ_i, θ_i and the variance of the error terms σ_i^2, i.e.,

$$\mathrm{Var}(X_n) = \frac{\left(1 + 2\phi_i\theta_i + \theta_i^2\right)\sigma_i^2}{1 - \phi_i^2}.$$

We consider a change by transforming the time series model defined by the parameter vector β_0 to one of the vectors β_i, $i = 1, 2, 3, 4$.

- Model 0: $\beta_0 = (\phi_0, \theta_0, \sigma_0) = (0.4, 0.2, 0.5)$,
- Model 1: $\beta_1 = (0.4, 0.2, 1)$,
- Model 2: $\beta_2 = (0.3, 0.3, 1.5)$,
- Model 3: $\beta_3 = (0.5, 0.3, 1.5)$,
- Model 4: $\beta_4 = (0.4, 0.2, 2)$.

We use Model 0 as the baseline. In Model 1 a small change in the error variance is introduced, which increases the uncertainty. Models 2 and 3 lead to medium changes in the variance and also transform the dependence structure between the r.v. On the other hand in Model 4 a large change is introduced by increasing the uncertainty.

In Table 1 we report the results of the simulation study. We depict the aggregate percentage of the CPs over the multitude of the simulations. For every test and each iteration we calculate the exact number of CPs detected:

- 0 when no CPs are detected, denoting the percentage of false negatives in all cases but the first (in which β_0 does not change); in this latter case it corresponds to the true success rate;
- 1 when a single CP is detected, denoting the true success rate in all cases but the first, in which it corresponds to a false positive rate;
- > 1 when more than one CPs are detected, denoting the percentage of false positives, in all cases other than the first. To obtain the overall false positive percentage, this value needs to be added to the false positive percentage above.

Furthermore, we denote by $\hat{cp}^{*}$ the median of the time instance of the identification of the true CP, evaluated in all cases but the first. The closest this number to the true point of the CP at 500, the quicker the detection and the better the responsiveness of the integrated algorithm.

Initially, we discuss the impact of the choice of the sensitivity parameter γ in the *L* and *NP* approaches. Studying Table 1, we conclude that $\gamma = 0$ is the most reasonable choice in the case of medium or more significant changes in the variance, since it leads to significantly lower false positive rates. On the other hand, in the case of only small changes in the variance, captured in our study in the

TABLE 1. Results from an ARMA generating process and for one change in the variance.

		ARMA(1,1)											
β	γ	Non parametric approach (NP)				ARMA approach (L)				GARCH approach (NL)			
		Detected CPs			$\hat{cp}^*$	Detected CPs			$\hat{cp}^*$	Detected CPs			$\hat{cp}^*$
		0	1	>1	med	0	1	>1	med	0	1	>1	med
β_0	0	**0.99**	0.01	0	-	**0.99**	0.01	0	-	**0.98**	0.02	0	-
	0.25	**0.95**	0.05	0	-	**0.98**	0.02	0	-				
β_1	0	0.49	**0.5**	0.01	-	0.48	**0.52**	0	554	0.74	**0.26**	0	-
	0.25	0.18	**0.76**	0.06	549	0.07	**0.93**	0	548				
β_2	0	0.04	**0.94**	0.02	550	0.03	**0.95**	0.02	546	0.15	**0.83**	0.02	549
	0.25	0	**0.93**	0.07	531	0	**0.96**	0.04	521				
β_3	0	0.01	**0.96**	0.03	536	0.01	**0.98**	0.01	535	0	**0.97**	0.03	548
	0.25	0	**0.92**	0.08	521	0	**0.97**	0.03	521				
β_4	0	0	**0.97**	0.03	533	0	**0.99**	0.01	530	0.01	**0.97**	0.02	544
	0.25	0	**0.93**	0.07	519	0	**0.97**	0.03	513				

TABLE 2. Results from a GARCH generating process and for one change in the variance.

		GARCH(1,1)											
θ	γ	non parametric approach (NP)				ARMA approach (L)				GARCH approach (NL)			
		Detected CPs			$\hat{cp}^*$	Detected CPs			$\hat{cp}^*$	Detected CPs			$\hat{cp}^*$
		0	1	>1	med	0	1	>1	med	0	1	>1	med
θ_0	0	**0.85**	0.15	0	-	**0.75**	0.25	0	-	**0.9**	0.1	0	-
	0.25	**0.65**	0.35	0	-	**0.42**	0.58	0	-				
θ_1	0	0.16	**0.8**	0.04	527	0.03	**0.77**	0.23	528	0.04	**0.92**	0.04	550
	0.25	0	**0.87**	0.13	521	0	**0.6**	0.4	515				
θ_2	0	0.03	**0.87**	0.1	524	0.01	**0.76**	0.23	521	0.01	**0.93**	0.06	544
	0.25	0.01	**0.85**	0.14	516	0	**0.56**	0.44	510				
θ_3	0	0	**0.93**	0.07	511	0	**0.7**	0.3	511	0	**0.93**	0.07	531
	0.25	0	**0.81**	0.19	508	0	**0.58**	0.42	505				

transformation from the β_0 to the β_1 model, a higher value of γ is needed (intuitively, for smaller changes a larger sensitivity is required). Therefore, depending on whether smaller or larger deviations need to be rapidly detected we can fine-tune the value of γ. For the sake of simplicity, in the following we focus on $\gamma = 0$ (larger deviations).

According to Table 1, the three approaches provide appropriate empirical sizes, and the false alarm rates are in all cases close to the significance level $\alpha = 0.01$. Overall the L procedure outperforms the *NP* and the *NL*, both in terms of the true alarm rates as well as in terms of the detection time; this is intuitive as in this first experiment the underlying process is generated by a linear ARMA(1,1) model and therefore a linear parametric model is excellently suited to capture the underlying dynamics. Furthermore, comparing the *NP* and the *NL* approaches, Table 1 illustrates that the *NP* is more sensitive than the *NL* approach, leading to more accurate detection for small changes at the cost of increased false positive rates in the case of larger changes. The opposite is true for the *NL* approach that appears to be more "conservative". Moreover, the fact that the *NP* procedure is statistically more sensitive leads to a quicker detection of a CP as captured through $\hat{cp}^*$.

We proceed to the more challenging case of a GARCH(1,1) generating model, with parameter vector $\theta_i = (\omega_i, \alpha_i, \beta_i)$ that fully describes the model and unconditional variance,

$$\text{Var}(X_n) = \frac{\omega_i}{(1 - \alpha_i - \beta_i)}.$$

To examine the alarm rates we assume the following models,

- Model 0: $\theta_0 = (\omega_0, \alpha_0, \beta_0) = (0.05, 0.4, 0.3)$,
- Model 1: $\theta_1 = (0.5, 0.2, 0.1)$,
- Model 2: $\theta_2 = (0.5, 0.3, 0.2)$,
- Model 3: $\theta_3 = (1, 0.3, 0.2)$.

GARCH is a varying volatility model, allowing volatility changes over time. Being more elaborate and complex in terms of the dependence of the variance on the model parameters, the higher false alarm and the lower true alarm rates in Table 2 are reasonable. In this case, the L procedure seems fully inappropriate irrespective of the choice of $\gamma = 0$ or $\gamma = 0.25$, suffering from very high false positive rates, since constant variance is assumed. The *NL* procedure, as expected, surpasses both the L and the *NP* procedures, as it is excellently suited to capture the GARCH process. More specifically, the true alarm rate estimation is stable for the different magnitudes of changes, with a detection time lag

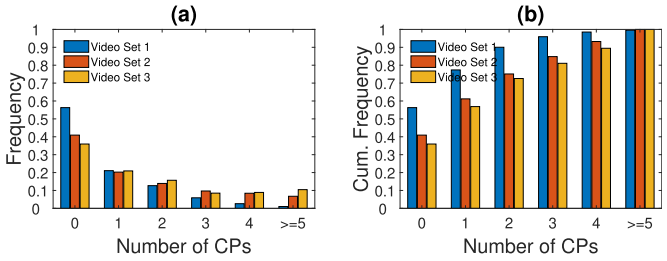


FIGURE 3. Estimated a) frequency and b) cumulative frequency of the number of CPs per time series, for three different Video Sets.

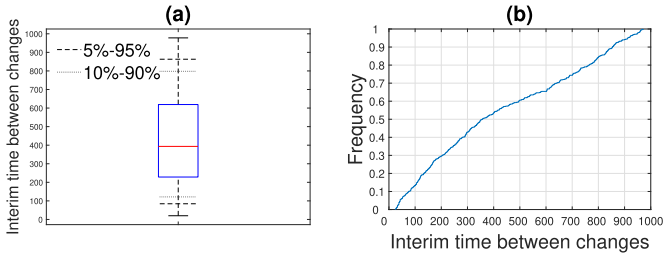


FIGURE 4. Interim time between consecutive CPs: a) Boxplot including the interval (5% – 95%) (dashed line) and (10% – 90%) interval (dotted line), b) Cumulative frequency for the interim time of consecutive CPs.

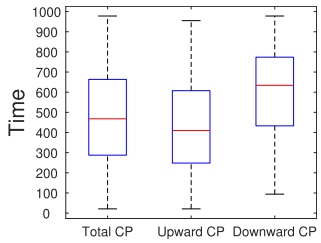


FIGURE 5. Boxplot of the number of upward and downward CPs, per time series.

ranging from 50 instances for small changes to 31 instances for larger changes. On the other hand, the *NP* procedure appears to capture well the actual changes for $\gamma = 0$, with success rates relatively close to the those of the *NL* procedure, especially for medium/large changes. However, for $\gamma = 0.25$, the approach leads to ineligible false positive rates, despite the fact that it can identify small changes more efficiently. The *NP* method also achieves faster detection of changes, with $\hat{cp}^*$ ranging from 5 to 28 time instances.

Based on the analysis of the Monte Carlo results for the three procedures under the two different time series generating models, we can synthesize our overall conclusions in the following two points:

1) The *NL* and the *NP* approaches adapt better to a wider range of models and underlying assumptions; if there are indications of a highly nonlinear underlying procedure the *NP* approach could render better results;
2) The *L* approach is strongly related to the ARMA model assumptions and therefore it is advisable to be applied only if these can be readily shown to hold.

VI. ILLUSTRATION OF THE INTEGRATED ALGORITHM USING REAL DATA

Finally, we study the performance of the proposed algorithms on monitoring real YouTube video traces provided within the framework of the CONGAS project [55]; the dataset consists of 882 videos traces and the observation period is of $N = 1000$ time instances.

In this Section, we only adopt the non parametric (*NP*) and the GARCH (*NL*) approaches. We exclude the ARMA (*L*) approach from the evaluation, based on the conclusions of the previous Section. We work with the centered simple returns of the content popularity time series,

$$Y_n = (X_{n+1} - X_n) - \frac{1}{900}\sum_{n=1}^{900}(X_{n+1} - X_n), \quad n = 1, \ldots, 900$$

and then apply the methods on Y_n.

In order to clarify some general characteristics of the dataset, in terms of changing content dynamics, we first apply the off-line algorithm to the video traces. In Fig. 3, we consider three video sets; Video Set 1 contains the whole dataset, Video Set 2 contains the videos with average number of visits $\mathrm{E}((Y(1) : Y(1000)) \geq 10$ and Video Set 3 contains the videos with average number of visits greater or equal to 20. Fig. 3, depicts a high percentage of rejecting the H_0 hypothesis, for a significance level of $\alpha = 0.05$. Especially for the Video Sets 1 and 2, the rejection of the assumption of normal behavior exceeds 60% and 65% of the time series, respectively. This result confirms that a sufficiently high number of time series provide content popularity anomalies, for example in Video Set 3, in 10% of the cases there are over than four CPs per time series. This small analysis confirms the suitability of change point analysis as a viable approach for the detection of changes in video content popularity dynamics.

Subsequently, in Fig. 4, we analyze the interim time between consecutive CPs. The respective boxplot diagrams illustrate the existence of sufficiently large intervals between consecutive changes; this fact supports our subtle assumption in Section III regarding the existence of a sufficient gap between two consecutive CPs (e.g., $>$ 80 instances). In particular, 90% and 95% of the intervals correspond to consecutive CPs exceeding 100 and 80 time instances, respectively. This outcome assures that a sufficiently large training window after a detected change can be applied, denoted by the parameter d.

Additionally, Fig. 5, illustrates the time instances of upward (increase in volatility) and downward changes (decrease in volatility) in the form of a boxplot. It is shown that upward changes occur earlier in time than downward changes.

We consider now the performance of the on-line approach, by illustrating the estimated CPs in the second order characteristics of different time series. We choose the beginning of the monitoring period at $m_s = 200$, the sensitivity parameter

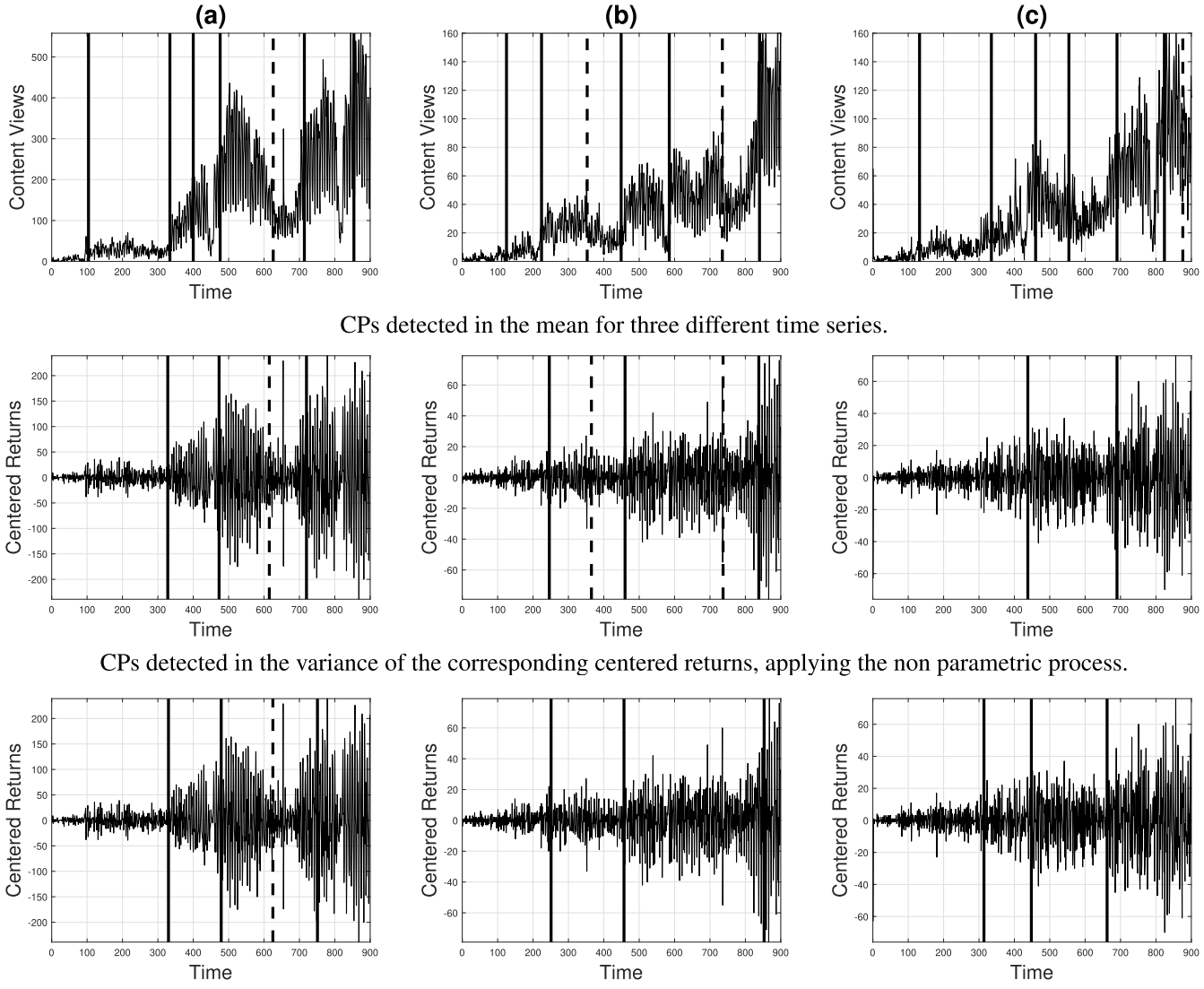


FIGURE 6. CPs detected in the mean (first row) and variance (second and third row) for three different content views time series. Solid and dashed lines represent an upward and a downward change, respectively.

$\gamma = 0$ and the significance level $\alpha = 0.05$. To fit a GARCH(p, q) model we consider all the possible combinations of the $p, q = 1, \cdots, 4$ and choose the orders p, q that minimize the Akaike information criterion (AIC).

The corresponding results are depicted in Fig. 6 at the top of the next page. The first row of results represent the detected changes in the mean value by using the RCPD algorithm presented in [11]. In the second and third row the estimated CPs in the variance are depicted, for the same time series, by applying on the first order differences Y_n the non parametric (*NP*) approach and the GARCH (*NL*) approach, respectively. Solid lines represent upwards changes while dashed lines represent downward changes.

Firstly, we observe that the variance changes are closely connected to a corresponding mean change. In particular, variance changes are less in multitude and seem to be related to the most significant mean changes, which can be intuitively explained by considering that if the average number of views changes significantly, the variance in the number of views at the respective interval will follow a similar trend. The importance of jointly studying the changes in the mean and the variance value is also depicted in Fig. 6. For instance, in Fig. 6a, to describe or handle the content popularity dynamics it is crucial to estimate quickly the‘‘explosion’’ in variance after time instances 500 or 700, that leads to a high instability of the values from the mean. On the other hand, variance ‘‘reduction’’ detection is also important, as it implies that values remain relatively constant, like in Fig. 6a between time instances 600 and 700.

Both the *NP* and the *NL* approaches provide similar results in terms of the number of CPs and the detection time of the estimated CPs. More precisely, in Fig. 6a, both procedures detect the same number of changes, while the *NP* method gives a slightly quicker detection.

Focusing on the capability of the proposed algorithm to estimate the magnitude of a detected CP, we use the GARCH model. We estimate the parameters of the model considering 10 time instances before the detected change and forecast the variance for 10 time instances after the CP. For the time series in Fig. 6b, the actual variance after each change is 7.92, 12.51 and 38.66, while the predicted variance values are 7.39, 13.52 and 39.24, respectively. As we observe, in this case the *NL* algorithm can efficiently describe the post change variance behavior.

In the future, we will develop a joint approach identifying CPs simultaneously in the first and the second order characteristics, providing an aggregated and compact view of content popularity dynamics.

VII. CONCLUSION AND FUTURE WORK

In this paper, we presented an integrated algorithm for the detection of changes in the variance of a time series. We proposed to combine an off-line approach during which algorithmic and model parameters are learned. Subsequently, during the on-line part of the algorithm, changes in the variance of the time series are identified using a stopping time procedure. Whenever the value of a test statistic surpasses a predefined critical value, a change is declared.

To develop the test statistic we proposed three different approaches: i) a non-parametric approach, ii) a parametric approach using an ARMA model, and, iii) a parametric approach using a nonlinear GARCH model. Our studies using synthetic data indicated that the ARMA parametric approach does not generalize well. Due to this fact, we only performed experiments on real data using the non-parametric and the GARCH approaches. We concluded that both can equally well identify large deviations in the variance and that in the general case the non-parametric approach can provide quicker detection of CPs in the datasets studied in this work. In the future, we will develop joint detectors for the mean and the variance of video content popularity.

REFERENCES

[1] P. Sermpezis and T. Spyropoulos, "Effects of content popularity on the performance of content-centric opportunistic networking: An analytical approach and applications," *IEEE/ACM Trans. Netw.*, vol. 24, no. 6, pp. 3354–3368, Dec. 2016.

[2] M. Z. Shafiq, A. R. Khakpour, and A. X. Liu, "Characterizing caching workload of a large commercial content delivery network," in *Proc. IEEE INFOCOM, 35th Annu. IEEE Int. Conf. Comput. Commun.*, San Francisco, CA, USA, Apr. 2016, pp. 1–9.

[3] M. A. Salahuddin, J. Sahoo, R. Glitho, H. Elbiaze, and W. Ajib, "A survey on content placement algorithms for cloud-based content delivery networks," *IEEE Access*, vol. 6, pp. 91–114, 2018.

[4] S. Li, J. Xu, M. Van Der Schaar, and W. Li, "Popularity-driven content caching," in *Proc. IEEE INFOCOM, 35th Annu. IEEE Int. Conf. Comput. Commun.*, San Francisco, CA, USA, Apr. 2016, pp. 1–9.

[5] Y. Liu, T. Zhi, H. Xi, X. Duan, and H. Zhang, "A novel content popularity prediction algorithm based on auto regressive model in information-centric IoT," *IEEE Access*, vol. 7, pp. 27555–27564, 2019.

[6] F. Guillemin, B. Kauffmann, S. Moteau, and A. Simonian, "Experimental analysis of caching efficiency for YouTube traffic in an ISP network," in *Proc. 25th Int. Teletraffic Congr. (ITC)*, Shanghai, China, 2013, pp. 1–9.

[7] W.-X. Liu, J. Zhang, Z.-W. Liang, L.-X. Peng, and J. Cai, "Content popularity prediction and caching for ICN: A deep learning approach with SDN," *IEEE Access*, vol. 6, pp. 5075–5089, 2018.

[8] Y. Zheng, L. Liu, L. Wang, and X. Xie, "Learning transportation mode from raw GPS data for geographic applications on the Web," in *Proc. 17th Int. Conf. World Wide Web (WWW)*, Beijing, China, Apr. 2008, pp. 247–256.

[9] J. Song, M. Sheng, T. Q. S. Quek, C. Xu, and X. Wang, "Learning-based content caching and sharing for wireless networks," *IEEE Trans. Commun.*, vol. 65, no. 10, pp. 4309–4324, Oct. 2019.

[10] F. Jiang, Z. Yuan, C. Sun, and J. Wang, "Deep Q-learning-based content caching with update strategy for fog radio access networks," *IEEE Access*, vol. 7, pp. 97505–97514, 2019.

[11] S. Skaperas, L. Mamatas, and A. Chorti, "Early video content popularity detection with change point analysis," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Abu Dhabi, UAE, Dec. 2018, pp. 1–7.

[12] S. Skaperas, L. Mamatas, and A. Chorti, "Real-time video content popularity detection based on mean change point analysis," *IEEE Access*, vol. 7, pp. 142246–142260, 2019.

[13] V. Jandhyala, S. Fotopoulos, I. Macneill, and P. Liu, "Inference for single and multiple change-points in time series," *J. Time Anal.*, vol. 34, pp. 423–446, May 2013.

[14] A. Aue and L. Horváth, "Structural breaks in time series," *J. Time Anal.*, vol. 34, no. 1, pp. 1–16, Jan. 2013.

[15] C. Callegari, A. Coluccia, A. D'Alconzo, W. Ellens, S. Giordano, M. Mandjes, M. Pagano, T. Pepe, F. Ricciato, and P. Zuraniewski, "A methodological overview on anomaly detection," in *Data Traffic Monitoring and Analysis*. Berlin, Germany: Springer-Verlag, 2013, pp. 148–183.

[16] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, Jul. 2009.

[17] S. Aminikhanghahi and D. J. Cook, "A survey of methods for time series change point detection," *Knowl. Inf. Syst.*, vol. 51, no. 2, pp. 339–367, May 2017.

[18] D. Choudhary, A. Kejariwal, and F. Orsini, "On the runtime-efficacy trade-off of anomaly detection techniques for real-time streaming data," Oct. 2017, *arXiv:1710.04735*. [Online]. Available: https://arxiv.org/abs/1710.04735

[19] J. Chen and A. K. Gupta, *Parametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance*. Cambridge, MA, USA: Birkhäuser 2000.

[20] G. Tsechpenakis, D. N. Metaxas, C. Neidle, and O. Hadjiliadis, "Robust online change-point detection in video sequences," in *Proc. Comput. Vis. Patt. Recog. Work.*, New York, NY, USA, Jun. 2006, p. 155.

[21] V. Konev and S. Vorobeychikov, "Quickest detection of parameter changes in stochastic regression: Nonparametric CUSUM," *IEEE Trans. Inf. Theory*, vol. 63, no. 9, pp. 5588–5602, Sep. 2017.

[22] A. G. Tartakovsky, A. S. Polunchenko, and G. Sokolov, "Efficient computer network anomaly detection by change point detection methods," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 1, pp. 4–11, Feb. 2013.

[23] A. Soule, K. Salamatian, and N. Taft, "Combining filtering and statistical methods for anomaly detection," in *Proc. 5th ACM SIGCOMM Conf. Internet Meas. (IMC)*, New York, NY, USA, Oct. 2005, p. 31.

[24] I. Nevat, D. M. Divakaran, S. G. Nagarajan, P. Zhang, L. Su, L. L. Ko, and V. L. L. Thing, "Anomaly detection and attribution in networks with temporally correlated traffic," *IEEE/ACM Trans. Netw.*, vol. 26, no. 1, pp. 131–144, Feb. 2018.

[25] G. Thatte, U. Mitra, and J. Heidemann, "Parametric methods for anomaly detection in aggregate traffic," *IEEE/ACM Trans. Netw.*, vol. 19, no. 2, pp. 512–525, Apr. 2011.

[26] S. Lee, J. Ha, O. Na, and S. Na, "The cusum test for parameter change in time series models," *Scandin. J. Statist.*, vol. 30, no. 4, pp. 781–796, Dec. 2003.

[27] I. Berkes, E. Gombay, L. Horváth, and P. Kokoszka, "Sequential change-point detection in GARCH (p, q) models," *Econ. Theory*, vol. 20, no. 6, pp. 1140–1167, Dec. 2004.

[28] Y. Borghol, S. Mitra, S. Ardon, N. Carlsson, D. Eager, and A. Mahanti, "Characterizing and modelling popularity of user-generated videos," *Perform. Eval.*, vol. 68, no. 11, pp. 1037–1055, Nov. 2011.

[29] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Commun. ACM*, vol. 53, no. 8, pp. 80–88, Aug. 2010.

[30] J. Xu, M. Van Der Schaar, J. Liu, and H. Li, "Forecasting popularity of videos using social media," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 2, pp. 330–343, Mar. 2015.

[31] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "Youtube traffic characterization: A view from the edge," in *Proc. 7th ACM SIGCOMM Conf. Internet Meas.*, San Diego, CA, USA, 2007, pp. 15–28.
[32] S. Tanwir and H. Perros, "A survey of VBR video traffic models," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 4, pp. 1778–1802, 4th Quart., 2013.
[33] M. Tsagkias, W. Weerkamp, and M. De Rijke, "News comments: Exploring, modeling, and online prediction," in *Proc. Eur. Conf. Inform. Retr.*, Milton Keynes, U.K., Mar. 2010, pp. 191–203.
[34] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "Analyzing the video popularity characteristics of large-scale user generated content systems," *IEEE/ACM Trans. Netw.*, vol. 17, no. 5, pp. 1357–1370, Oct. 2009.
[35] T. Trzcinski and P. Rokita, "Predicting popularity of online videos using support vector regression," *IEEE Trans. Multimedia*, vol. 19, no. 11, pp. 2561–2570, Nov. 2017.
[36] S. Romano and H. Elaarag, "A neural network proxy cache replacement strategy and its implementation in the Squid proxy server," *Neural Comput. Appl.*, vol. 20, no. 1, pp. 59–78, Feb. 2011.
[37] W. Ali and S. M. Shamsuddin, "Intelligent client-side Web caching scheme based on least recently used algorithm and neuro-fuzzy system," in *Proc. Int. Symp. Neural Netw. (ISNN)*, Wuhan, China, May 2009, pp. 70–79.
[38] G. Gursun, M. Crovella, and I. Matta, "Describing and forecasting video access patterns," in *Proc. IEEE INFOCOM*, Shanghai, China, Apr. 2011, pp. 16–20.
[39] D. Niu, Z. Liu, B. Li, and S. Zhao, "Demand forecast and performance prediction in peer-assisted on-demand streaming systems," in *Proc. IEEE INFOCOM*, Shanghai, China, Apr. 2011, pp. 421–425.
[40] C. Katris and S. Daskalaki, "Generation of synthetic video traffic using time series," *Simul. Model. Pract. Theory*, vol. 75, pp. 127–145, Jun. 2017.
[41] B. Zhou, D. He, Z. Sun, and W. H. Ng, "Network traffic modeling and prediction with ARIMA/GARCH," in *Proc. HET-NETs Conf.*, IIkley, U.K., Jul. 2005, pp. 1–10.
[42] C. Katris and S. Daskalaki, "Dynamic bandwidth allocation for video traffic using FARIMA-based forecasting models," *J Netw. Syst. Manage.*, vol. 27, no. 1, pp. 39–65, Jan. 2019.
[43] D. Niu, H. Xu, and B. Li, "Resource auto-scaling and sparse content replication for video storage systems," *ACM Trans. Model. Perform. Eval. Comput. Syst.*, vol. 2, no. 4, pp. 1–30, Nov. 2017.
[44] P. Valsamas, S. Skaperas, and L. Mamatas, "Elastic content distribution based on unikernels and change-point analysis," in *Proc. 24th Eur. Wireless Conf. (EW)*, Catania, Italy, 2018, pp. 1–7.
[45] A. Aue, S. Hörmann, L. Horváth, and M. Reimherr, "Break detection in the covariance structure of multivariate time series models," *Ann. Statist.*, vol. 37, no. 6B, pp. 4046–4087, Dec. 2009.
[46] D. W. K. Andrews, "Heteroskedasticity and autocorrelation consistent covariance matrix estimation," *Econometrica*, vol. 59, no. 3, pp. 817–858, May 1991.
[47] C. Inclan and G. C. Tiao, "Use of cumulative sums of squares for retrospective detection of changes of variance," *J. Amer. Stat. Assoc.*, vol. 89, no. 427, pp. 913–923, Sep. 1994.
[48] K. Pape, D. Wied, and P. Galeano, "Monitoring multivariate variance changes," *J. Empirical Finance*, vol. 39, pp. 54–68, Dec. 2016.
[49] A. Aue, C. Dienes, S. Fremdt, and J. Steinebach, "Reaction times of monitoring schemes for ARMA time series," *Bernoulli*, vol. 21, no. 2, pp. 1238–1259, May 2015.
[50] O. Na, Y. Lee, and S. Lee, "Monitoring parameter change in time series models," *Stat. Methods Appl.*, vol. 20, no. 2, pp. 171–199, Jun. 2011.
[51] J. Davidson, *Stochastic Limit Theory: An Introduction for Econometricians*. New York, NY, USA: Oxford Univ. Press, 1994.
[52] D. Wied, M. Arnold, N. Bissantz, and D. Ziggel, "A new fluctuation test for constant variances with applications to finance," *Metrika*, vol. 75, no. 8, pp. 1111–1127, Nov. 2012.
[53] C. Francq and J.-M. Zakoïan, "Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes," *Bernoulli*, vol. 10, no. 4, pp. 605–637, Aug. 2004.
[54] W. L. N. Leung, S. Him, and C. Y. Yau, "Sequential change-point detection in time series models based on pairwise likelihood," *Statistica Sinica*, vol. 27, no. 2, pp. 575–605, Apr. 2017.
[55] F. De Pellegrini, M. Zeni, and D. Miorandi, "YOUStatAnalyzer: A tool for analysing the dynamics of YouTube content popularity," in *Proc. 7th Int. Conf. Perform. Eval. Methodol. Tools*, Torino, Italy, Dec. 2013, pp. 286–289.

SOTIRIS SKAPERAS (Student Member, IEEE) received the B.Sc. degree in mathematics and the M.Sc. degree in statistics and modeling from the Department of Mathematics, Aristotle University of Thessaloniki, Greece, in 2013 and 2016, respectively. He is currently pursuing the Ph.D. degree in the area of resource management in 5G networks with the Department of Applied Informatics, University of Macedonia, Thessaloniki. He involved in the areas of resource allocation/load balancing for 5G networks using time-series/change point analysis and stochastic modeling.

LEFTERIS MAMATAS (Member, IEEE) received the Diploma and Ph.D. degrees from the Department of Electrical and Computer Engineering, Democritus University of Thrace, Greece, in 2003 and 2008, respectively. He was a Researcher with the University College London, U.K., the Space Internetworking Center/ Democritus University of Thrace, Greece, and the DoCoMo Euro-Labs, Munich, Germany. He is currently an Assistant Professor with the Department of Applied Informatics, University of Macedonia, Thessaloniki, Greece, where he leads the Softwarized and Wireless Networks Research Group. He has published more than 60 articles in international journals and conferences. His research interests include the areas of software-defined networks, the Internet of Things, 5G networks, and multiaccess edge computing. He participated in many international research projects, such as NECOS (H2020), FED4FIRE+OC4 (H2020), WiSHFUL OC2 (H2020), MONROE OC2 (H2020), Dolfin(FP7), UniverSELF (FP7), and Extending Internet into Space (ESA). He served as the General Chair for the WWIC2016 Conference and the INFO-COM SWFAN 2016 Workshop and as the TPC Chair for the INFOCOM SWFAN 2017, E-DTN 2009, and IFIP WWIC 2012 conferences/workshops. He is also a Guest Editor for the Journal *Ad Hoc Networks* (Elsevier).

ARSENIA CHORTI (Member, IEEE) received the M.Eng. degree in electrical and electronic engineering from the University of Patras, Greece, the D.E.A. degree in electronics from the University Pierre et Marie Curie, Paris VI, France, and the Ph.D. degree in electrical engineering from Imperial College London, U.K., in November 2005. She held postdoctoral positions with the University of Southampton, U.K., Technical University of Crete, Greece, and University College London, U.K., from 2005 to 2008. She served as a Senior Lecturer in communications for Middlesex University, U.K., from December 2008 to April 2010. From 2010 to 2013, she was a Marie Curie IOF Researcher with Princeton University, NJ, USA, and with the Institute of Computer Science-FORTH, Greece. From 2013 to 2017, she was a Lecturer with the University of Essex, U.K. She is currently an Associate Professor with ETIS, UMR 8051, University Paris Seine, University Cergy-Pontoise, ENSEA, CNRS, Cergy, France. She is also a Visiting Research Fellow with the University of Essex, U.K. Her work has so far been disseminated in more than 60 journals and international conferences and one book.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LWC.2020.2990563, IEEE Wireless Communications Letters

Performance Analysis of Uplink NOMA-Relevant Strategy Under Statistical Delay QoS Constraints

Mylene Pischella, *Senior Member, IEEE*, Arsenia Chorti, *Member, IEEE* and Inbar Fijalkow, *Senior Member, IEEE*

***Abstract*—A new multiple access (MA) strategy, referred to as non orthogonal multiple access - Relevant (NOMA-R), allows selecting NOMA when this increases *all* individual rates, i.e., it is beneficial for both strong(er) and weak(er) individual users. This letter provides a performance analysis of the NOMA-R strategy in uplink networks with statistical delay constraints. Closed-form expressions of the effective capacity (EC) are provided in two-users networks, showing that the strong user always achieves a higher EC with NOMA-R. Regarding the network's sum EC, there are distinctive gains with NOMA-R, particularly under stringent delay constraints.**

***Index Terms:* NOMA, effective capacity, QoS delay constraint.**

I. Introduction

Due to the strict delay requirements of many emerging ultra reliable low latency communication (URLLC) applications in beyond fifth generation (B5G) networks, the investigation of the interplay between statistical delay quality of service (QoS) constraints and wireless propagation conditions is highly timely. In this context, employing the link layer metric of the effective capacity (EC) [1], [2] – which indicates the maximum achievable rate under a target delay-outage probability threshold – emerges as a natural choice.

In parallel, non-orthogonal multiple access (NOMA) [3] has been consistently shown to achieve higher sum spectral efficiencies when compared to OMA or other schemes [4]. Moreover, NOMA may be required in B5G networks for very large users densities. Up to now, EC analyses in NOMA networks have focused primarily on the downlink [5]–[7]. With respect to the uplink, in [8] it was shown that in two-user networks NOMA is more efficient than OMA at low signal to noise ratios (SNRs), whereas the opposite conclusion holds at large SNRs, due to the interference experienced by the strong user. Adaptive multiple access (MA) strategies could therefore enhance the performance; to the best of our knowledge, [9] is the first attempt to propose an adaptive MA strategy, called NOMA-Relevant (NOMA-R). In NOMA-R, clusters of users employ NOMA only when it is beneficial for all of them in terms of their individual rates.

This letter is the first EC performance analysis of adaptive MA strategies. Its contributions are the following: (i) we evaluate the probability of using NOMA when NOMA-R is employed; (ii) we use the EC as the performance metric[1] , provide new analytic expressions of the EC with NOMA-R and compare them to those derived in the two-users case for NOMA and OMA in [8]; (iii) we prove that NOMA-R is the strategy that maximizes the EC of the strong user, whereas it always outperforms OMA but not NOMA for the weak user; (iv) with respect to this latter aspect, this loss in EC for the weak user becomes negligible under stringent delay constraints; (v) numerical results also show that this conclusion holds for a larger number of users. This letter consequently proves that NOMA-R is a very efficient strategy for delay constrained applications.

Mylene Pischella is with CNAM CEDRIC, France and is a visiting researcher at ETIS, UMR 8051 (contact: mylene.pischella@cnam.fr); Arsenia Chorti and Inbar Fijalkow are with ETIS UMR 8051 CY Cergy Paris Université, ENSEA, CNRS, France.

II. Probability of using NOMA with NOMA-R

A. System model

Let us consider a network with K users employing either OMA, NOMA or NOMA-R in the uplink. The achievable rate of user $k \in \mathcal{S}_K = \{1, ..., K\}$ is denoted in the following by R_k, $\tilde{R}_k$ and $\hat{R}_k$ for NOMA, OMA and NOMA-R, respectively. We assume that the independent and identically distributed (i.i.d.) fading channel coefficients in the links to the base station (BS), denoted by h_k, follow unit variance Rayleigh distributions. The channel gains $x_k = |h_k|^2$ are assumed ordered in decreasing order, so that $x_k \leq x_{k+1} \forall k \in \{1, ..., K-1\}$, and, their distributions can be found by using the theory of order statistics [10]. We denote by $\rho = \frac{1}{N}$ the transmit SNR with N the additive white Gaussian noise power in each link (assumed the same for all links for simplicity). The transmit power used by user k is denoted by P_k so that its received SNR is ρP_k, $k \in \mathcal{S}_K$. Let us assume that a cluster $\mathcal{S}$ of users in $\mathcal{S}_K$ with cardinality $|\mathcal{S}|$ is chosen for NOMA. The achievable rates (in bits/s/Hz) of the kth user in $\mathcal{S}$, assuming perfect successive interference cancellation (SIC) decoding, can be expressed as:

$$R_k = \frac{|\mathcal{S}|}{K} \log_2 \left(1 + \frac{\rho P_k x_k}{1 + \sum_{\substack{j \in \mathcal{S} \\ j<k}} \rho P_j x_j} \right), \forall k \in \mathcal{S}. \quad (1)$$

On the other hand, for the kth user in $\mathcal{S}_K \setminus \mathcal{S}$ the achievable rates with OMA are given as

$$\tilde{R}_k = \frac{1}{K} \log_2 \left(1 + \rho P_k x_k \right), \forall k \in \mathcal{S}_K \setminus \mathcal{S}. \quad (2)$$

The coefficients $|\mathcal{S}|/K$ and $1/K$ account for fair division of the resources between the users employing NOMA and OMA,

[1]We note that [9] focused on proportional fairness.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LWC.2020.2990563, IEEE Wireless Communications Letters

respectively. Although in a standard NOMA network all users will employ NOMA, i.e., $\mathcal{S}$ and $\mathcal{S}_K$ coincide, in NOMA-R, $\mathcal{S}$ is a subset of $\mathcal{S}_K$. The formalization of the NOMA-R criterion stems from the requirement that $\mathcal{S}$ includes all users whose achievable rates are greater with NOMA than with OMA, i.e.,

$$1+\frac{\rho P_k x_k}{1+\sum_{\substack{j\in\mathcal{S}\\ j<k}}\rho P_j x_j} \geq (1+\rho P_k x_k)^{\frac{1}{|\mathcal{S}|}}. \quad (3)$$

Users in $\mathcal{S}_K \setminus \mathcal{S}$ employ OMA. Consequently, the data rate in NOMA-R is $\hat{R}_k = R_k \ \forall k \in \mathcal{S}$ and $\hat{R}_k = \tilde{R}_k \ \forall k \notin \mathcal{S}$. In terms of implementation, the BS identifies the subsets $\mathcal{S}$ using its knowledge of the network's full channel state information (CSI). It then transmits to each user a one-bit feedback indicating whether they should use OMA or NOMA and either the user's index in the OMA subset or the NOMA cluster index if several disjoint NOMA clusters are selected. Therefore, NOMA-R imposes at most one-bit signalling overhead with respect to OMA.

The EC in bits/s/Hz of user $k \in \mathcal{S}_K$ is defined as [1]:

$$E_c^k = -\frac{1}{\theta_k T_f B}\ln\left(\mathbb{E}[e^{-\theta_k T_f B r_k}]\right) = \frac{1}{\beta_k}\log_2\left(\mathbb{E}[e^{\beta_k \ln(2) r_k}]\right)$$

where r_k is the achievable rate of user k (equal to R_k, $\tilde{R}_k$, or $\hat{R}_k$ if the user employs NOMA, OMA or NOMA-R, respectively), T_f is the symbol period and B is the occupied bandwidth. θ_k, known as the QoS exponent [1], is the exponent of the exponential decay of the buffer overflow probability. Under a constant packet arrival rate assumption, the EC is defined as the maximum achievable rate such that a target delay-bound violation probability is met. The more stringent the delay requirement, the larger the delay exponent θ_k. To simplify the notation, we define $\beta_k = -\frac{\theta_k T_f B}{\ln(2)}$ as the negative QoS exponent. Closed form expressions of the EC when OMA and NOMA are employed were derived in [8] for $K = 2$ and are not repeated in the present for compactness.

B. Probability of using NOMA while in NOMA-R for $K \geq 2$

When NOMA-R selects NOMA for $k \in \mathcal{S}$, the sum rate can easily be shown to be equal to $\frac{|\mathcal{S}|}{K}\log_2\left(1+\sum_{k\in\mathcal{S}}\rho P_k x_k\right)$. Consequently, the subset $\mathcal{S}$ of $\mathcal{S}_K$ that maximizes the sum rate while satisfying (3) is selected by NOMA-R. Several disjoint clusters may also be selected if they independently verify (3).

The probability of using NOMA when NOMA-R is employed, denoted by τ_K in the following, is the union of the probabilities to verify (3) for any subset $\mathcal{S} \subseteq \mathcal{S}_K$ with $|\mathcal{S}| \geq 2$ and its analytical derivation is very evolved. As an illustrative example, let us assume $\mathcal{S} = \{1, ..., k\}$ with $k \leq K$. Let $y_k = \rho P_k x_k$ be the weighted kth order statistics and $z_k = \sum_{j=1}^{k-1}\rho P_j x_j$ the weighted sum of the lowest $(k-1)$th order statistics. Then τ_k is equal to:

$$\tau_k = Pr\left(\bigcap_{i=2:k}\left(\frac{(1+y_i)-(1+y_i)^{\frac{1}{k}}}{(1+y_i)^{\frac{1}{k}}-1} \geq z_i\right)\right). \quad (4)$$

For the specific case $k = K$, the joint probability density function (pdf) of (y_K, z_K) can be derived by using the moment generating function (MGF) of the weighted sum of the lowest order statistics, denoted as $\mathcal{M}_{z_K}$, as in [10]. The pdf of z_k can be obtained by using the following properties: $\mathcal{M}_{z_K} = \prod_{j=1}^{K-1}\mathcal{M}_{x_j}(\rho P_j x_j)$ and $\mathcal{L}^{-1}\left(\mathcal{M}_{x_j}(\rho P_j x_j)\right)(t) = \frac{1}{\rho P_j} f_{x_j}\left(\frac{t}{\rho P_j}\right)$, where $\mathcal{L}^{-1}$ is the inverse Lagrange transform. Consequently, the joint pdf $f_{z_K, y_K=\bar{y}}$ can be derived from that of z_k and y_K in [10, eq. (3.41)]. However, a closed-form expression of $Pr\left(\frac{(1+y_K)-(1+y_K)^{\frac{1}{K}}}{(1+y_K)^{\frac{1}{K}}-1} \geq z_K\right)$ cannot be obtained, and similarly to the conclusion in [11, Section V.D], it should be calculated with a mathematical software. Moreover when $k < K$, to the best of our knowledge, the joint pdf of (y_K, z_K) is yet unknown. Consequently, when $K > 2$, (4) cannot be evaluated analytically with reasonable effort. For all these reasons, our analytical study is limited to $K = 2$ while we provide numerical results for $K > 2$.

Finally, examining the case non i.i.d. channel coefficients, we note that the case of non-identical exponential distributions can be treated as in [12], while the case of non independent coefficients as in [13]. If full CSI is not available at the BS, CSI uncertainties can be inserted in users' distributions to derive an MA selection strategy [13], and, when $K = 2$, (3) can be formulated as a binary hypothesis testing problem [14].

C. Probability of using NOMA while in NOMA-R for $K = 2$

In the following, we consider the two-users case and call user 2 the strong user, and user 1 the weak user. As $R_1 \geq \tilde{R}_1$ is always fulfilled, the NOMA-R strategy is used whenever $R_2 \geq \tilde{R}_2$. The NOMA-R condition consequently simplifies to $x_2 \geq \frac{\rho^2 x_1^2 P_1^2 - 1}{\rho P_2}$ and $\tau_2 = \tau(\rho) = Pr\left(x_2 \geq \frac{\rho^2 x_1^2 P_1^2 - 1}{\rho P_2}\right)$. Using the theory of order statistics, the pdf of x_1 is $2e^{-2x_1}$, the pdf of x_2 is $2e^{-x_2}(1-e^{-x_2})$ and the joint pdf of (x_1, x_2) is $2e^{-x_1}e^{-x_2}$. Then $\tau(\rho)$ is equal to:

$$\begin{aligned}\tau &= \int_{x_1=0}^{\frac{P_2+\sqrt{P_2^2+4P_1^2}}{2\rho P_1^2}}\int_{x_2=x_1}^{+\infty} 2e^{-x_1}e^{-x_2}dx_2dx_1 \\ &+ \int_{x_1=\frac{P_2+\sqrt{P_2^2+4P_1^2}}{2\rho P_1^2}}^{+\infty}\int_{x_2=\frac{\rho^2 x_1^2 P_1^2-1}{\rho P_2}}^{+\infty} 2e^{-x_1}e^{-x_2}dx_2dx_1 \\ &= f(\rho) + g(\rho) \end{aligned} \quad (5)$$

where

$$f(\rho) = 1 - e^{-\frac{P_2+\sqrt{P_2^2+4P_1^2}}{\rho P_1^2}} \quad (6)$$

$$g(\rho) = \frac{\sqrt{\pi}e^{\frac{4P_1^2+P_2^2}{4\rho P_2 P_1^2}}\left(1-\mathrm{erf}\left(\frac{2P_2+\sqrt{P_2^2+4P_1^2}}{2\sqrt{P_2\rho P_1}}\right)\right)\sqrt{P_2}}{P_1\sqrt{\rho}} \quad (7)$$

and $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$. The boundary in both integrals in (5) is due to the fact that x_2 should always be such that $x_2 \geq x_1$, but $\frac{\rho^2 x_1^2 P_1^2 - 1}{\rho P_2}$ is lower than x_1 if $x_1 \leq \frac{P_2+\sqrt{P_2^2+4P_1^2}}{2\rho P_1^2}$. $\tau(\rho)$ is a monotonically decreasing function with respect to ρ. (5) is validated with Monte-Carlo simulations, shown in Fig. 1, assuming that $P_1 + P_2 = 1$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LWC.2020.2990563, IEEE Wireless Communications Letters

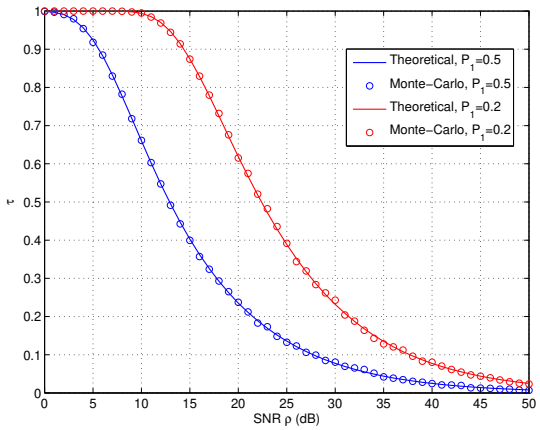


Fig. 1. Validation of the closed-form expression of τ

Lemma 1. *$\tau(\rho)$ tends to 1 when ρ tends to 0 and $\tau(\rho)$ tends to 0 when $\rho >> 1$.*

Proof. When $\rho \to 0$, $f(\rho) \to 1$. Moreover, $\text{erf}(x) \approx 1 - e^{-x^2}/(x\sqrt{\pi})$ when $x >> 1$, which implies that $g(\rho) \approx ae^{-b/\rho}$ with (a, b) two strictly positive constants. Therefore, $g(\rho) \to 0$ when $\rho \to 0$. Furthermore, when $\rho >> 1$, $f(\rho) \to 0$. As $\text{erf}(x) \approx \frac{2}{\sqrt{\pi}} x e^{-x^2}$ when $x \to 0$, $g(\rho) \approx a_1 \frac{e^{a_2/\rho}}{\rho} + b_1 \frac{e^{b_2/\rho}}{\rho}$ where (a_1, a_2, b_1, b_2) are constants and $g(\rho) \to 0$ □

III. NOMA-R Effective Rates

The ECs of user $k = 1, 2$ when employing NOMA, OMA or NOMA-R are denoted by $E^k_{c,N}$, $E^k_{c,O}$ and $E^k_{c,R}$, respectively.

A. NOMA-R EC of user 1

We hereafter derive closed-form expressions of the EC with NOMA-R. As τ corresponds to the proportion of time spent in NOMA and $(1-\tau)$ is the proportion of time spent in OMA, the EC of user $k \in \{1, 2\}$ when the NOMA-R strategy is employed is given as:

$$E^k_{c,R} = \frac{1}{\beta_k} \log_2 \left(\mathbb{E} \left[e^{\beta_k \tau R_k + \beta_k (1-\tau) \tilde{R}_k} \right] \right). \quad (8)$$

For user 1, the EC achieved with the NOMA-R strategy is:

$$\begin{aligned} E^1_{c,R} &= \frac{1}{\beta_1} \log_2 \left(\mathbb{E}[(1+\rho P_1 x_1)^{\frac{\beta_1(\tau+1)}{2}}] \right) \\ &= \frac{1}{\beta_1} \log_2 \left(\frac{2}{\rho P_1} U \left(1, 2 + \frac{\beta_1(\tau+1)}{2}, \frac{2}{\rho P_1} \right) \right) \end{aligned} \quad (9)$$

where $U(a, b, z) = \frac{1}{\Gamma(a)} \int_0^\infty e^{-zt} t^{a-1} (1+t)^{(b-a-1)} dt$ denotes the confluent hypergeometric function.

Lemma 2. *$E^1_{c,R}$ is monotonically increasing with ρ.*

Proof. Let us consider ρ_1 and ρ_2 such that $\rho_1 < \rho_2$. For any value of β_1, let us define: $\beta_{1,a} = \frac{\beta_1(1+\tau(\rho_1))}{2}$ and $\beta_{1,b} = \frac{\beta_1(1+\tau(\rho_2))}{2}$. Then from [8, eq.(11)] and (9), $E^1_{c,R}(\beta_1, \rho_1) = E^1_{c,N}(\beta_{1,a}, \rho_1)$ and $E^1_{c,R}(\beta_1, \rho_2) = E^1_{c,N}(\beta_{1,b}, \rho_2)$. As $\tau(\rho)$ is a decreasing function with respect to ρ, and β are negative, $\beta_{1,a} \leq \beta_{1,b}$. Moreover, $E^1_{c,R}(\beta, \rho)$ is increasing both with respect to β and to ρ according to [8]. Consequently, $E^1_{c,N}(\beta_{1,a}, \rho_1) \leq E^1_{c,N}(\beta_{1,a}, \rho_2) \leq E^1_{c,N}(\beta_{1,b}, \rho_2)$ and:

$$E^1_{c,R}(\beta_1, \rho_1) \leq E^1_{c,R}(\beta_1, \rho_2) \quad \forall \rho_1 < \rho_2 \quad (10)$$

□

B. NOMA-R EC of user 2

For user 2, the NOMA-R EC is given by

$$E^2_{c,R} = \frac{1}{\beta_2} \log_2 \left(\mathbb{E} \left[\left(1 + \frac{\rho P_2 x_2}{1 + \rho P_1 x_1} \right)^{\beta_2 \tau} (1 + \rho P_2 x_2)^{\frac{\beta_2(1-\tau)}{2}} \right] \right) \quad (11)$$

Lemma 3. *When ρ tends to 0, the EC with the NOMA-R strategy becomes equivalent to that of NOMA given in [8].*

Proof. When $\rho \to 0$, $\tau(\rho) \to 1$ and therefore $(1 + \rho P_2 x_2)^{\frac{\beta_2(1-\tau)}{2}} \to 1$. Consequently, $E^2_{c,R}$ tends to $E^2_{c,N}$. □

Lemma 4. *When $\rho >> 1$, the EC with the NOMA-R strategy becomes equivalent to that of OMA and its closed-form expression is given by:*

$$E^2_{c,R} \approx \frac{1}{\beta_2} \log_2 \left(\Gamma \left(\frac{\beta}{2} + 1 \right) (\rho P_2)^{\frac{\beta}{2}} (2 - 2^{-\frac{\beta}{2}}) \right). \quad (12)$$

Proof. When $\rho >> 1$, $\left(1 + \frac{\rho P_2 x_2}{1+\rho P_1 x_1}\right)^{\beta_2 \tau} \to 1$ because $\tau \to 0$. Then using $(1+x)^\alpha \approx x^\alpha$, the EC of user 2 becomes:

$$\begin{aligned} E^2_{c,R} &\approx \frac{1}{\beta_2} \log_2 \left(\mathbb{E} \left[(\rho P_2 x_2)^{\frac{\beta_2}{2}} \right] \right) \\ &\approx \frac{1}{\beta_2} \log_2 \left(\int_0^\infty 2 (\rho P_2 x_2)^{\frac{\beta_2}{2}} e^{-x_2} (1 - e^{-x_2}) dx_2 \right). \end{aligned}$$

The integral's closed-form expression leads to (12). □

Theorem 1. *The EC of user 1 is always larger with NOMA than with NOMA-R, while OMA is the worst strategy in terms of EC. Moreover, the EC of user 2 is always larger with the NOMA-R strategy than with NOMA or OMA.*

Proof. The NOMA-R instantaneous rate of user 1 is equal to $\hat{R}_1 = \frac{(1+\tau)}{2} \log_2(1 + \rho P_1 x_1)$, according to (9). Therefore $\tilde{R}_1 \leq \hat{R}_1 \leq R_1$ Then as β_1 is negative, $e^{\beta_1 R_1} \leq e^{\beta_1 \hat{R}_1} \leq e^{\beta_1 \tilde{R}_1}$, and $\mathbb{E}\left[e^{\beta_1 R_1}\right] \leq \mathbb{E}\left[e^{\beta_1 \hat{R}_1}\right] \leq \mathbb{E}\left[e^{\beta_1 \tilde{R}_1}\right]$, so that

$$E^1_{c,N} \geq E^1_{c,R} \geq E^1_{c,O} \quad (13)$$

The NOMA-R rate of user 2 is $\hat{R}_2 = \max\{R_2, \tilde{R}_2\}$. Following the same steps as for user 1, we conclude that

$$E^2_{c,R} \geq E^2_{c,N} \text{ and } E^2_{c,R} \geq E^2_{c,O} \quad (14)$$

□

Remark: $E^2_{c,R}$ asymptotically tends to either $E^2_{c,N}$ or $E^2_{c,O}$, both of which are monotonically increasing with ρ. Moreover, contrary to the NOMA strategy, the NOMA-R strategy does not lead to a saturation of the EC of the strong user because of (14) and because $E^2_{c,O}$ increases without bound with ρ [8].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LWC.2020.2990563, IEEE Wireless Communications Letters

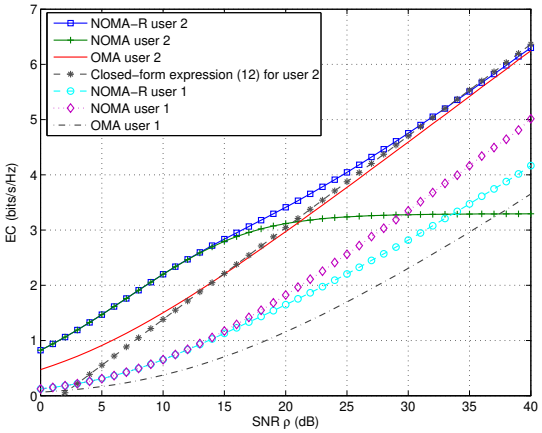


Fig. 2. EC per user versus SNR ρ when $K = 2$

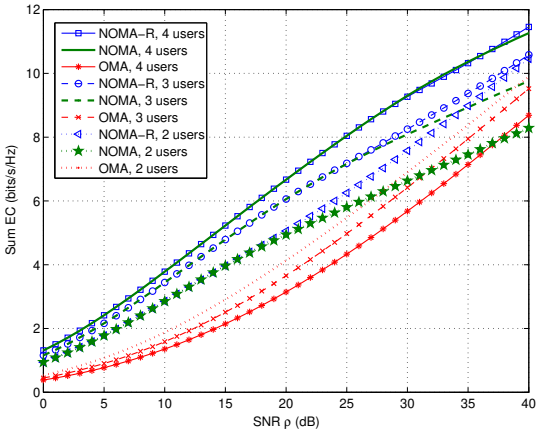


Fig. 3. Sum EC versus SNR ρ

IV. Numerical Results

Performance comparisons with respect to EC with the NOMA-R strategy are provided for $P = [0.2, 0.8]$, $P = [0.05, 0.15, 0.8]$ and $P = [0.01, 0.04, 0.15, 0.8]$ when $K = 2, 3$ and 4, respectively. Unless otherwise stated, $\beta_k = -2$, $\forall k$. Fig. 2 validates the results of Theorem 1. Fig. 3 shows that the largest sum EC is always achieved with NOMA-R whatever the value of K. Moreover, the sum EC with NOMA-R coincides with that of NOMA when the SNR is lower than a minimum value $\rho_{\min}$ that increases with K, because constraint (3) becomes less stringent when K increases. Fig. 4 shows the dependency of the EC on to β_1 when the SNR is equal to 35 dB, $\beta_K = -2$ and $\beta_j = \beta_1$, $\forall j < K$. The sum EC is larger with NOMA-R, except when β_1 approaches 0. The NOMA-R strategy is consequently more favorable when the target delay-bound violation probabilities are more stringent, especially for weak users.

V. Conclusions

In this letter the EC performance of an adaptive MA strategy, NOMA-R, was studied both analytically for $K = 2$ users and numerically for larger values of K. It was shown that

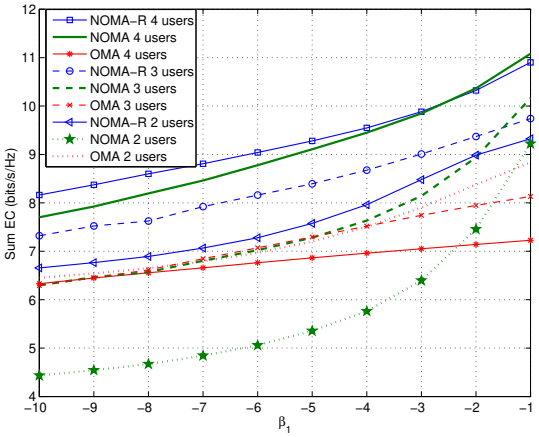


Fig. 4. Sum EC versus β_1

NOMA-R is an advantageous strategy for delay constrained applications in B5G, e.g., URLLC, particularly as the users' delay-outage probability constraints become more stringent.

References

[1] D. Wu and R. Negi, "Effective Capacity: a Wireless Link Model for Support of Quality of Service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, July 2003.

[2] M. Amjad, L. Musavian, and M. H. Rehmani, "Effective Capacity in Wireless Networks: A Comprehensive Survey," *IEEE Commun. Surveys Tuts.*, pp. 1–33 (early access), July 2019.

[3] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A Survey of Non-Orthogonal Multiple Access for 5G," *IEEE Commun. Surv. Tut.*, vol. 20, no. 3, pp. 2294–2323, thirdquarter 2018.

[4] Y. Kanaras, A. Chorti, M. Rodrigues, and I. Darwazeh, "An optimum detection for a spectrally efficient non orthogonal FDM system," in *Proc. 13th Int. OFDM WS*, Aug 2008, pp. 65–68.

[5] G. Liu, Z. Ma, X. Chen, Z. Ding, F. R. Yu, and P. Fan, "Cross-Layer Power Allocation in Non-Orthogonal Multiple Access Systems for Statistical QoS Provisioning," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 11388–11393, Dec 2017.

[6] W. Yu, L. Musavian, and Q. Ni, "Link-Layer Capacity of NOMA Under Statistical Delay QoS Guarantees," *IEEE Trans. Commun.*, vol. 66, no. 10, pp. 4907–4922, Oct 2018.

[7] C. Xiao, J. Zeng, W. Ni, R. P. Liu, X. Su, and J. Wang, "Delay Guarantee and Effective Capacity of Downlink NOMA Fading Channels," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 508–523, June 2019.

[8] B. Mouktar, W. Yu, A. Chorti, and L. Musavian, "Performance Analysis of NOMA Uplink Networks under Statistical QoS Delay Constraints," in *IEEE International Conference on Communications*, 2020, pp. 1–7.

[9] M. Pischella and D. Le Ruyet, "NOMA-Relevant Clustering and Resource Allocation for Proportional Fair Uplink Communications," *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 873–876, June 2019.

[10] H.-C. Yang and M.-S. Alouini, *Order Statistics in Wireless Communications: Diversity, Adaptation, and Scheduling in MIMO and OFDM Systems*, Cambridge University Press, 2011.

[11] Y. Ko, H. Yang, S. Eom, and M. Alouini, "Adaptive Modulation with Diversity Combining Based on Output-Threshold MRC," *IEEE Trans. Wir. Commun.*, vol. 6, no. 10, pp. 3728–3737, October 2007.

[12] N. Balakrishnan, "Order statistics from non-identical exponential random variables and some applications," *Commun. Stat. - Theor. Meth.*, vol. 18, no. 2, pp. 203–253., 1994.

[13] A. Alsawah and I. Fijalkow, "Practical radio link resource allocation for fair QoS-provision on OFDMA downlink with partial channel-state information," *EURASIP J. Adv. Signal Proc.*, pp. 1–16, 2009.

[14] S. W. H. Shah, M. M. U. Rahman, A. N. Mian, A. Imran, S. Mumtaz, and O. A. Dobre, "On the impact of mode selection on effective capacity of device-to-device communication," *IEEE Wireless Communications Letters*, vol. 8, no. 3, pp. 945–948, 2019.