Collecte de traces WiFi publiques: de la protection de la vie privée à l'analyse de trajectoires

Fernando Molano Ortiz, Abhishek Kumar Mishra, F. D. de M. Silva, Nadjib Achir, Aline Carneiro Viana, Anne Fladenmuller, L. H. M. K. Costa

Collecte de traces WiFi publiques: de la protection de la vie privée à l'analyse de trajectoires[†]

F. M. Ortiz[1], A. Mishra[1], F. D. de M. Silva[4], N. Achir[1,2], A. C. Viana[1], A. Fladenmuller[3], et L. H. M. K. Costa[4]

[1]*INRIA,* [2]*Université Sorbonne Paris Nord,* [3]*Sorbonne Université, France.* [4]*Universidade Federal do Rio de Janeiro, Brésil.*

Dans le paysage actuel marqué par l'omniprésence des smartphones et des réseaux sans fil, la génération d'empreintes numériques est devenue courante, révélant les habitudes quotidiennes des utilisateurs. Cet article présente un ensemble d'outils pour collecter et analyser des traces WiFi. Ces outils relèvent différents défis, tels que la gestion des associations des adresses MAC des smartphones, la reconstruction de trajectoires des utilisateurs et la protection de leur confidentialité. En abordant systématiquement ces défis, ces outils visent à faciliter la compréhension de la mobilité des individus et à établir des contacts plausibles entre divers appareils.

Mots-clefs : Communication sans fil, Outil de collecte, Mesure passive expérimentale, Mobilité humaine

1 Introduction

Un volume significatif de trafic réseau est engendré par les appareils équipés d'interfaces WiFi, tels que les smartphones, les montres intelligentes et autres. Ces appareils envoient des trames de gestion pour signaler leur présence aux points d'accès (AP) environnants, dans le cadre du processus de découverte des réseaux [Boa21]. Par conséquent, ces trames peuvent constituer une source utile pour la communauté de recherche afin de modéliser la mobilité humaine et en analyser les déplacements et interactions. Toutefois, l'analyse des trames de gestion pose plusieurs défis. Tout d'abord un défi technique lié à la mise en place du processus de collecte. Le deuxième défi concerne l'anonymisation des données collectées du fait que les trames contiennent des informations privées telles que les adresses MAC. Le troisième défi consiste à identifier un équipement même s'il recourt à un mécanisme de randomisation des adresses MAC. Le dernier défi est de pouvoir reconstruire la trajectoire d'un équipement. Pour combler ces challenges, nous présentons dans cet article une **approche expérimentale**, incluant l'infrastructure physique et les outils pour la collecte passive de trames de gestion WiFi. Notre approche utilise des **mesures passives non intrusives** pour déduire la mobilité des utilisateurs et leurs interactions potentielles lors de leurs déplacements.

2 Architecture

La Figure 1 illustre l'architecture complète du système, comprenant : (*i*) un outil de collecte des trames de gestion WiFi (*probe-requests*) d'appareils individuels aléatoires détectés dans une zone de couverture définie ; (*ii*) un outil de traitement a la volée des trames collectées pour la protection de l'anonymat des utilisateurs ; (*iii*) un outil d'association des adresses MAC randomisées, compte tenu de l'utilisation par certains appareils mobiles de stratégies de randomisation MAC ; (*iv*) et finalement, outils de reconstruction de trajectoire des appareils mobiles.

Collecte : Cet outil nécessite le déploiement d'un matériel spécifique appelé *super-sniffers*, une stratégie pour améliorer la qualité du système de mesure passif utilisant des *sniffers* redondants. Un *super-sniffer* est composé de cinq *sniffers* (basé sur l'analyse de redondance dans [SFD22]), comme illustré dans la Figure 2. Chaque *sniffer* est un Raspberry Pi disposant d'une antenne WiFi configurée en mode moniteur

[†]Ce travail a été partiellement financé par le projet ANR MITIK, Agence Nationale de la Recherche (ANR), PRC AAPG2019.

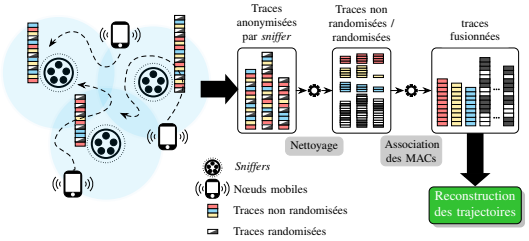
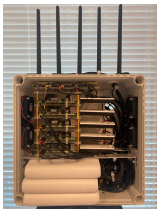


Fig. 1: Collecte passive des trames de gestion WiFi.



(a)

Équipement	Modèle
Backbone	Raspberry Pi 4B 8 GB RAM
Module WiFi	TPLink WN722-N
Antenne	2.4 GHz, 6.5 dBi
Récepteur GPS	U-Blox NEO-7M
Batterie	Belkin 20k
Carte SD	32 GB

(b)

Fig. 2: Dispositifs utilisés pour le processus de sniffing.

pour la capture de données sur un canal prédéfini ou plusieurs canaux en utilisant un *round-robin*, et un GPS pour maintenir une synchronisation temporelle en plus de la localisation physique du *sniffer*.

Un outil de gestion basé sur `Ansible` est utilisé pour configurer les *sniffers* avec les paramètres requis pour gérer le système de configuration, de manière synchrone, depuis un ordinateur distant. Chaque *sniffing* utilise l'outil `Scapy`, une bibliothèque open source basée sur `Python`, pour la manipulation de trames collectées. Une fois que les *sniffers* ont fini de capturer les données, ces traces sont stockées dans des fichiers `.pcap`, et envoyées d'une manière sécurisée à un serveur distant pour une analyse *offline*.

Anonymisation des données : Conformément au Règlement Général sur la Protection des Données (RGPD), pour éviter de stocker les identifiants annoncés des appareils et empêcher de potentiels attaquants sur des données privées et sensibles, une stratégie d'anonymisation des données est proposée [dMSMV+22]. Pour chaque trame *probe-request* captée par le *sniffer*, une fonction de hachage (MD5 ou SHA256) est appliquée dans la RAM et avant son enregistrement dans la ROM. Cette adresse MAC hachée est par la suite tronquée à la même longueur que l'original (48 bits), avant d'être enregistrée sur le *sniffer*. Le même processus est également appliqué à toute information considérée comme privée et incluse dans la trame, comme par exemple des SSIDs. Ce processus apporte une anonymisation complète et difficilement réversible. Il est important de noter que le problème de collision lié à la fonction de hachage est toujours possible, mais très limité du fait du nombre faible d'échantillons (adresses MAC) collectés dans l'espace et le temps.

Association des adresses MAC : Pour préserver la confidentialité des utilisateurs, la norme WiFi recommande aux appareils mobiles de modifier périodiquement leur véritable adresse MAC physique, procédé appelé randomisation. Ce mécanisme est utilisé pour empêcher le suivi et l'identification des appareils dans leurs interactions avec les réseaux sans fil.

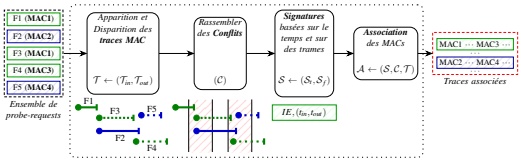


Fig. 3: Structure du cadre `Bleach`.

Cette randomisation des adresses MAC empêche toute reconstruction de trajectoire des appareils mobiles. Par conséquent, un outil d'association d'adresses MAC WiFi doit être utilisé pour permettre une corrélation de plusieurs adresses MAC aléatoires émises par le même appareil [MCRV16]. Dans notre architecture, nous utilisons une solution, nommée `Bleach`, proposée dans [Mis23] (Figure 3). Cette solution prend en entrée une trace de *probe-requests* avec des adresses MAC aléatoires et génère un dictionnaire d'adresses aléatoires associées à des périphériques particuliers. `Bleach` divise, dans un premier temps, la trace collectée en un ensemble de traces de séquences de trames *probe-request* utilisant la même adresse MAC. Nous appelons chaque ensemble un *trail*, ce qui permet de réduire le problème de l'association MAC à celui d'associer correctement la disparition d'un *trail* avec la réapparition d'un autre *trail*.

Par la suite, les *trails* sont séparés en sous-ensembles disjoints pour identifier les conflits. Un **conflit** ($\mathcal{C}$) fait référence à l'ensemble des nouvelles *trails* observées, dans une période de temps ($T_c^{\tau_i}$), à partir de la fin d'un *trail* avec une adresse MAC spécifique. Une bonne valeur de la période de conflit nous permet de considérer toutes les associations potentielles. $T_c^{\tau_i}$ peut être facilement identifié par mesure. Une fois l'ensemble des conflits identifié, `Bleach` extrait des **signatures** ($\mathcal{S}$) pour chaque *trail*. Deux types de signatures sont évalués : *i.*) une signature basée sur le temps ($\mathcal{S}_t$) en analysant le comportement temporel des trames *probe-requests* reçues. *ii.*) une signature basée sur des éléments d'information ($\mathcal{S}_f$) contenue dans le champ de contrôle des trames. L'objectif de ces deux signatures est de distinguer un appareil du reste de la population. Ces deux signatures sont par la suite utilisées dans un algorithme d'association MAC capable d'associer, les pairs de *trail* ayant les signatures les plus proches.

Reconstruction des trajectoires : La reconstruction de la trajectoire des appareils mobiles à partir de traces collectées est un défi important du fait que la localisation de ces équipements dépend fortement des caractéristiques du signal reçu par les *sniffers*. Bien que le *Received Signal Strength Indicator* (RSSI) est la métrique la plus largement utilisée pour l'inférence de localisation d'un équipement, elle souffre d'inexactitudes et d'erreurs provoquées par des facteurs environnementaux aléatoires. Pour contourner ces problèmes, nous utilisons dans notre architecture de collecte un outil appelé `Allies` [Mis23]. `Allies` introduit le concept de trajectoire bornée. Une trajectoire bornée désigne une zone où un utilisateur particulier est probablement présent dans le temps. Pour déduire cette borne dans la trajectoire, `Allies` propose une nouvelle caractérisation des erreurs dans l'estimation de la distance radiale (r) basée sur le RSSI observé par un *sniffer* (i) pour un utilisateur (j).

`Allies` propose trois étapes : La première étape consiste à obtenir **l'ensemble des observations** de différents *sniffers* pour un utilisateur particulier dans un petit intervalle de temps. Chaque entrée dans un ensemble est un *tuple* contenant un horodatage, la valeur du RSSI observé et l'identifiant du *sniffer* ayant effectué cette observation (*timestamp, RSSI, sniffer_id*) pour toutes les trames *probe-requests*.

La seconde étape consiste à estimer les **erreurs de distance radiale** ($E_{r_{ij}}$) provenant de l'ensemble des observations. $E_{r_{ij}}$ comporte deux composantes. La première, *l'erreur de distance* (E_{span}), modélise le comportement de l'erreur due à la distance qui sépare le *sniffer* de l'équipement. La deuxième, *l'erreur d'environnement* (E_{env}), capture les fluctuations observées sur cette erreur de distance due à l'environnement et a son impact sur la variabilité du RSSI.

La dernière étape de `Allies` consiste à estimer, pour chaque intervalle de temps, plusieurs localisations possibles en effectuant plusieurs échantillonnages des distributions des erreurs estimées dans l'étape précédente. L'idée est d'arriver à estimer un nuage de positions possible pour chaque équipement afin d'en déduire une localisation bornée. La concaténation de cet ensemble de localisations bornées de chaque utilisateur pour chaque intervalle de temps donne lieu à la trajectoire bornée de l'équipement.

3 Résultats

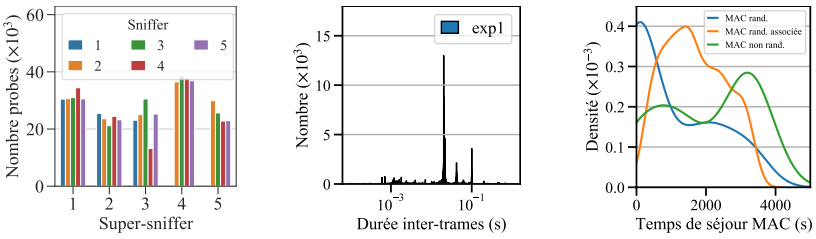
Plusieurs expériences de 60 minutes chacune ont été réalisées lors de campagnes de mesures à l'Université de La Rochelle. Chaque *sniffer* capture des données sur le canal 1 de la bande 2.4 GHz. Ci-dessous, nous analysons les résultats obtenus pour une des expériences. Lors de cette expérience, nous avons déployé 5 *super-sniffers* au sein du campus de l'université. La Figure 4a donne le nombre de *probe-requests* capturés par chaque *super-sniffer*. Ce nombre varie de 100 k à 160 k.

Nous avons utilisé, dans un premier temps, les traces collectées grâce à nos *super-sniffer* par l'outil d'association des adresses MAC. L'un des paramètres les plus importants pour l'outil d'association des adresses MAC est la durée qui caractérise la période de conflit pour un *trail* ($T_c^{\tau_i}$). Un *trail* est une succession de trames *probe-requests* avec la même adresse MAC.

Dans la Figure 4b, nous représentons la durée qui sépare deux *probe-requests* consécutif ayant la même adresse MAC. Deux pics distincts peuvent être observés. Le premier pic de l'ordre de la dizaine de millisecondes représente le temps qui sépare deux *probe-requests* appartenant à la même rafale et par conséquent au même *trail*. Le second pic, de l'ordre de la seconde, représente le temps qui sépare deux *probe-requests* appartenant à deux rafales différentes, mais ayant la même adresse MAC. Ces deux valeurs ont été utilisées pour identifier les *trails* et l'ensemble des conflits.

Une fois les *trails* identifiés, nous représentons la distribution des temps de séjour des adresses MAC statiques (non-randomisé) ainsi que des adresses MAC randomisé, après association (Figure 4c). En effet, certains anciens terminaux ne disposent pas encore du mécanisme de randomisation des adresses. Les MAC randomisés ont des temps de séjour inférieurs aux deux autres, en raison de la fréquence de changement d'adresse MAC. Les temps de séjour des appareils après association et ceux des MACs non randomisés sont similaires dans leurs distributions. Cependant, nous supposons qu'il existe dans la zone de collecte des dispositifs statiques qui génèrent un volume de *probe-requests* plus élevé que ceux qui sont en mouvement.

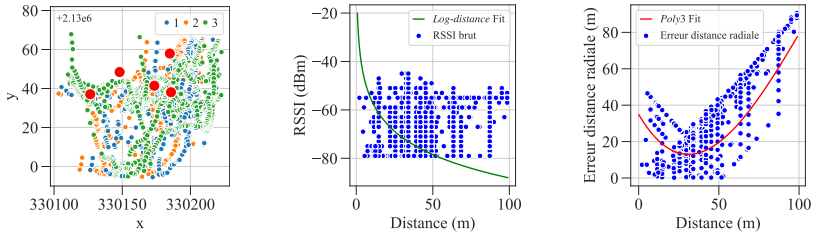
Une fois les adresses MAC associées, nous passons à la reconstruction des trajectoires. Pour cela, nous introduisons trois appareils avec MAC non randomisés, chacun configuré dotés d'un GPS, pour générer des traces réelles avec des informations de localisation. La Figure 5a montre le positionnement des *sniffers* et les positions des trois appareils en mouvement.



(a) (b) (c)

Fig. 4: Résultats de l'association MAC.

À partir de traces obtenues, nous générons les ensembles d'observations. Ils sont par la suite utilisés pour effectuer une caractérisation des erreurs dans l'estimation de la distance. La Figure 5b représente la dégradation du signal observé au niveau d'un *super-sniffer*. Nous observons que la qualité du signal (cf. points bleus) se dégrade en fonction de la distance radiale à partir de 50 m. La courbe verte représente l'ajustement de la qualité du signal en utilisant les valeurs de RSSI obtenue pour optimiser les paramètres du modèle *path-loss*. La dégradation de la qualité du signal entraîne également une augmentation de l'erreur d'estimation moyenne. D'autre part, l'erreur d'estimation (cf. ligne rouge obtenue en effectuant un ajustement polynomial d'ordre 3 ainsi que ses fluctuations (cf. valeurs d'erreurs de distance radiales associées aux points bleus)) augmentent avec l'augmentation de la distance radiale (Figure 5c). Les fluctuations sont dues à l'influence des conditions de l'environnement (ex. mouvement de véhicules ou des piétons).



(a) Localisation i, j. (b) Estimation du *path-loss* (c) $E_{r_{ij}} = E_{span} + E_{env}$

Fig. 5: Résultats de la reconstruction de trajectoires.

4 Conclusion

L'objectif de ce travail était de concevoir et de développer une architecture de collecte permettant d'analyser la mobilité des équipements dans les zones urbaines à travers une stratégie passive. Cette architecture inclut un premier outil d'anonymisation des données conformément aux recommandations RGPD. Un outil de d'association des adresses MAC du fait du mécanisme de randomisation. Les résultats obtenus montrent des temps de séjour des adresses MAC associées semblables adresses MAC réelles. Finalement, bien que nous ayons obtenu des échantillons très dispersés à partir d'équipements utilisés pour générer des traces réelles avec des informations de localisation des utilisateurs, le dernier outil consacré à la reconstruction des trajectoires montre une tendance cohérente de l'évolution de l'erreur moyenne par rapport à la distance radiale. La suite de notre travail consistera à classifier l'erreur d'estimation de distance pour obtenir des trajectoires bornées et à évaluer la qualité des bornes inférées.

References

[Boa21] IEEE Std. BOARD : Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. *IEEE Std 802.11*, 2021.

[dMSMV+22] F. D. de MELLO SILVA, A. K. MISHRA, A. C. VIANA, N. ACHIR, A. FLADENMULLER et L. H. M. K. COSTA : Performance analysis of a privacy-preserving frame sniffer on a Raspberry Pi. *Dans IEEE CSNet*, 2022.

[MCRV16] C. MATTE, M. CUNCHE, F. ROUSSEAU et M. VANHOEF : Defeating MAC address randomization through timing attacks. *Dans WiSec*, 2016.

[Mis23] A. K. MISHRA : *Revealing and exploiting privacy vulnerabilities in users' public wireless packets*. Thèse de doctorat, Institut Polytechnique de Paris, 2023.

[SFD22] M. I. SYED, A. FLADENMULLER et M. DIAS DE AMORIM : How much can sniffer redundancy improve Wi-Fi traffic? *Dans IEEE VTC*, 2022.