Aggregated monitoring of large-scale network systems and control of epidemics

Muhammad Umar B. Niazi

UNIVERSITÉ GRENOBLE ALPES

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE ALPES

Spécialité : **Automatique - Productique**

Arrêtée ministériel : 25 mai 2016

Présentée par

Muhammad Umar B. NIAZI

Thèse dirigée par **Carlos CANUDAS-DE-WIT**
et codirigée par **Alain KIBANGOU**

préparée au sein du **Laboratoire Grenoble Images Parole Signal Automatique (GIPSA)**
dans **l'École Doctorale Eléctronique Eléctrotechnique Automatique & Traitement du signal (EEATS)**

Aggregated Monitoring of Large-scale Network Systems and Control of Epidemics

Thèse soutenue publiquement le **12 juillet 2021**,
devant le jury composé de :

Didier GEORGES
Professeur des universités, Grenoble-INP, Président

Denis EFIMOV
Directeur de recherche, Inria Lille, Rapporteur

Jun-ichi IMURA
Professeur des universités, Tokyo Institute of Technology, Rapporteur

Jacquelien M. A. SCHERPEN
Professeur des universités, University of Groningen, Examinatrice

Karl Henrik JOHANSSON
Professeur des universités, KTH Royal Institute of Technology, Examinateur

Pierre-Alexandre BLIMAN
Directeur de recherche, Inria Paris, Invité

Abstract

This Ph.D. thesis is done mainly in the context of the European Research Council's (ERC) Advanced Grant project Scale-FreeBack and partially in the context of the Inria's COVID-19 Mission project Healthy-Mobility. The Scale-FreeBack project aims to develop a holistic, scale-free control approach to complex systems and to set new foundations for a theory dealing with complex physical networks with arbitrary dimensions. On the other hand, motivated by the onset of the COVID-19 pandemic, the Healthy-Mobility project aims to develop optimal control strategies for testing and urban human mobility to limit the epidemic spread. In relation to both projects, the contributions of the thesis are respectively divided into two parts.

In the first part of the thesis, we develop a theory for monitoring large-scale clustered network systems with limited computational and sensing equipment through a projected network system, which is of tractable dimension and is obtained through the aggregation of clusters of a network system. We propose a minimum-order average observer and provide its design criteria. Then, the notions of average reconstructability, average observability, and average detectability are defined and their necessary and sufficient conditions are provided. We also provide graph-theoretic interpretations of these notions through inter-cluster and intra-cluster graph topologies of a clustered network system. When a clustered network system does not meet the design criteria of the average observer, we devise an optimal design methodology to minimize the average estimation error. On the other hand, if the clusters are not pre-specified in a network system, we develop clustering algorithms to achieve minimum average estimation error. Finally, we propose a K-means type clustering approach to estimate the state variance of network systems, which is a nonlinear functional of the state vector and measures the squared deviation of state trajectories from their average mean. We illustrate the results through application examples of a building thermal system and an SIS epidemic spread over large networks.

In the second part of the thesis, we first study epidemic suppression through a testing policy. We develop a five-compartment epidemic model that incorporates the testing rate as a control input. We propose a best-effort strategy for testing (BEST), which is an epidemic suppression policy that provides a minimum testing rate from a certain day onward to stop the growth of the epidemic. The BEST policy is evaluated through its impact on the number of active intensive care unit (ICU) cases and the cumulative number of deaths for the COVID-19 case of France. Secondly, we develop a model of urban human mobility between residential areas and social destinations such as industrial areas, business parks, schools, markets, etc. for epidemic mitigation. We formulate two optimal control policies, the so-called optimal capacity control (OCC) and optimal schedule control (OSC), that aims to maximize the economic activity in an urban environment while keeping the number of active infected cases bounded. The OCC limits the epidemic spread by reducing the maximum number of people allowed at each destination category at any time of day, whereas the OSC limits the epidemic spread by reducing the daily business hours of each destination category.

Keywords Large-scale network systems, average observer, network clustering, state variance, epidemic spread, testing policy, urban human mobility

Resumé

Cette thèse de doctorat est réalisée principalement dans le cadre du projet Scale-FreeBack de l'European Research Council (ERC) Advanced Grant et partiellement dans le cadre du projet Healthy-Mobility de la mission COVID-19 de l'Inria. Le projet Scale-FreeBack vise à développer une approche holistique de contrôle sans échelle des systèmes complexes et à établir de nouvelles bases pour une théorie traitant des réseaux physiques complexes aux dimensions arbitraires. D'autre part, motivé par la pandémie COVID-19, le projet Healthy-Mobility vise à développer des stratégies de contrôle optimal pour les tests et la mobilité humaine urbaine pour limiter la propagation de l'épidémie. En relation avec ces deux projets, les contributions de la thèse sont respectivement divisées en deux parties.

Dans la première partie de la thèse, nous développons une théorie pour la surveillance de systèmes de réseaux en grappes à grande échelle avec des ressources de calcul et de détection limitées par le biais d'un système de réseau projeté, qui est de dimension traçable et est obtenu par l'agrégation de grappes d'un système de réseau. Nous proposons un observateur moyen d'ordre minimum et fournissons ses critères de conception. Ensuite, les notions de reconstructibilité moyenne, d'observabilité moyenne et de détectabilité moyenne sont définies et leurs conditions nécessaires et suffisantes sont fournies. Nous fournissons également des interprétations graph-théoriques de ces notions à travers les topologies de graphe inter-cluster et intra-cluster d'un système de réseau en grappe. Lorsqu'un système de réseau en grappe ne répond pas aux critères de conception de l'observateur moyen, nous concevons une méthodologie de conception optimale pour minimiser l'erreur d'estimation moyenne. D'autre part, si les clusters ne sont pas pré-spécifiés dans un système de réseau, nous développons des algorithmes de clustering pour atteindre une erreur d'estimation moyenne minimale. Enfin, nous proposons une approche de regroupement de type K-means pour estimer la variance d'état des systèmes en réseau, qui est une fonction non linéaire du vecteur d'état et mesure l'écart au carré des trajectoires d'état par rapport à leur moyenne. Nous illustrons les résultats par des exemples d'application d'un système thermique de bâtiment et d'une épidémie de SIS répandue sur de grands réseaux.

Dans la deuxième partie de la thèse, nous étudions d'abord la suppression des épidémies par une politique de test. Nous développons un modèle épidémique à cinq compartiments qui incorpore le taux de test comme donnée de contrôle. Nous proposons une stratégie de best-effort pour le test (BEST), qui est une politique de suppression d'épidémie qui fournit un taux de test minimum à partir d'un certain jour pour arrêter la croissance de l'épidémie. La politique BEST est évaluée à travers son impact sur le nombre de cas actifs dans les unités de soins intensifs (USI) et le nombre cumulé de décès pour le cas COVID-19 en France. Deuxièmement, nous développons un modèle de mobilité humaine urbaine entre les zones résidentielles et les destinations sociales telles que les zones industrielles, les parcs d'affaires, les écoles, les marchés, etc. pour l'atténuation des épidémies. Nous formulons deux politiques de contrôle optimal, le contrôle optimal de la capacité et le contrôle optimal de l'horaire, qui visent à maximiser l'activité économique dans un environnement urbain tout en maintenant le nombre de cas d'infection actifs limité. Le contrôle de la capacité optimale limite la propagation de l'épidémie en réduisant le nombre maximum de personnes autorisées dans chaque catégorie de destination à tout moment de la journée, tandis que le contrôle de l'horaire optimal limite la propagation de l'épidémie en réduisant les heures d'ouverture quotidiennes de chaque catégorie de destination.

Keywords Systèmes de réseaux à grande échelle, observateur moyen, regroupement de réseaux, variance d'état, propagation épidémique, mobilité humaine urbaine

To my son, Nuh.

Acknowledgment

All praise belongs to God, the Infinitely Good, the All-Merciful, and the Origin and Originator of all that is. I thank Thee for elucidating the obscurities, easing the difficulties, and providing strength and patience during my Ph.D.

I express my gratitude to Carlos Canudas-de-Wit and Alain Kibangou for their support and guidance throughout my thesis. Without their continuous assistance and supervision, this thesis would not have been possible. I learned from them many valuable lessons that not only enhanced my scholarly abilities but also instilled professionalism in my character.

I am very thankful to the jury members Didier Georges, Denis Efmiov, Jun-ichi Imura, Jacquelien M. A. Scherpen, Karl H. Johansson, and Pierre-Alexandre Bliman for their valuable comments and suggestions that significantly improved this thesis.

My family has always been supportive of me in my academic pursuit. I really appreciate and thank my wife, Isra, for her patience and support during the writing of this thesis. I will forever be indebted to my parents for their continuous love and care.

During my Ph.D., I was fortunate to have the company of Martin Rodriguez-Vega, Stephane Mollier, Vadim Bertrand, Diego Deplano, Ujjwal Pratap, Nicolas Martin, Liudmila Tumash, and Denis Nikitin. All the board games that we played together will always be part of my beautiful memories – especially the afternoons at Inria's café with pleasant coffee smell and delightful ambiance, where we kept optimizing Hanabi strategies and linking them to the concepts in multi-agent systems and cooperative game theory. I shared the office with Martin and had long and deep discussions with him on many relevant/irrelevant but important topics, which was a refreshing as well as an enjoyable experience. The family get-togethers and picnics with Elena & Tommaso and Laura & Martin were full of fun and laughter.

This thesis was mainly funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (ERCAdG no. 694209, Scale-FreeBack, website: http://scale-freeback.eu/) and partially by Inria, France, under the framework of the Inria Mission Covid-19 (Project HealthyMobility).

Contents

General Introduction

In this introductory chapter, we present the topics treated in this thesis, identify the challenges, and motivate our work with the help of real-world applications. The thesis is divided into two parts. The first part deals with aggregated monitoring of large-scale network systems and the second part with modeling and control of epidemics. We also summarize how each part of the thesis is organized and enlist our peer-reviewed publications.

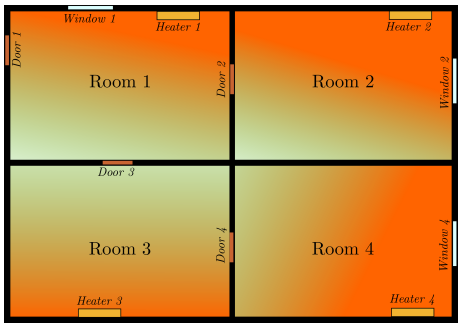
Part I Aggregated monitoring of large-scale network systems

System monitoring provides information about the system's state to assess its performance and give feedback to the controller. It requires the state to be estimated by an observer using the input and output measurements from the system. For large-scale network systems, however, monitoring requires tremendous amounts of computational and sensing resources, which becomes impractical under a limited budget. This is because the complexity of a large-scale network system may challenge the available computational resources, while a limited number of sensors may render the system unobservable. Such limitations make the state estimation task infeasible.

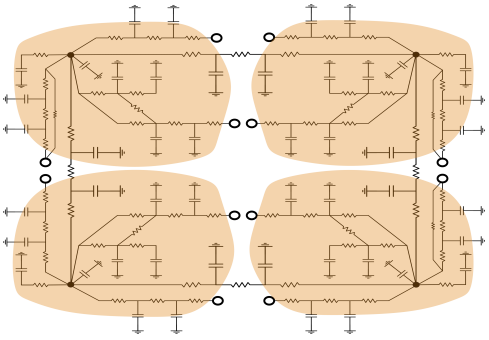
To deal with this issue, we propose aggregated monitoring based on the estimation of aggregated state profiles of a network system, for example, multi-cluster average and variance of the state vector. The multi-cluster average provides the mean state trajectory of each cluster in a network system, while the multi-cluster variance provides a measure of squared deviation of state trajectories in each cluster. Aggregated monitoring is reasonable for large-scale network systems such as building thermal systems [Deng2010, Deng2014], power modules [Murdock2006, Sakhraoui2018], urban traffic networks [Ramezani2015, Rodriguez-Vega2020], and epidemic spread over large networks [Martin2020]. We elaborate these examples below in the context of aggregated monitoring.

Building thermal systems

The first example is the thermal monitoring of residential buildings, which is important because of a significant share of the residential sector in energy consumption and greenhouse gas emissions [Hache2017, Lévy2018]. However, in addition to the limitations in sensing resources for providing temperature measurements, the sensing capability of thermistors is also limited. This is because thermistors can be placed in specific locations on walls and ceilings, where they can provide temperature measurements corresponding to only small areas around them. However, due to the diffusive nature of heat, it can be argued that those measurements are sufficient for thermal monitoring. Nonetheless, it has been shown that such temperature measurements fail to capture the temperature variance in the rooms, which can affect human comfort significantly [Boduch2009]. Therefore, estimating and regulating the mean operative temperature of each room ensures not only human comfort but also facilitates thermal monitoring [Niazi2020a].

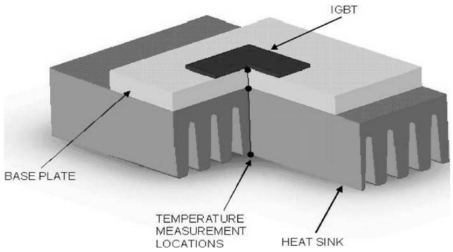


(a) Thermal system of a four-room building

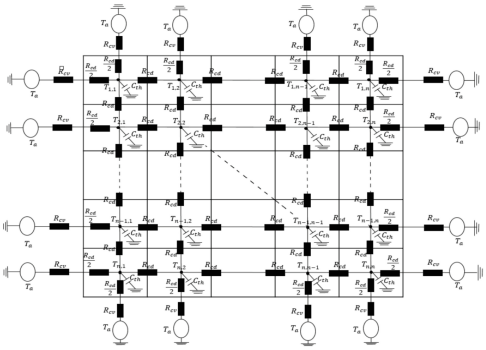


(b) Clustered RC-network model

Figure 1: An RC-network model of a building thermal system.

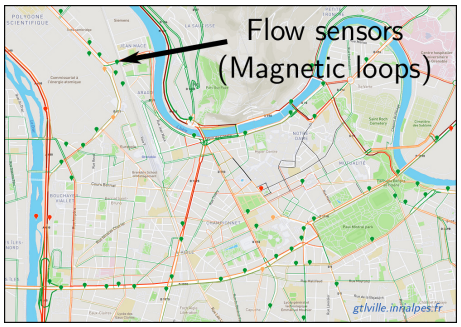


(a) Representation of a power module by [Murdock2006]

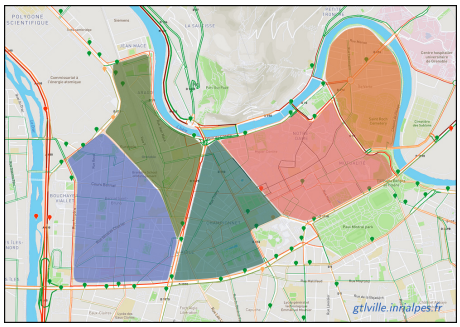


(b) Spatially-discrete 2D heated plate model by [Sakhraoui2018]

Figure 2: Power module as spatially discretized 2D heated plate.

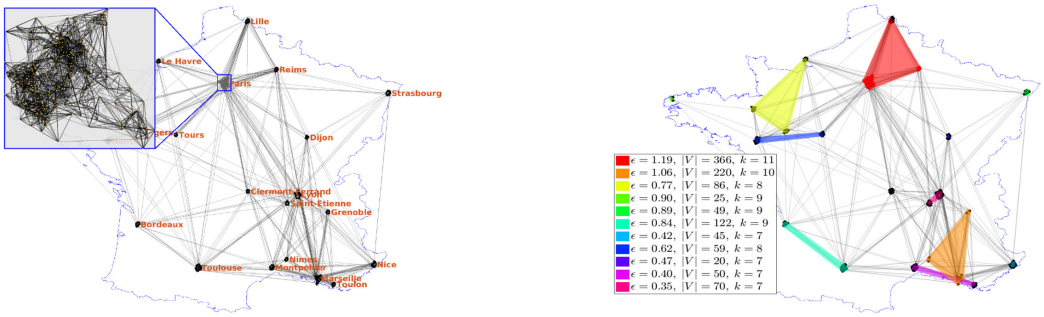


(a) Urban traffic network with flow sensors



(b) Cluster selection for aggregated monitoring

Figure 3: Urban traffic network of Grenoble, France.



(a) Network of main cities of France

(b) Cluster selection for aggregated monitoring by [Martin2020]

Figure 4: Epidemic process over a network of main cities of France.

In the second example, controlling thermal operating performance can avoid potentially damaging stresses on power modules [Murdock2006]. This entails that the precise knowledge of the local temperature of different areas of power modules, modeled as spatially-discrete 2D heated plate [Sakhraoui2018], must be obtained through estimation.

In the third example of urban traffic networks, estimating the traffic density of every road in an urban traffic network is often not possible [Ramezani2015]. Therefore, estimating the average traffic densities in multiple sectors of a network helps to monitor the congestion effectively [Rodriguez-Vega2020, Rodriguez-Vega2021].

Finally, in the event of an epidemic spread over large networks, it is challenging to measure the whole network for monitoring the epidemic situation. In such a case, it is more reasonable to identify clusters in the network and estimate the average number of infected people in each cluster [Martin2020]. Such a strategy can help, for instance, to devise preventive measures for controlling the epidemic spread in a country based on the clusters of several towns, which are connected through human mobility due to work or other purposes.

Part II Modeling and control of epidemics

The motivation behind the second part of this thesis is the COVID-19 epidemic. Started in Wuhan, China, at the end of 2019, it was soon declared to be a pandemic by the World Health Organization (WHO) on March 11, 2020. The common symptoms of the disease include fever, cough, fatigue, shortness of breath, and loss of sense of smell, where complications may lead to pneumonia and respiratory distress known as a severe acute respiratory syndrome (SARS). During the first year of the pandemic, the primary mode of treatment had been symptomatic and supportive therapy [Cao2020, Baden2020], and no vaccine or specific antiviral treatment had been approved.

The pandemic shook the economy of the whole world with significant reductions in exports, a decline in tourism, mass unemployment, and business closures [Loayza2020]. Governments and health authorities worldwide responded by implementing non-pharmaceutical intervention (NPI) policies that include travel restrictions, lockdown strategies, social distancing measures, workplace hazard controls, closing down of schools and workplaces, curfew strategies, and cancellation of public events. Many countries also upgraded existing infrastructure and personnel to increase testing capabilities and facilities for focused isolation. People were instructed to wash hands several times a day, cover mouth and nose when coughing or sneezing, maintain a certain physical distance from other people, wear a face

mask in public places/gatherings, and monitor and self-isolate if the disease symptoms appear. The extent to which such policies and measures have been implemented in a certain country is called the stringency index of its government's response [Hale2020a, Hale2020b]. Each government responded in its capacity to find a suitable balance between saving lives and saving livelihoods, which [Glover2020] termed as a perplexing problem of health versus wealth. Livelihoods can be saved through the implementation of suitable relief and recovery measures for people and small businesses. On the other hand, lives can be saved through the implementation of good testing and NPI policies. In other words, there is a direct relationship between the stringency index of the government and saving lives.

All the above policies implemented by the governments are control mechanisms for an epidemic. Such strategies fall under two categories: mitigation and suppression [Ferguson2020, Walker2020]. The mitigation strategies slow down the rate of transmission of disease or, in popular terms, 'flatten the curve.' However, they do not necessarily stop the disease spread, which is the goal of suppression strategies.

In addition to the NPI policies, testing and isolating the infected population from the susceptible population is one of the most important strategies to control the epidemic spread. For instance, it has been reported that COVID-19 was eliminated from the Italian village Vo'Euganeo through testing both symptomatic and asymptomatic cases [Romagnani2020, Day2020]. Moreover, in his media briefing[1] of March 16, 2020, Dr. Tedros Adhanom Ghebreyesus, the Director-General of WHO, urged the following:

> *"Social distancing measures can help to reduce transmission and enable health systems to cope. Hand-washing and coughing into your elbow can reduce the risk for yourself and others. But on their own, they are not enough to extinguish this pandemic. It's the combination that makes the difference. As I keep saying, all countries must take a comprehensive approach. But the most effective way to prevent infections and save lives is by breaking the chains of transmission. And to do that, you must test and isolate. You cannot fight a fire blindfolded. And we cannot stop this pandemic if we don't know who is infected. We have a simple message for all countries: TEST, TEST, TEST."*

Following the recommendation of the WHO director, with different levels of setups, many governments increased their testing capacities, while others feared the economic burden of intensive testing policy. However, [Eichenbaum2020,Salathé2020] show that such a burden is only short-term and, on the contrary, intensive testing reduces the overall cost of the epidemic in the long run because it enables the government to gain rapid control of the epidemic and revive the economy of a country. Since testing enables the health authority to identify and isolate the infected people from the susceptible population, thus limiting the disease transmission, it is considered to be a crucial control mechanism for the epidemic [Chowell2003]. However, as we will show in the literature review in Chapter 6, few attempts have been dedicated to studying the testing policies for epidemics from a control-theoretic perspective.

Another forefront for fighting epidemics is through the control of urban human mobility. Although it plays a vital role in a country's economy, it facilitates the spread of disease by allowing contact between infected and susceptible populations. If not controlled, urban human mobility can result in a huge number of infected cases, which could overwhelm the hospitals and cause the loss of lives. On the other hand, strict restrictions on mobility can halt the economy and result in the loss of livelihoods. Given the required objectives (e.g., socio-economic costs) and constraints (e.g., restricting the number of infected people under

[1] Website: https://www.who.int/dg/speeches/detail/

a given bound), the framework of optimal control theory can be employed to find optimal strategies for urban human mobility. For this, the models of urban human mobility that incorporate the process of disease spread play a vital role in the analysis, understanding, and mitigation of epidemics.

Thesis Organization

The first part of the thesis concerns the estimation of aggregated state profiles of large-scale network systems. In particular, we provide methodologies to estimate the multi-cluster average states and the state variance of large-scale network systems. In Chapter 1, we provide the literature review and identify our contributions related to aggregated monitoring. The problem is formulated in Chapter 2. Chapter 3 studies the notion under which the average states of pre-specified multiple clusters of a network system can be estimated asymptotically. When asymptotic estimation is not possible, we provide an optimal estimation methodology in Chapter 4 to estimate the average states with minimal error. Finally, Chapter 5 presents clustering techniques for optimal average estimation, open-loop average estimation, and variance estimation.

The second part of the thesis concerns the modeling and design of suppression and mitigation policies for epidemics. In particular, we study testing policies and control of urban human mobility in the event of an epidemic. After providing the literature review and identifying our contributions in Chapter 6, we provide a best-effort strategy for testing as a suppression policy for epidemics in Chapter 7. This policy provides a minimum testing rate to stop the growth of the epidemic. Chapter 8 provides a model of human mobility in an urban environment incorporating the process of an epidemic spread. Two optimal control policies related to the operating capacities and schedules of destinations are also devised using the framework of optimal control theory.

Publications

Journal articles

- M.U.B. Niazi, A.Y. Kibangou, C. Canudas-de-Wit, D. Nikitin, L. Tumash, and P.-A. Bliman. “Modeling and control of COVID-19 epidemic through testing policies." (Submitted to *Annual Reviews in Control*, 2020)

- M.U.B. Niazi, C. Canudas-de-Wit, and A.Y. Kibangou. “Average state estimation in large-scale clustered network systems." *IEEE Transactions on Control of Network Systems*, Vol. 7, No. 4, p. 1736-1745, 2020.

- M.U.B. Niazi, D. Deplano, C. Canudas-de-Wit, and A.Y. Kibangou. “Scale-free estimation of the average state in large-scale systems." *IEEE Control Systems Letters*, Vol. 4, No. 1, p. 211-216, 2019.

Conference papers

- M.U.B. Niazi, C. Canudas-de-Wit, A.Y. Kibangou, and P.-A. Bliman. “Optimal control of urban human mobility for epidemic mitigation." (Submitted to *60th IEEE Conference on Decision and Control (CDC)*, 2021)

- M.U.B. Niazi, A.Y. Kibangou, C. Canudas-de-Wit, D. Nikitin, L. Tumash, and P.-A. Bliman. “Testing policies for epidemic control." (Submitted to *60th IEEE Conference on Decision and Control (CDC)*, 2021)

- M.U.B. Niazi, C. Canudas-de-Wit, and A.Y. Kibangou. “State variance estimation in large-scale network systems." *59th IEEE Conference on Decision and Control (CDC)*, p. 6052-6057, 2020.
- M.U.B. Niazi, C. Canudas-de-Wit, and A.Y. Kibangou. “Thermal monitoring of buildings by aggregated temperature estimation." *IFAC World Congress*, Vol. 53, No. 2, pp. 4132-4137, 2020.
- M.U.B. Niazi, X. Cheng, C. Canudas-de-Wit, and J.M.A. Scherpen. “Structure-based clustering algorithm for model reduction of large-scale network systems." it 58th IEEE Conference on Decision and Control (CDC), p. 5038-5043, 2019.
- M.U.B. Niazi, C. Canudas-de-Wit, and A.Y. Kibangou. “Average observability of large-scale network systems." *18th European Control Conference (ECC)*, p. 1506-1511, 2019.

Notations

$\mathbb{N}$	The set of natural numbers
$\mathbb{Z}$	The set of integers
$\mathbb{R}$	The set of real numbers
$\mathbb{C}$	The set of complex numbers
$\mathbb{R}_{\geq 0}, \mathbb{R}_{>0}$	The sets of non-negative and positive real numbers, respectively
$\mathbb{R}_{\leq 0}, \mathbb{R}_{<0}$	The sets of non-positive and negative real numbers, respectively
$\mathbb{C}_{\geq 0}, \mathbb{C}_{>0}$	The closed and open right-half complex plane, respectively
$\mathbb{C}_{\leq 0}, \mathbb{C}_{<0}$	The closed and open left-half complex plane, respectively
$\mathbf{a}$	A vector is denoted by a bold lower-case letter
$\mathbf{a}^\intercal$	Transpose of vector $\mathbf{a} \in \mathbb{R}^n$
$[\mathbf{a}]_i$	The i-th entry of vector $\mathbf{a}$
$\mathbf{1}_n, \mathbf{0}_n$	The vector of ones and the vector of zeros of dimension $n \times 1$
A	A matrix is denoted by an upper-case letter
$A^\intercal$	Transpose of matrix $A \in \mathbb{R}^{m \times n}$
I_n	Identity matrix of dimension $n \times n$
A^{-1}	Inverse of a square, non-singular matrix $A \in \mathbb{R}^{n \times n}$, i.e., $A^{-1}A = AA^{-1} = I_n$
A^+	Left pseudo-inverse of a tall matrix $A \in \mathbb{R}^{m \times n}$ with $m > n$, i.e., $A^+A = I_n$
$A^\dagger$	Right pseudo-inverse of a fat matrix $A \in \mathbb{R}^{m \times n}$ with $m < n$, i.e., $AA^\dagger = I_m$
$[A]_{ij}$	The ij-th entry of matrix A
$\mathtt{im}(A)$	Image (or range) of $A \in \mathbb{R}^{m \times n}$
$\mathtt{ker}(A)$	Kernel (or nullspace) of $A \in \mathbb{R}^{m \times n}$
$\mathtt{rank}(A)$	Rank of $A \in \mathbb{R}^{m \times n}$, i.e., dimension of $\mathtt{im}(A)$
$\mathtt{nullity}(A)$	Nullity of $A \in \mathbb{R}^{m \times n}$, i.e., dimension of $\mathtt{ker}(A)$
$\mathtt{trace}(A)$	Trace of $A \in \mathbb{R}^{n \times n}$, i.e., $[A]_{11} + [A]_{22} + \cdots + [A]_{nn}$
$\mathtt{eig}(A)$	Spectrum of $A \in \mathbb{R}^{n \times n}$, i.e., the set of eigenvalues of A
$\|\mathbf{a}\|$	The Euclidean norm of $\mathbf{a} \in \mathbb{R}^n$, i.e., $\|\mathbf{a}\| = \sqrt{\mathbf{a}^\intercal \mathbf{a}}$
$\|\|\|A\|\|\|$	Norm of $A \in \mathbb{R}^{m \times n}$ induced by Euclidean norm
$\|\mathbf{f}\|_2, \|\mathbf{f}\|_\infty$	L_2 and L_∞ norm of a vector-valued function $\mathbf{f}$, respectively

Part I

Aggregated Monitoring of Large-scale Network Systems

1 State of the Art

This chapter provides a literature review of state estimation and aggregated monitoring of large-scale network systems, and describes our contributions in relation to the existing literature.

1.1 Literature Review

For monitoring and control of dynamical systems, knowledge of the system's state is undoubtedly necessary. However, due to limited sensing capability and resources, the complete state of a system is usually not accessible. An observer is therefore needed to estimate the state by using the knowledge of sensor measurements. In this regard, a seminal paper by Kalman and Bucy [Kalman1961] treats the problem of optimal filtering and state estimation for linear systems when the sensor measurements are corrupted by white noise. This paper builds on the prior paper of Kalman [Kalman1960] and proposes an optimal state estimator known as Kalman filter (or Kalman-Bucy filter), which depends on a time-varying gain matrix obtained in real-time by solving a matrix Ricatti differential equation. However, avoiding the use of differentiators for practical reasons [Bongiorno Jr1968] led to the development of Luenberger observer [Luenberger1964, Luenberger1966], which considers a constant gain matrix under the assumption that noise in sensor measurements is negligible. For linear time-invariant (LTI) systems, it was then proved that the existence of a gain matrix that ensures an asymptotic convergence of Luenberger observer to the true state at an arbitrary rate is equivalent to the observability of system [Kalman1963, Wonham1967, Gopinath1971, Luenberger1971].

With limited computational and sensing resources, it is often challenging and sometimes impossible to monitor large-scale systems by estimating the complete state [Antoulas2005]. This is because limited computational resources result in the computational intractability of the Luenberger observer, whereas limited sensing resources result in the unobservability of the system. However, it is sometimes unnecessary to estimate the complete state, and it suffices in many applications to estimate linear functionals of the state

vector. Examples of such cases include static state-feedback control [Kailath1980, Kautsky1985, Khargonekar1988], output-feedback control [Anderson1975, Kimura1977], decentralized control [Wang1973, Corfmat1976], and reduced-order dynamic control [Anderson1989, McFarlane1990].

A linear functional observer is also proposed by Luenberger in [Luenberger1971], where the order of the functional observer estimating a single linear functional is equal to the observability index of the system minus one. Two years later, [Murdoch1973] points out that the order of the functional observer is not minimum and can be reduced. A similar design procedure was also proposed in [Aldeen1994, Aldeen1999] for the order reduction of functional observers. Finally, Darouach [Darouach2000] provides a necessary and sufficient condition for the existence and design of a functional observer of minimum order that is possible to achieve, where the design procedure was recently improved in [Darouach2019]. The minimum order of Darouach observer is equal to the number of linear functionals that the observer estimates. However, for general linear systems, which do not satisfy the necessary and sufficient condition of [Darouach2000], finding a minimum-order functional observer whose dimension is greater than Darouach observer remained an open problem [Trinh2011]. This led to the development of functional observability in [Fernando2010a, Fernando2010b, Jennings2011] followed by [Rotella2011, Rotella2015, Rotella2016], proposing different methodologies to design a minimum-order functional observer. However, these methods find a minimum order by working within their framework, which makes it difficult to prove whether the obtained order is in fact minimum in an absolute sense. Moreover, the proposed methods are iterative and require computations of several observability matrices and the ranks of their concatenation at each iteration, which may not be computationally feasible for large-scale systems.

In light of the above, it is therefore reasonable to resort to methods based on aggregation of large-scale systems. In this regard, [Aoki1968, Wei1969, Coxson1984] introduced the notion of lumpability of large-scale systems, which can have multiple interpretations. Algebraically, an LTI system is lumpable if and only if the aggregation matrix is A-invariant [Atay2017], where A is the state matrix of the system. For network systems, on the other hand, a necessary condition for lumpability is that the aggregated clusters in a network are chosen according to an (almost) equitable partition [Ji2007, Martini2010, Egerstedt2012, Monshizadeh2014, Aguilar2017]. Therefore, the lumpability of network systems can be achieved by using clustering algorithms for equitable partition. However, for large-scale network systems, one may have constraints on the number of sensors and clusters, which makes network clustering very challenging. Moreover, when dealing with physical network systems such as urban traffic networks, building thermal systems, and water distribution networks, one may have yet another constraint, which is to obtain clusters that are physically connected [Martin2019].

Another line of research employs model reduction tools [Sandberg2004, Antoulas2005, Gugercin2008, Sandberg2009, Antoulas2010] that are known to be quite effective in reducing the complexity of large-scale systems [van der Schaft2013, Deng2014, Cheng2018b]. For network systems, however, in addition to the dynamical properties of the system, preserving the topological structure of the network is also critical. In this regard, clustering-based model reduction techniques [Ishizaki2014, Ishizaki2015, Cheng2017, Cheng2018a, Cheng2021] have shown promising results by not only preserving the topological structure but also providing technical tools to quantify model approximation error.

The main goal of clustering-based model reduction is to identify clusters in a network system and aggregate them for reducing the dimension of the system. That is, the optimal clustering yields a reduced system with minimum model approximation error, which is characterized in terms of $\mathcal{H}_2$ or $\mathcal{H}_\infty$ norms of the difference between frequency responses of both systems. The idea is to obtain a reduced system with a tractable dimension and whose

input-output behavior is similar to the input-output behavior of the original large-scale network system. Therefore, one can monitor network systems by estimating aggregated state profiles from the reduced system. In this regard, [Sadamoto2017] presents an average state observer by a clustering-based model reduction technique of [Ishizaki2015]. Such a technique is reasonable because the states of the reduced system approximate the average states of clusters in the original network system, where the clustering algorithm can be adapted to achieve a specified error bound. Thus, it suffices to design an average observer using the model of the reduced system to estimate the average states of the original network system. However, this approach becomes irrelevant when the clusters are a priori specified. In such a case, it is reasonable to study conditions under which the average states of the clusters can be reconstructed and/or asymptotically estimated.

1.2 Our Contributions

We study the estimation of aggregated state profiles of a large-scale network system such as multi-cluster average and variance of the state vector. The average states of multiple clusters of a network system are linear functionals of the state vector, whereas the variance is a nonlinear functional of the state that measures the spread of the state around its average mean. Since a large-scale network system poses difficulties when the computational resources are limited, we project its state on lower-dimensional state space and obtain a projected network system, which is an aggregated, tractable representation of the original network system. The projected network system considers each cluster as a single supernode and provides the dynamics of the average states of clusters. However, we show that the average deviation vector influences the dynamics of the projected network system by acting as a structured unknown input. Nonetheless, the goal of obtaining an aggregated representation of a large-scale network system is to attain computational tractability when designing an average observer for estimating the average states of the clusters.

First, we consider a clustered network system, where the clusters are specified a priori, and study the problem of average state estimation. We provide design criteria of an average observer with a minimum order, which is equal to the number of specified clusters, and define three notions: (i) average reconstructability, (ii) average observability, and (iii) average detectability. After providing necessary and sufficient conditions of these notions, we show that these notions are related to the convergence rate of the average observer. Under average reconstructability, the average states of clusters can be exponentially estimated by the average observer at an arbitrary rate. Average observability, on the other hand, allows for a high-gain type of average observer, where the average states of clusters can be exponentially estimated at a rate that approaches infinity. Finally, average detectability allows for the exponential open-loop estimation of average states at a fixed rate that depends on the eigenvalues of the projected network system. The graph-theoretic interpretations of these notions through the inter-cluster and intra-cluster graph topologies of the network system are also provided. We also provide some remarks about the graph structures of a network system that can be suitable for each of these notions.

When a clustered network system does not satisfy the necessary and sufficient condition of average reconstructability, we provide an optimal design of average observer to estimate the average states of clusters with minimal error. This optimal design is achieved by minimizing the effect of average deviation and ensuring the stability of the average observer. It is worth noticing that the main assumption in the literature for showing stability is to consider a network system with a strongly connected digraph, however, we show that this

assumption can be relaxed to weak connectivity of the digraph. The optimal average estimation problem is formulated as a convex optimization problem with a single decision parameter, which is solved by gradient descent and incremental search algorithms. The methodology of optimal average estimation is then applied to an application example of thermal monitoring in a four-room building.

For the case where the clusters are not specified a priori in a network system, we first provide a clustering algorithm for optimal average estimation by minimizing the distance from average reconstructability. For illustration, the effectiveness of this clustering methodology for average estimation is used in an application example of SIS epidemics over large networks. Another clustering algorithm for open-loop average estimation is provided that minimizes the distance from average lumpability, which is shown to be related to average detectability. Finally, we study the estimation of the state variance of a network system by using a K-means type clustering technique, which approximates the state variance by identifying the nodes whose state trajectories are closer to each other.

The part I of the thesis is organized as follows: In Chapter 2, we formulate the problem. The design of average observer and the notions of average reconstructability, average observability, and average detectability are studied in Chapter 3. Chapter 4 presents an optimal design of average observer. Finally, Chapter 5 presents clustering algorithms for optimal average estimation, open-loop average estimation, and state variance estimation.

2
Problem Formulation

This chapter formulates the problem of aggregated monitoring of a clustered network system. After defining a clustered network system in section 2.1, we describe the inter-cluster and intra-cluster graph topologies in section 2.2. Then, in section 2.3, through the aggregation of clusters, we obtain a projected network system, which is a tractable representation of a large-scale clustered network system. Finally, in section 2.4, we provide problem statements that are studied in the first part of this thesis.

Contents

In this chapter, we define a class of network systems studied in the first part of this thesis. If the nodes of a network system are partitioned into multiple clusters, then it is called a clustered network system. We describe the inter-cluster and intra-cluster topologies describing the induced subgraph structures between each pair of clusters and within each cluster, respectively. Through the aggregation of clusters, we project the state of a clustered network system on a lower-dimensional state-space and obtain a projected network system, which is of tractable dimension and computationally feasible for aggregated monitoring. Finally, we list the problem statements studied in part I of the thesis.

2.1 Clustered Network System

Consider a weighted digraph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with the set of nodes $\mathcal{V}$ and the set of directed edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. A directed edge (i, j) is an arc from node j to i, i.e., $i \leftarrow j$, with a weight $a_{ij} > 0$ associated to it. Now, consider a linear, time-invariant network system defined over a digraph $\mathcal{G}$ with the dynamics of each node $i \in \mathcal{V}$ given by

$$\dot{x}_i(t) = a_{ii}x_i(t) + \sum_{j \in \mathcal{N}_{i \leftarrow \mathcal{V}}} a_{ij}x_j(t) + \sum_{l=1}^{p} b_{il}u_l(t) \tag{2.1}$$

where $x_i(t) \in \mathbb{R}$ is the state of $i \in \mathcal{V}$, $u_l(t) \in \mathbb{R}$, for $l = 1, \ldots, p$, are the inputs to i with weights $b_{il} \in \mathbb{R}$, and

$$\mathcal{N}_{i \leftarrow \mathcal{V}} = \{j \in \mathcal{V} : (i, j) \in \mathcal{E},\ j \neq i\}$$

is the set of i's in-neighbors in $\mathcal{V}$. The first term on the right-hand side of (2.1) refers to the local damping at each node i with the weight $a_{ii} \in \mathbb{R}_{\leq 0}$. The second term corresponds to the aggregated inflow to node i from all its in-neighbors $j \in \mathcal{N}_{i \leftarrow \mathcal{V}}$, each of which is weighted by $a_{ij} \in \mathbb{R}_{>0}$, respectively. Finally, the third term depicts the aggregated influence of the inputs $u_l(t)$, for $l = 1, \ldots, p$, at i, each of which is weighted by $b_{il} \in \mathbb{R}$, respectively.

The local damping corresponds to self-loop of a node in the digraph $\mathcal{G}$. However, for the sake of simplicity, we shall omit the self loops in the figures and consider them implicit when illustrating digraphs. We further remark that the self-loop weights a_{ii}, for all $i \in \mathcal{V}$, considered in (2.1) are quite general. There might be a network structure associated to these scalars as in consensus-seeking multi-agent systems, [Olfati-Saber2007], where

$$a_{ii} = -\sum_{j \in \mathcal{N}_{i \leftarrow \mathcal{V}}} a_{ij}$$

or in spatially-discrete reaction-diffusion systems, [Ishizaki2014], where

$$a_{ii} = -\left(r_i + \sum_{j \in \mathcal{N}_{i \leftarrow \mathcal{V}}} a_{ij}\right)$$

with r_i being the reaction rate (chemical dissolution) at i, or in linear flow networks, [Walter1999], where

$$a_{ii} = -\sum_{j \in \mathcal{N}_{i \rightarrow \mathcal{V}}} a_{ji}$$

with

$$\mathcal{N}_{i\to\mathcal{V}} = \{j \in \mathcal{V} : (j,i) \in \mathcal{E}, j \neq i\} \tag{2.2}$$

the set of node i's out-neighbors in $\mathcal{V}$.

The nodes of the network system, without loss of generality, are partitioned into measured and unmeasured nodes. A measured node, also known as a gateway node [Bullo2018], is a node where a dedicated sensor is placed that provides its state measurements with respect to time. On the other hand, there are no dedicated sensors for unmeasured nodes. Let

$$\mathcal{V}_1 = \{\mu_1, \ldots, \mu_m\}, \quad \mathcal{V}_2 = \{\nu_1, \ldots, \nu_n\}$$

be the sets of m measured and n unmeasured nodes, respectively, where $\mathcal{V}_1 \cup \mathcal{V}_2 = \mathcal{V}$ and $\mathcal{V}_1 \cap \mathcal{V}_2 = \emptyset$. Similarly, let

$$\mathbf{x}_1(t) = [\, x_{\mu_1}(t) \;\ldots\; x_{\mu_m}(t) \,]^\intercal \in \mathbb{R}^m, \quad \mathbf{x}_2(t) = [\, x_{\nu_1}(t) \;\ldots\; x_{\nu_n}(t) \,]^\intercal \in \mathbb{R}^n$$

be the state vectors of measured nodes $\mathcal{V}_1$ and unmeasured nodes $\mathcal{V}_2$, respectively. We shall also refer $\mathbf{x}_1(t)$ as the measured state vector and $\mathbf{x}_2(t)$ as the unmeasured state vector. By considering

$$\mathbf{x}(t) = [\; \mathbf{x}_1^\intercal(t) \quad \mathbf{x}_2^\intercal(t) \;]^\intercal \in \mathbb{R}^{m+n}, \quad \mathbf{u}(t) = [\, u_1(t) \;\ldots\; u_p(t) \,]^\intercal \in \mathbb{R}^p$$

to be the network's state vector and the input vector, we represent the network system (2.1) in vector-form as

$$\boldsymbol{\Sigma} : \begin{cases} \dot{\mathbf{x}}(t) &= A\mathbf{x}(t) + B\mathbf{u}(t) \\ \mathbf{y}(t) &= C\mathbf{x}(t) \end{cases} \tag{2.3}$$

where $\mathbf{y}(t) = \mathbf{x}_1(t) \in \mathbb{R}^m$ is the measured state vector or, simply, the output vector. The ij-th entry of state matrix $A \in \mathbb{R}^{(m+n)\times(m+n)}$ is

$$[A]_{ij} = \begin{cases} a_{ij} > 0, & \text{if } (i,j) \in \mathcal{E} \text{ and } i \neq j \\ a_{ii} \leq 0, & \text{if } i = j \\ 0\,, & \text{otherwise.} \end{cases}$$

In other words, the off-diagonal entries of A constitute the graph structure $\mathcal{G}$ of the network system and the diagonal entries correspond to the local damping at its nodes. The input matrix $B \in \mathbb{R}^{(m+n)\times p}$ contains the weighted input configurations of the network, namely $[B]_{il} = b_{il} \in \mathbb{R}$, for $i = 1, \ldots, m+n$ and $l = 1, \ldots, p$. The output matrix $C \in \mathbb{R}^{m\times(m+n)}$ is given by $C = [\, \mathbf{e}_{\mu_1} \;\ldots\; \mathbf{e}_{\mu_m} \,]^\intercal$, where $\mu_1, \ldots, \mu_m$ are the measured nodes and $\mathbf{e}_{\mu_i} \in \mathbb{R}^{m+n}$ is a standard basis vector given by the μ_i-th column of the identity matrix I_{m+n}. Since the states can be reordered by transforming the network system $\boldsymbol{\Sigma}$ by a permutation matrix, we can assume, without loss of generality, that the nodes in the network are arranged as follows: $\mathcal{V} = \{\mu_1, \ldots, \mu_m, \nu_1, \ldots, \nu_n\}$. This gives the following block partition of the state matrices

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \qquad C = \begin{bmatrix} I_m & 0_{m\times n} \end{bmatrix} \tag{2.4}$$

where $A_{11} \in \mathbb{R}^{m\times m}$, $A_{22} \in \mathbb{R}^{n\times n}$, $A_{12} \in \mathbb{R}^{m\times n}_{\geq 0}$, $A_{21} \in \mathbb{R}^{m\times n}_{\geq 0}$, $B_1 \in \mathbb{R}^{m\times p}$, and $B_2 \in \mathbb{R}^{n\times p}$.

We shall abide by the following assumption throughout part I of the thesis:

Assumption 2.1. We assume that $\texttt{rank}(A_{12}) = m$.

To interpret Assumption 2.1, suppose $A_{12} \in \{0,1\}^{m\times n}$ is a structured matrix. Then, the full-row rank of A_{12} means that (i) no measured node is a disconnected node in $\mathcal{G}$ and (ii) no pair of measured nodes has the same set of unmeasured nodes as their in-neighbors. To elucidate further, if there is a disconnected measured node μ_i in $\mathcal{G}$, this implies that a corresponding row i of A_{12} will be zero. Moreover, if a pair of measured nodes μ_i, μ_j have the same set of unmeasured nodes as their in-neighbors, i.e., $\mathcal{N}_{\mu_i \leftarrow \mathcal{V}_2} \cap \mathcal{V}_2 = \mathcal{N}_{\mu_j \leftarrow \mathcal{V}_2} \cap \mathcal{V}_2$, then two corresponding rows i and j of A_{12} can be linearly dependent. Both of these cases lead to a rank deficiency of A_{12}. In other words, Assumption 2.1 supposes that, since they have a significant economic cost, the sensors are placed strategically to maximize the coverage for network system monitoring. Nonetheless, we consider that the network system $\mathbf{\Sigma}$ is a large-scale system with the number of nodes very large. Moreover, there is a limited number of available sensors so that $m \ll n$—i.e., the number of measured nodes is very small as compared to the number of unmeasured nodes. This implies that $\mathbf{\Sigma}$ may not be observable, which means that the observability matrix

$$\mathcal{O} = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{m+n-1} \end{bmatrix}$$

is rank deficient.

Now, to define a clustered network system, we suppose that the unmeasured nodes $\mathcal{V}_2$ are partitioned into k disjoint clusters, where $k < n$. That is to say that we are given a partition, or clustering, $\mathcal{Q} = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$ of the network's unmeasured nodes $\mathcal{V}_2$ such that $\mathcal{C}_1 \cup \mathcal{C}_2 \cup \cdots \cup \mathcal{C}_k = \mathcal{V}_2$ and, for any $\alpha, \beta \in \{1, \cdots, k\}$ and $\alpha \neq \beta$, we have $\mathcal{C}_\alpha \cap \mathcal{C}_\beta = \emptyset$.

Definition 2.1. A network system $\mathbf{\Sigma}$ with a set of measured nodes $\mathcal{V}_1$ and a clustering $\mathcal{Q} = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$ of unmeasured nodes $\mathcal{V}_2$ is called a clustered network system, which is denoted by $\mathbf{\Sigma}_{\mathcal{V}_1, \mathcal{Q}}$.

Let

$$\mathfrak{C}_{n,k} = \{X \in \{0,1\}^{n\times k} : X\mathbf{1}_k = \mathbf{1}_n\}$$

be the set of characteristic matrices of all clusterings with k clusters of n nodes. Then, for a given clustering $\mathcal{Q}$, the characteristic matrix $Q \in \mathfrak{C}_{n,k}$ is defined as follows: For $i \in \{1, \ldots, n\}$ and $\alpha \in \{1, \ldots, k\}$,

$$[Q]_{i\alpha} = \begin{cases} 1, & \text{if } \nu_i \in \mathcal{C}_\alpha \\ 0, & \text{otherwise.} \end{cases} \tag{2.5}$$

Let $n_\alpha = |\mathcal{C}_\alpha|$ be the number of nodes in cluster $\mathcal{C}_\alpha$, then $Q^\intercal Q = \texttt{diag}(n_1, \ldots, n_k)$ and the left pseudo-inverse Q^+ of Q, i.e., $Q^+Q = I_k$, is given by

$$Q^+ = (Q^\intercal Q)^{-1} Q^\intercal \in \mathbb{R}^{k\times n}$$

where

$$[Q^+]_{\alpha i} = \begin{cases} \dfrac{1}{n_\alpha}, & \text{if } \nu_i \in \mathcal{C}_\alpha \\ 0, & \text{otherwise.} \end{cases} \tag{2.6}$$

2.2 Inter-Cluster and Intra-Cluster Graph Topologies

In relation to the clusters $\mathcal{C}_1, \ldots, \mathcal{C}_k$, we consider the following induced subgraphs of $\mathcal{G}$:

- For $\alpha \in \{1, \ldots, k\}$, the topology from cluster $\mathcal{C}_\alpha$ to the set of measured nodes $\mathcal{V}_1$ is captured by an induced bipartite subgraph $\mathcal{G}_{\mu\alpha} = (\mathcal{V}_1, \mathcal{C}_\alpha, \mathcal{E}_{\mu\alpha})$, where $\mathcal{E}_{\mu\alpha} = \mathcal{E} \cap (\mathcal{V}_1 \times \mathcal{C}_\alpha)$ is the set of directed edges from all the nodes in $\mathcal{C}_\alpha$ to all the nodes in $\mathcal{V}_1 = \{\mu_1, \ldots, \mu_m\}$. The set of all such subgraphs is $\{\mathcal{G}_{\mu 1}, \ldots, \mathcal{G}_{\mu k}\}$.

- For $\alpha, \beta \in \{1, \ldots, k\}$ and $\alpha \neq \beta$, the inter-cluster topology from cluster $\mathcal{C}_\beta$ to cluster $\mathcal{C}_\alpha$ is captured by an induced bipartite subgraph $\mathcal{G}_{\alpha\beta} = (\mathcal{C}_\alpha, \mathcal{C}_\beta, \mathcal{E}_{\alpha\beta})$, where $\mathcal{E}_{\alpha\beta} = \mathcal{E} \cap (\mathcal{C}_\alpha \times \mathcal{C}_\beta)$ is the set of directed edges from $\mathcal{C}_\beta$ to $\mathcal{C}_\alpha$. The set of all such subgraphs is $\{\mathcal{G}_{12}, \ldots, \mathcal{G}_{1k}, \mathcal{G}_{21}, \mathcal{G}_{23}, \ldots, \mathcal{G}_{2k}, \ldots, \mathcal{G}_{k1}, \ldots, \mathcal{G}_{k(k-1)}\}$.

- For $\alpha \in \{1, \ldots, k\}$, the intra-cluster topology of cluster $\mathcal{C}_\alpha$ is captured by induced subgraph $\mathcal{G}_{\alpha\alpha} = (\mathcal{C}_\alpha, \mathcal{E}_{\alpha\alpha})$, where $\mathcal{E}_{\alpha\alpha} = \mathcal{E} \cap (\mathcal{C}_\alpha \times \mathcal{C}_\alpha)$ is the set of directed edges from all the nodes in $\mathcal{C}_\alpha$ to their out-neighbors in $\mathcal{C}_\alpha$. The set of all such subgraphs is $\{\mathcal{G}_{11}, \ldots, \mathcal{G}_{kk}\}$.

An induced bipartite subgraph $\mathcal{G}_{\mu\alpha}$ is illustrated in Figure 2.1(b), and the inter-cluster induced bipartite subgraphs $\mathcal{G}_{\alpha\beta}$ and the intra-cluster induced subgraphs $\mathcal{G}_{\alpha\alpha}$ are illustrated in Figure 2.1(c) and (d), respectively, where the black nodes are the measured nodes and the colored nodes are the unmeasured nodes. Notice that, other than $\mathcal{G}_{32}$, all the inter-cluster induced bipartite subgraphs are empty (or edgeless) in the example of Figure 2.1(a).

For a given clustering $\mathcal{Q}$, if we reorder the unmeasured nodes

$$\mathcal{V}_2 = \{\nu_1, \ldots, \nu_n\} \to \{\tilde{\nu}_1, \ldots, \tilde{\nu}_n\}$$

such that

$$\begin{aligned} \mathcal{C}_1 &= \{\tilde{\nu}_1, \ldots, \tilde{\nu}_{n_1}\} \\ \mathcal{C}_2 &= \{\tilde{\nu}_{n_1+1}, \ldots, \tilde{\nu}_{n_1+n_2}\} \\ &\vdots \qquad\quad \vdots \\ \mathcal{C}_k &= \{\tilde{\nu}_{n_1+\cdots+n_{k-1}+1}, \ldots, \tilde{\nu}_{n_1+\cdots+n_{k-1}+n_k}\} \end{aligned}$$

then the characteristic matrix

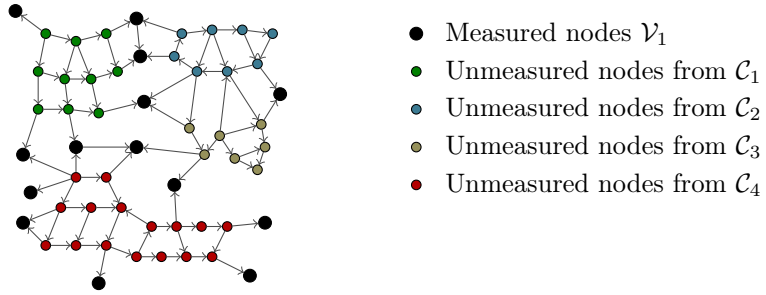
$$Q = \mathtt{diag}(\mathbf{1}_{n_1}, \mathbf{1}_{n_2}, \ldots, \mathbf{1}_{n_k}).$$

We can divide two matrix blocks of the state matrix A in (2.4) according to clusters as

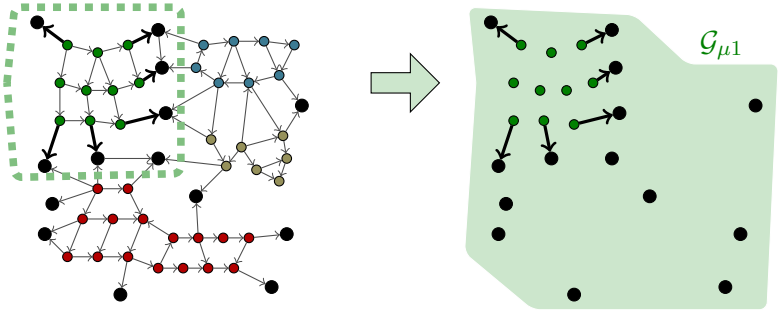
$$A_{12} := \begin{bmatrix} \tilde{A}_{\mu 1} & \ldots & \tilde{A}_{\mu k} \end{bmatrix}, \quad A_{22} := \begin{bmatrix} \tilde{A}_{11} & \ldots & \tilde{A}_{1k} \\ \vdots & \ddots & \vdots \\ \tilde{A}_{k1} & \ldots & \tilde{A}_{kk} \end{bmatrix}.$$

The sub-matrices $\tilde{A}_{\mu\alpha}$ of A_{12}, for $\alpha = 1, \ldots, k$, contain the weighted edge configurations from $\mathcal{C}_\alpha$ to $\mathcal{V}_1$. In other words, $\tilde{A}_{\mu\alpha}$ is the biadjacency matrix of the induced bipartite subgraph $\mathcal{G}_{\mu\alpha}$ with

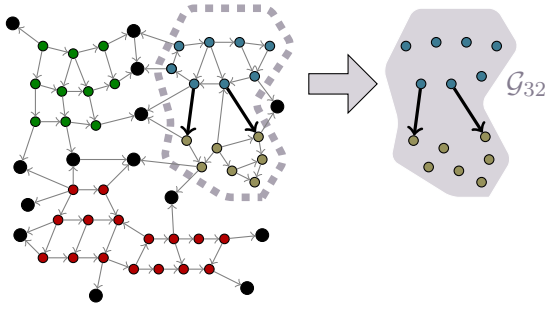
$$\begin{cases} [\tilde{A}_{\mu\alpha}]_{ij} > 0, & \text{if } (i,j) \in \mathcal{E}_{\mu\alpha} \\ [\tilde{A}_{\mu\alpha}]_{ij} = 0, & \text{otherwise.} \end{cases}$$



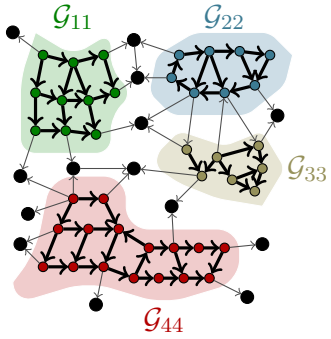
(a) The graph $\mathcal{G}$ of the clustered network system $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$, where $\mathcal{Q} = \{\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4\}$.



(b) An induced bipartite subgraph $\mathcal{G}_{\mu 1}$ that contains directed edges from cluster $\mathcal{C}_1$ to the measured nodes $\mathcal{V}_1$.



(c) An induced bipartite subgraph $\mathcal{G}_{32}$ capturing inter-cluster topology from cluster $\mathcal{C}_2$ to cluster $\mathcal{C}_3$.



(d) All the induced subgraphs $\mathcal{G}_{\alpha\alpha}$ capturing intra-cluster topology of clusters $\mathcal{C}_\alpha$, for $\alpha = 1, 2, 3, 4$, respectively.

Figure 2.1: Different topologies embedded in a clustered network system $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$.

Similarly, for $\alpha, \beta = 1, \ldots, k$ and $\alpha \neq \beta$, the sub-matrices $\tilde{A}_{\alpha\alpha}$ and $\tilde{A}_{\alpha\beta}$ of A_{22} correspond to the intra-cluster induced subgraphs $\mathcal{G}_{\alpha\alpha}$ and the biadjacency matrix of the induced bipartite subgraph $\mathcal{G}_{\alpha\beta}$, respectively, where

$$\begin{cases} [\tilde{A}_{\alpha\alpha}]_{ij} > 0, & \text{if } (i,j) \in \mathcal{E}_{\alpha\alpha} \\ [\tilde{A}_{\alpha\alpha}]_{ij} = 0, & \text{otherwise;} \end{cases}$$
$$\begin{cases} [\tilde{A}_{\alpha\beta}]_{ij} > 0, & \text{if } (i,j) \in \mathcal{E}_{\alpha\beta} \\ [\tilde{A}_{\alpha\beta}]_{ij} = 0, & \text{otherwise.} \end{cases}$$

Finally, we define the *outflow centrality* of $i \in \mathcal{V}_2$ as

$$c_{i\to\mathcal{V}} = a_{ii} + \sum_{j \in \mathcal{N}_{i\to\mathcal{V}}} a_{ji}$$

where $\mathcal{N}_{i\to\mathcal{V}}$ is the set of i's out-neighbors defined in (2.2). Then, the *relative outflow centrality* of $i \in \mathcal{V}$ with respect to cluster $\mathcal{C}_\alpha$, for $\alpha \in \{1, \ldots, k\}$, is defined as

$$c_{i\to\mathcal{C}_\alpha} = \begin{cases} a_{ii} + \displaystyle\sum_{j \in \mathcal{C}_\alpha \cap \mathcal{N}_{i\to\mathcal{V}}} a_{ji}, & \text{if } i \in \mathcal{C}_\alpha \\ \displaystyle\sum_{j \in \mathcal{C}_\alpha \cap \mathcal{N}_{i\to\mathcal{V}}} a_{ji}, & \text{if } i \notin \mathcal{C}_\alpha. \end{cases} \tag{2.7}$$

2.3 Projected Network System

The average state of each cluster $\mathcal{C}_\alpha$ is defined as

$$z_\alpha(t) := \frac{1}{n_\alpha} \sum_{i \in \mathcal{C}_\alpha} x_i(t)$$

for $\alpha = 1, \ldots, k$. In other words, the average state vector

$$\mathbf{z}_\mathtt{a}(t) = [\, z_1(t) \; \cdots \; z_k(t) \,]^\intercal \in \mathbb{R}^k$$

is defined through the aggregation of each cluster $\mathcal{C}_\alpha$ as

$$\mathbf{z}_\mathtt{a}(t) = Q^+ \mathbf{x}_2(t). \tag{2.8}$$

By considering each cluster $\mathcal{C}_\alpha$ as a single supernode, we project the unmeasured state vector $\mathbf{x}_2(t) \in \mathbb{R}^n$ on a lower dimensional state space $\mathbb{R}^k$, where $k < n$, and obtain a projected network system

$$\mathring{\mathbf{\Sigma}}_{\mathcal{V}_1,\mathcal{Q}} : \begin{cases} \dot{\mathbf{z}}(t) &= E\mathbf{z}(t) + F\boldsymbol{\sigma}(t) + G\mathbf{u}(t) \\ \mathbf{0}_k &= Q^+\boldsymbol{\sigma}(t) \\ \mathbf{y}(t) &= H\mathbf{z}(t) \end{cases} \tag{2.9}$$

with the state vector

$$\mathbf{z}(t) = [\; \mathbf{x}_1^\intercal(t) \quad \mathbf{z}_\mathtt{a}^\intercal(t) \;]^\intercal \in \mathbb{R}^{m+k}$$

and the average deviation vector

$$\boldsymbol{\sigma}(t) := (I_n - QQ^+)\mathbf{x}_2(t) \in \mathbb{R}^n.$$

The system matrices of $\mathring{\boldsymbol{\Sigma}}_{\mathcal{V}_1,\mathcal{Q}}$ are

$$\begin{aligned} E &= \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix} := \begin{bmatrix} A_{11} & A_{12}Q \\ Q^+A_{21} & Q^+A_{22}Q \end{bmatrix} \\ F &= \begin{bmatrix} F_1 \\ F_2 \end{bmatrix} := \begin{bmatrix} A_{12} \\ Q^+A_{22} \end{bmatrix} \\ G &= \begin{bmatrix} G_1 \\ G_2 \end{bmatrix} := \begin{bmatrix} B_1 \\ Q^+B_2 \end{bmatrix} \\ H &= \begin{bmatrix} H_1 & H_2 \end{bmatrix} := \begin{bmatrix} I_m & 0_{m\times k} \end{bmatrix}. \end{aligned} \tag{2.10}$$

Note that both $\mathring{\boldsymbol{\Sigma}}_{\mathcal{V}_1,\mathcal{Q}}$ and $\boldsymbol{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ yield the same output $\mathbf{y}(t)$ because

$$\mathbf{y}(t) = C\mathbf{x}(t) = H\mathbf{z}(t) = \mathbf{x}_1(t)$$

with

$$\mathbf{z}(t) = \begin{bmatrix} I_m & 0_{m\times n} \\ 0_{k\times m} & Q^+ \end{bmatrix} \mathbf{x}(t)$$

and

$$\mathbf{x}(t) = \begin{bmatrix} I_m & 0_{m\times k} \\ 0_{n\times m} & Q \end{bmatrix} \mathbf{z}(t) + \begin{bmatrix} 0_{m\times n} \\ I_n \end{bmatrix} \boldsymbol{\sigma}(t).$$

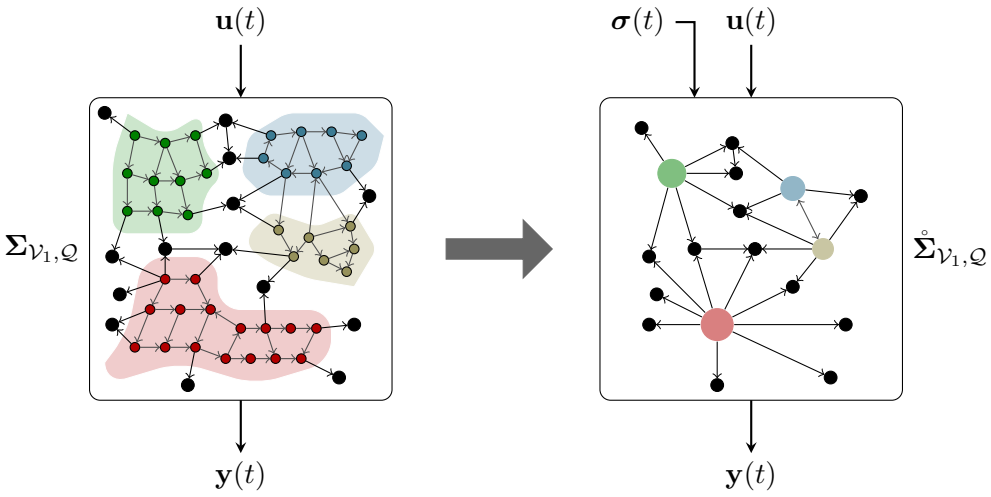


Figure 2.2: Obtaining a projected network system $\mathring{\boldsymbol{\Sigma}}_{\mathcal{V}_1,\mathcal{Q}}$ from a clustered network system $\boldsymbol{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ by aggregating the clusters of unmeasured nodes.

The average deviation vector $\boldsymbol{\sigma}(t)$ is a cluster-wise mean-centered vector given by

$$[\boldsymbol{\sigma}(t)]_i = [\mathbf{x}_2(t)]_i - z_\alpha(t) \tag{2.11}$$

for $\nu_i \in \mathcal{C}_\alpha$. Recall that $\boldsymbol{\sigma} = J_Q \mathbf{x}_2$, where the matrix $J_Q := I_n - QQ^+$ is symmetric ($J_Q^\intercal = J_Q$) and idempotent ($J_Q^2 = J_Q$). The columns of J_Q form a complete basis of $\texttt{ker}(Q^+)$ because $Q^+ J_Q = 0_{k\times n}$ and $\texttt{nullity}(Q^+) = \texttt{rank}(J_Q) = n-k$. Therefore, $\boldsymbol{\sigma}(t) \in \texttt{ker}(Q^+)$ which means $Q^+\boldsymbol{\sigma} \equiv \mathbf{0}_k$. Finally, note that $\boldsymbol{\sigma}(t)$ acts as a structured unknown input in the projected system $\mathring{\boldsymbol{\Sigma}}_{\mathcal{V}_1,\mathcal{Q}}$. It is structured because it satisfies $Q^+\boldsymbol{\sigma} \equiv \mathbf{0}_k$ and unknown because it is a function of the unmeasured state $\mathbf{x}_2(t)$.

The purpose of $\mathring{\boldsymbol{\Sigma}}_{\mathcal{V}_1,\mathcal{Q}}$ is to attain computational feasibility since the dimension of its state space is much lower than that of the clustered network system $\boldsymbol{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ when $k \ll n$. We base our analysis in the following chapters on $\mathring{\boldsymbol{\Sigma}}_{\mathcal{V}_1,\mathcal{Q}}$, where we don't consider the dynamics of $\boldsymbol{\sigma}(t)$—we only consider its structural property, i.e., $Q^+\boldsymbol{\sigma} \equiv \mathbf{0}_k$, which is the only known information we assume about $\boldsymbol{\sigma}(t)$. Even though considering the dynamics $\dot{\boldsymbol{\sigma}}(t)$ may allow for more information about the average deviation, however, it will require the dimension of $\mathring{\boldsymbol{\Sigma}}_{\mathcal{V}_1,\mathcal{Q}}$ to be equal to $m+k+n$, which doesn't attain computational feasibility since this dimension is even larger than the dimension of clustered network system $\boldsymbol{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$.

2.4 Problem Statement

In part I of the thesis, we study the following problems:

Problem 1 Given a clustered network system $\boldsymbol{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ with the set of measured nodes $\mathcal{V}_1$ and the clustering $\mathcal{Q}$ of unmeasured nodes $\mathcal{V}_2$, under what conditions on the inter-cluster, intra-cluster, and clusters-to-measured nodes topologies is it possible to estimate the average state vector $\mathbf{z}_\mathtt{a}(t)$ of clusters from the projected network system $\mathring{\boldsymbol{\Sigma}}_{\mathcal{V}_1,\mathcal{Q}}$ such that $\mathbf{z}_\mathtt{a}(t) - \hat{\mathbf{z}}_\mathtt{a}(t)$ converges to $\mathbf{0}_k$ asymptotically when $t \to \infty$, where $\hat{\mathbf{z}}_\mathtt{a}(t)$ is an estimated average state vector?

Problem 2 Given a clustered network system $\boldsymbol{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ with the set of measured nodes $\mathcal{V}_1$ and the clustering $\mathcal{Q}$ of unmeasured nodes $\mathcal{V}_2$, if the conditions obtained in Problem 1 are not satisfied, is it possible to devise an optimal methodology such that the average estimation error $\mathbf{z}_\mathtt{a}(t) - \hat{\mathbf{z}}_\mathtt{a}(t)$ is as small as possible when $t \to \infty$?

Problem 3 Given a network system $\boldsymbol{\Sigma}$ with the sets of measured nodes $\mathcal{V}_1$ and unmeasured nodes $\mathcal{V}_2$, find a clustering $\mathcal{Q}$ of $\mathcal{V}_2$ such that the obtained clustered network system $\boldsymbol{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ provides an optimal solution of Problem 2.

Problem 4 Given a network system $\boldsymbol{\Sigma}$ with the sets of measured nodes $\mathcal{V}_1$ and unmeasured nodes $\mathcal{V}_2$, devise a methodology to estimate the state variance of unmeasured nodes of $\boldsymbol{\Sigma}$ defined as

$$x_\mathtt{v}(t) = \frac{1}{n}\sum_{i\in\mathcal{V}_2}\left(x_i(t) - \frac{1}{n}\sum_{j\in\mathcal{V}_2} x_j(t)\right)^2$$

which measures the spread of unmeasured states around their average mean.

In Chapter 3 and 4, we study problems 1 and 2, respectively. In sections 5.1, 5.2, and 5.3 of Chapter 5, we study problem 3. Finally, section 5.4 of Chapter 5 studies problem 4.

3

Design of Average Observer for Clustered Network Systems

This chapter provides conditions under which the average states of pre-specified clusters of a network system can be asymptotically estimated. In section 3.1, we provide a minimum-order average observer and its design criteria. Then, in sections 3.2, 3.3, and 3.4, we introduce the notions of average reconstructability, average observability, and average detectability, respectively, and relate them to the convergence rate of the average observer. We also provide the design of average observer under these notions and their graph-theoretic interpretations with respect to inter-cluster and intra-cluster graph topologies of the clustered network system. Finally, in section 3.5, we provide some remarks on the relation of average reconstructability and average observability with scale-free networks, and a scaling property of open-loop estimation when the necessary and sufficient condition of average detectability is not satisfied.

Contents

In this chapter, we propose a minimum-order average observer that estimates the average states of clusters in a clustered network system and provide its design criteria. Several notions that allow asymptotic estimation of the average states are defined. For computational tractability, these notions are linked with the model of the projected network system.

3.1 Design Criteria of Average Observer

In this section, we provide a minimum-order average observer for a clustered network system and its design criteria as a necessary and sufficient condition that allows for the asymptotic estimation of the average states of clusters. Consider a clustered network system $\boldsymbol{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ as defined in section 2.1 of Chapter 2 and whose dynamics are given by the equation (2.3), where $\mathcal{V}_1 = \{\mu_1, \ldots, \mu_m\}$ is the set of m measured nodes and $\mathcal{Q} = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$ is the clustering of n unmeasured nodes $\mathcal{V}_2 = \{\nu_1, \ldots, \nu_n\}$. The characteristic matrix $Q \in \mathfrak{C}_{n,k} \subset \{0,1\}^{n\times k}$ of a clustering $\mathcal{Q}$ is defined in (2.5). One obtains a projected network system $\mathring{\boldsymbol{\Sigma}}_{\mathcal{V}_1,\mathcal{Q}}$ as in the equation (2.9) simply by aggregating the clusters $\mathcal{C}_1, \ldots, \mathcal{C}_k$ and considering them as supernodes. In fact, the projected network system describes the dynamics of the average state vector $\mathbf{z}_\mathsf{a}(t)$ of clusters, and is obtained by projecting the clustered network system on a lower dimensional state space with the aim of achieving computational tractability.

To estimate the average state vector $\mathbf{z}_\mathsf{a}(t)$ from the projected network system $\mathring{\boldsymbol{\Sigma}}_{\mathcal{V}_1,\mathcal{Q}}$ in real-time, we consider an average observer $\boldsymbol{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$ of dimension equal to the number of clusters in $\mathcal{Q}$, which takes the measurement vector $\mathbf{y}(t) \in \mathbb{R}^m$ and the known input $\mathbf{u}(t) \in \mathbb{R}^p$ as inputs and gives the estimated average state vector $\hat{\mathbf{z}}_\mathsf{a}(t) \in \mathbb{R}^k$ as an output. The average observer is given by

$$\boldsymbol{\Omega}_{\mathcal{V}_1,\mathcal{Q}} : \begin{cases} \dot{\mathbf{w}}(t) &= M\mathbf{w}(t) + K\mathbf{y}(t) + N\mathbf{u}(t) \\ \hat{\mathbf{z}}_\mathsf{a}(t) &= \mathbf{w}(t) + L\mathbf{y}(t) \end{cases} \tag{3.1}$$

where the matrices $M \in \mathbb{R}^{k\times k}$, $N \in \mathbb{R}^{k\times p}$, and $K, L \in \mathbb{R}^{k\times m}$ with k being the number of clusters. In the following, we provide the design criteria of average observer that is a necessary and sufficient condition for the asymptotic estimation of the average states.

Let us define the average estimation error

$$\boldsymbol{\zeta}(t) = \mathbf{z}_\mathsf{a}(t) - \hat{\mathbf{z}}_\mathsf{a}(t) \tag{3.2}$$

where $\mathbf{z}_\mathsf{a}(t) \in \mathbb{R}^k$ is the average state vector and $\hat{\mathbf{z}}_\mathsf{a}(t) \in \mathbb{R}^k$ is the estimated average state vector of clusters. Then, by taking a derivative on both sides, it follows

$$\begin{aligned} \dot{\boldsymbol{\zeta}}(t) &= \dot{\mathbf{z}}_\mathsf{a}(t) - \dot{\hat{\mathbf{z}}}_\mathsf{a}(t) \\ &= \dot{\mathbf{z}}_\mathsf{a}(t) - \dot{\mathbf{w}}(t) - L\dot{\mathbf{y}}(t) \end{aligned}$$

where $\dot{\mathbf{w}}(t)$ is given in (3.1) and, from (2.9) and (2.10),

$$\dot{\mathbf{z}}_\mathsf{a}(t) = Q^+A_{22}Q\mathbf{z}_\mathsf{a}(t) + Q^+A_{22}\boldsymbol{\sigma}(t) + Q^+A_{21}\mathbf{y}(t) + Q^+B_2\mathbf{u}(t) \tag{3.3}$$

and

$$\dot{\mathbf{y}}(t) = A_{11}\mathbf{y}(t) + A_{12}Q\mathbf{z}_{\mathtt{a}}(t) + A_{12}\boldsymbol{\sigma}(t) + B_1\mathbf{u}(t). \tag{3.4}$$

Furthermore, we add and subtract the term $(Q^+A_{22}Q - LA_{12}Q)\hat{\mathbf{z}}_{\mathtt{a}}(t)$ to the expression of $\dot{\boldsymbol{\zeta}}(t)$, which is equivalent to adding $(Q^+A_{22}Q - LA_{12}Q)(\mathbf{w} + L\mathbf{y} - \hat{\mathbf{z}}_{\mathtt{a}}(t))$ because $\hat{\mathbf{z}}_{\mathtt{a}}(t) = \mathbf{w}(t) + L\mathbf{y}(t)$, i.e.,

$$(Q^+A_{22}Q - LA_{12}Q)\hat{\mathbf{z}}_{\mathtt{a}} - (Q^+A_{22}Q - LA_{12}Q)\hat{\mathbf{z}}_{\mathtt{a}} \equiv (Q^+A_{22}Q - LA_{12}Q)(\mathbf{w} + L\mathbf{y} - \hat{\mathbf{z}}_{\mathtt{a}}).$$

Thus,

$$\begin{aligned}
\dot{\boldsymbol{\zeta}} &= Q^+A_{22}Q\mathbf{z}_{\mathtt{a}} + Q^+A_{21}\mathbf{y} + Q^+A_{22}\boldsymbol{\sigma} + Q^+B_2\mathbf{u} \\
&\quad -M\mathbf{w} - K\mathbf{y} - N\mathbf{u} - LA_{11}\mathbf{y} - LA_{12}Q\mathbf{z}_{\mathtt{a}} - LA_{12}\boldsymbol{\sigma} - LB_1\mathbf{u} \\
&\quad +(Q^+A_{22}Q - LA_{12}Q)(\mathbf{w} + L\mathbf{y} - \hat{\mathbf{z}}_{\mathtt{a}}) \\
&= (Q^+A_{22}Q - LA_{12}Q)(\mathbf{z}_{\mathtt{a}} - \hat{\mathbf{z}}_{\mathtt{a}}) + (Q^+A_{21} - LA_{11} - K)\mathbf{y} \\
&\quad +(Q^+B_2 - LB_1 - N)\mathbf{u} + (Q^+A_{22} - LA_{12})\boldsymbol{\sigma} \\
&\quad -M\mathbf{w} + (Q^+A_{22}Q - LA_{12}Q)(\mathbf{w} - L\mathbf{y})
\end{aligned}$$

where the dependence on t is omitted for brevity and should be considered implicit.

Define

$$R_L = Q^+A_{22} - LA_{12} \tag{3.5}$$

and consider

$$\begin{aligned}
M &= R_LQ \\
N &= Q^+B_2 - LB_1 \\
K &= Q^+A_{21} - LA_{11} + ML
\end{aligned} \tag{3.6}$$

then the dynamics of the average estimation error can be simplified to

$$\dot{\boldsymbol{\zeta}}(t) = R_L\left[Q\boldsymbol{\zeta}(t) + \boldsymbol{\sigma}(t)\right] \tag{3.7}$$

where $L \in \mathbb{R}^{k\times m}$ is the main design matrix of the average observer $\boldsymbol{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$ that should be chosen such that average state vector $\mathbf{z}_{\mathtt{a}}(t)$ can be estimated exponentially by the average observer $\boldsymbol{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$. More precisely, our goal is to find $L \in \mathbb{R}^{k\times m}$ such that the average estimation error $\boldsymbol{\zeta}(t)$ converges to $\mathbf{0}_k$ exponentially as $t \to \infty$, for any initial error $\boldsymbol{\zeta}_0 = \boldsymbol{\zeta}(0) \in \mathbb{R}^k$. Equivalently, there exists $a = a(\boldsymbol{\zeta}_0) > 0$ and $\gamma > 0$ such that $\|\boldsymbol{\zeta}(t)\| \leq ae^{-\gamma t}$.

Proposition 3.1. The average estimation error $\boldsymbol{\zeta}(t)$ converges to $\mathbf{0}_k$ exponentially as $t \to \infty$ for any $\boldsymbol{\sigma}(t) \in \ker(Q^+)$ *if and only if* there exists a matrix $L \in \mathbb{R}^{k\times m}$ such that both the following conditions are satisfied

(i) $\ker(R_L) \supseteq \ker(Q^+)$

(ii) R_LQ is Hurwitz

where R_L is given in (3.5) and $R_LQ = M$ by (3.6).

Proof of sufficiency. Assume (i) holds, then $R_L\boldsymbol{\sigma} \equiv \mathbf{0}_k$ since $\boldsymbol{\sigma}(t) \in \ker(Q^+)$ for all $t \in \mathbb{R}_{\geq 0}$. The dynamics of the average estimation error are then given by $\dot{\boldsymbol{\zeta}}(t) = R_LQ\boldsymbol{\zeta}(t)$. Assume (ii) holds, then the solution $\boldsymbol{\zeta}(t) = \exp(R_LQt)\boldsymbol{\zeta}(0)$ is such that, for every $\boldsymbol{\zeta}_0 =$

$\boldsymbol{\zeta}(0) \in \mathbb{R}^k$, there exists $a = a(\boldsymbol{\zeta}_0) \in (0, \infty)$ and $\gamma > 0$ such that $\|\boldsymbol{\zeta}(t)\| \leq ae^{-\gamma t}$. Thus, $\boldsymbol{\zeta}(t) \to \mathbf{0}_k$ as $t \to \infty$.

Proof of necessity. Assume (i) holds but (ii) does not hold. Then, the dynamics $\dot{\boldsymbol{\zeta}}(t) = R_L Q \boldsymbol{\zeta}(t)$ is unstable and $\boldsymbol{\zeta}(t) \to \infty$ as $t \to \infty$, which proves the necessity of (ii). Secondly, to prove the necessity of (i), assume (ii) holds but (i) does not hold. Then, the solution of (3.7) is given by

$$\boldsymbol{\zeta}(t) = \exp(R_L Q t)\boldsymbol{\zeta}(0) + \int_0^t \exp[R_L Q(t-\tau)] R_L \boldsymbol{\sigma}(\tau) d\tau$$

where the average deviation vector $\boldsymbol{\sigma}(t) \in \mathbb{R}^n$ is neither equal to $\mathbf{0}_k$ nor does it converge to $\mathbf{0}_k$ necessarily. Moreover, $\exp(R_L Q t)$ is always non-singular for all t and, since (i) does not hold by assumption, there exists a time interval (t_1, t_2) such that $R_L \boldsymbol{\sigma}(t^*) \neq \mathbf{0}_k$, for every $t^* \in (t_1, t_2)$. Therefore,

$$\int_0^t \exp[R_L Q(t-\tau)] R_L \boldsymbol{\sigma}(\tau) d\tau \neq \mathbf{0}_k$$

and $\limsup_{t\to\infty} \|\boldsymbol{\zeta}(t)\| \neq 0$, which concludes the proof.

□

The above proposition gives the design criteria of the average observer $\boldsymbol{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$. In particular, the matrix L of $\boldsymbol{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$ should be chosen such that it cancels out the effect of average deviation vector $\boldsymbol{\sigma}(t)$ and stabilizes the dynamics of average estimation error $\boldsymbol{\zeta}(t)$ given in (3.7). If L satisfies both criteria, then we are in the ideal case where the average observer asymptotically estimates the average states of clusters. On the other hand, if L fails to satisfy the first criterion, then we are in a non-ideal case where our goal is to design L such that $\limsup_{t\to\infty} \|\boldsymbol{\zeta}(t)\|$ is minimum in a sense that will be defined precisely in Chapter 4. Finally, if L fails to satisfy the second criterion, then the average observer is unstable and the average states cannot be estimated.

The second criterion, Proposition 3.1 (ii), holds if and only if the pair $(A_{12}Q, Q^+A_{22}Q)$ is detectable, i.e.,

$$\texttt{rank}\left(\begin{bmatrix} sI_k - Q^+A_{22}Q \\ A_{12}Q \end{bmatrix}\right) = k$$

for all $s \in \mathbb{C}_{\geq 0}$. In other words, the above equality is satisfied for all the marginally stable and the unstable eigenvalues of $Q^+A_{22}Q$. If the above rank is deficient for some eigenvalues of $Q^+A_{22}Q$, then those eigenvalues must be stable, i.e., in the open left half complex plane $\mathbb{C}_{<0}$. If the second criterion is not satisfied, then there is no hope and the average estimation error $\boldsymbol{\zeta}(t)$ grows unboundedly.

3.2 Average Reconstructability

The notion of average reconstructability is defined through the convergence of average observer to the true average states of clusters at an arbitrary rate. In general, the notion of 'reconstructability' allows for the reconstruction of the current state of the system from the knowledge of its past output and input [Antsaklis2006, Chapter 5]. Similarly, we define average reconstructability as a notion that allows for the reconstruction of current average states of the clusters from the knowledge of past output and input of the projected network

system. We provide a necessary and sufficient condition for average reconstructability and give it a graph-theoretical interpretation through the induced graph structure between the measured and the clusters of unmeasured nodes. Finally, we provide a design of average observer under average reconstructability.

3.2.1 Necessary and sufficient condition

The average state vector $\mathbf{z}_{\mathtt{a}}(t)$ can be estimated exponentially or reconstructed from the past output $\mathbf{y}(\tau)$ and input $\mathbf{u}(\tau)$ of the projected network system $\mathring{\mathbf{\Sigma}}_{\mathcal{V}_1,\mathcal{Q}}$, for $\tau \in [0,t]$, by employing the average observer $\mathbf{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$ if the average estimation error $\boldsymbol{\zeta}(t) = \mathbf{z}_{\mathtt{a}}(t) - \hat{\mathbf{z}}_{\mathtt{a}}(t)$, for some constants $a, \gamma > 0$, satisfies $\|\boldsymbol{\zeta}(t)\| \leq ae^{-\gamma t}$. That is to say that the estimated average state vector $\hat{\mathbf{z}}_{\mathtt{a}}(t)$ obtained from the average observer $\mathbf{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$ exponentially converges to the original average state vector $\mathbf{z}_{\mathtt{a}}(t)$ as $t \to \infty$.

Average reconstructability allows for the exponential convergence of the average observer but with an arbitrary rate. What we mean by the arbitrary rate of convergence is that, for any arbitrary $\gamma > 0$, there exists the design matrix $L \in \mathbb{R}^{k\times m}$ of the average observer $\mathbf{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$ such that, for any initial error $\boldsymbol{\zeta}_0 = \boldsymbol{\zeta}(0) \in \mathbb{R}^k$ and some constant $a = a(\boldsymbol{\zeta}_0) > 0$, the average estimation error satisfies $\|\boldsymbol{\zeta}(t)\| \leq ae^{-\gamma t}$.

Definition 3.1. The clustered network system $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is said to be *average reconstructable* if—for every $t \in \mathbb{R}_{\geq 0}$, $\mathbf{z}_{\mathtt{a}}(0) \in \mathbb{R}^k$, and $\boldsymbol{\sigma}(t) \in \mathbb{R}^n$ such that $Q^+\boldsymbol{\sigma} \equiv \mathbf{0}_k$—the average state vector $\mathbf{z}_{\mathtt{a}}(t)$ can be estimated exponentially with an arbitrary rate $\gamma > 0$ by the average observer $\mathbf{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$.

In the theorem below, we suppose that the set of measured nodes $\mathcal{V}_1 = \{\mu_1, \ldots, \mu_m\}$ and the clustering $\mathcal{Q} = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$ of the set of unmeasured nodes $\mathcal{V}_2 = \{\nu_1, \ldots, \nu_n\}$ are given for $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$. Furthermore, the system matrices of $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ are partitioned as in (2.4), and the characteristic matrix Q of $\mathcal{Q}$ and its left pseudo-inverse Q^+ are defined in (2.5) and (2.6), respectively.

Theorem 3.2. The clustered network system $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average reconstructable *if and only if*

$$\mathtt{rank}\left(\begin{bmatrix} A_{12} \\ Q^+A_{22} \\ Q^+ \end{bmatrix}\right) = \mathtt{rank}(A_{12}). \tag{3.8}$$

Proof of sufficiency. We prove that if (3.8) holds then $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average reconstructable. Assume (3.8) holds, which implies that, for any arbitrary matrix $V \in \mathbb{R}^{k\times k}$,

$$\mathtt{rank}\left(\begin{bmatrix} A_{12} \\ Q^+A_{22} - VQ^+ \end{bmatrix}\right) = \mathtt{rank}(A_{12}).$$

Thus,

$$L = (Q^+A_{22} - VQ^+)A_{12}^{\dagger} \tag{3.9}$$

is a solution to $LA_{12} = Q^+A_{22} - VQ^+$. Such a choice of L implies $R_L = VQ^+$, which gives $R_L\boldsymbol{\sigma} = VQ^+\boldsymbol{\sigma} \equiv \mathbf{0}_k$ and $R_LQ = V$. Therefore, the average estimation error $\boldsymbol{\zeta}(t) = \exp(Vt)\boldsymbol{\zeta}_0$, where $\boldsymbol{\zeta}_0 = \boldsymbol{\zeta}(0) \in \mathbb{R}^k$. Let $V = -\gamma I_k$, for some arbitrary $\gamma > 0$. Then, $\|\boldsymbol{\zeta}(t)\| \leq \|\boldsymbol{\zeta}_0\|e^{-\gamma t}$, which implies average reconstructability of $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$.

Proof of necessity. We prove that if $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average reconstructable then (3.8) holds. Assume $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average reconstructable, i.e., for any arbitrary $\gamma > 0$, the average estimation error satisfies $\|\boldsymbol{\zeta}(t)\| \leq \|\boldsymbol{\zeta}_0\|e^{-\gamma t}$. This means that the design matrix $L \in \mathbb{R}^{k\times m}$

of the average observer $\mathbf{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$ can be chosen such that the eigenvalues of $M = R_L Q$ can be assigned arbitrarily and $R_L \boldsymbol{\sigma} \equiv \mathbf{0}_k$. Let $L_1, L_2 \in \mathbb{R}^{k\times m}$ be such that $M_1 = R_{L_1} Q$ and $M_2 = R_{L_2} Q$ are two different Hurwitz matrices and $R_{L_1}\boldsymbol{\sigma} = R_{L_2}\boldsymbol{\sigma} \equiv \mathbf{0}_k$. Moreover, since M_1 and M_2 are arbitrary, we can choose L_1 and L_2 such that i-th row of M_1 is linearly independent from the i-th row of M_2, for $i = 1, \ldots, k$. Since $R_{L_1}\boldsymbol{\sigma} = R_{L_2}\boldsymbol{\sigma} \equiv \mathbf{0}_k$, therefore $L_1 A_{12} = Q^+ A_{22} - V_1 Q^+$ and $L_2 A_{12} = Q^+ A_{22} - V_2 Q^+$, where $V_1 = M_1$ and $V_2 = M_2$, which implies

$$\mathtt{rank}\left(\begin{bmatrix} A_{12} \\ Q^+A_{22} - M_1Q^+ \\ Q^+A_{22} - M_2Q^+ \end{bmatrix}\right) = \mathtt{rank}(A_{12}). \tag{3.10}$$

Using the fact that $\mathtt{rank}(M_1) = \mathtt{rank}(M_2) = k$ because they are Hurwitz, and that the corresponding rows of M_1 and M_2 are linearly independent, we show that

$$\mathtt{rank}\left(\begin{bmatrix} I_k & -M_1 \\ I_k & -M_2 \end{bmatrix}\right) = 2k.$$

That is,

$$\begin{bmatrix} I_k & -M_1 \\ I_k & -M_2 \end{bmatrix}$$

is invertible, which implies that

$$\begin{bmatrix} I_k & 0_{k\times k} & 0_{k\times k} \\ 0_{k\times k} & I_k & -M_1 \\ 0_{k\times k} & I_k & -M_2 \end{bmatrix}$$

is invertible. Finally, notice that

$$\begin{bmatrix} A_{12} \\ Q^+A_{22} - M_1Q^+ \\ Q^+A_{22} - M_2Q^+ \end{bmatrix} = \begin{bmatrix} I_k & 0_{k\times k} & 0_{k\times k} \\ 0_{k\times k} & I_k & -M_1 \\ 0_{k\times k} & I_k & -M_2 \end{bmatrix} \begin{bmatrix} A_{12} \\ Q^+A_{22} \\ Q^+ \end{bmatrix}$$

therefore,

$$\mathtt{rank}\begin{bmatrix} A_{12} \\ Q^+A_{22} - M_1Q^+ \\ Q^+A_{22} - M_2Q^+ \end{bmatrix} = \mathtt{rank}\begin{bmatrix} A_{12} \\ Q^+A_{22} \\ Q^+ \end{bmatrix}. \tag{3.11}$$

From (3.10) and (3.11), we obtain (3.8), which concludes the proof.

□

3.2.2 Graph-theoretic interpretation of average reconstructability

Recall the measured nodes $\mathcal{V}_1 = \{\mu_1, \ldots, \mu_m\}$, the unmeasured nodes $\mathcal{V}_2 = \{\nu_1, \ldots, \nu_n\}$, and the induced bipartite subgraphs $\mathcal{G}_{\mu 1}, \ldots, \mathcal{G}_{\mu k}$ that capture the graph topology from clusters $\mathcal{C}_1, \ldots, \mathcal{C}_k$ to the measured nodes, respectively. That is, for $\alpha \in \{1, \ldots, k\}$, each edge $(\mu_i, \nu_j) \in \mathcal{E}_{\mu\alpha}$ in $\mathcal{G}_{\mu\alpha}$ is such that $\mu_i \in \mathcal{V}_1$ and $\nu_j \in \mathcal{C}_\alpha \subset \mathcal{V}_2$. The edges in subgraph $\mathcal{G}_{\mu\alpha}$ are the directed arcs from the nodes in cluster $\mathcal{C}_\alpha$ to their out-neighbors that are measured nodes $\mathcal{V}_1$.

Definition 3.2. The set of measured nodes $\mathcal{V}_1$ is said to *span* the clustering $\mathcal{Q} = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$ of unmeasured nodes if, for every $\alpha \in \{1, \ldots, k\}$ and for every $\nu_j \in \mathcal{C}_\alpha$, there exists $\mu_i \in \mathcal{V}_1$ such that $(\mu_i, \nu_j) \in \mathcal{E}_{\mu\alpha}$ is an edge of $\mathcal{G}_{\mu\alpha} = (\mathcal{V}_1, \mathcal{C}_\alpha, \mathcal{E}_{\mu\alpha})$.

In other words, all the unmeasured nodes have at least one out-neighbor that is a measured node.

Corollary 3.2.1. If the clustered network system $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average reconstructable, then the set of measured nodes $\mathcal{V}_1$ spans the clustering $\mathcal{Q}$ of unmeasured nodes.

Proof. Assume $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average reconstructable, then, by (3.8), the rows of Q^+ are linearly dependent to the rows of $A_{12} \in \mathbb{R}^{m\times n}_{\geq 0}$. This implies that no column of A_{12} is equal to a zero vector because no column of Q^+ is equal to a zero vector and having a zero column in A_{12} contradicts (3.8). Furthermore, for $\mu_i \in \mathcal{V}_1$ and $\nu_j \in \mathcal{V}_2$, we have $[A_{12}]_{ij} > 0$ if and only if $(\mu_i, \nu_j) \in \mathcal{E}$. Therefore, for every $\alpha \in \{1, \ldots, k\}$, A_{12} captures the edge configurations of the induced bipartite subgraph $\mathcal{G}_{\mu\alpha} = (\mathcal{V}_1, \mathcal{C}_\alpha, \mathcal{E}_{\mu\alpha})$ because $\mathcal{E}_{\mu\alpha} = \mathcal{E} \cap (\mathcal{V}_1 \times \mathcal{C}_\alpha)$. Since the columns of A_{12} correspond to the unmeasured nodes and no column of A_{12} is equal to a zero vector, therefore, for every $\nu_j \in \mathcal{V}_2 \cap \mathcal{C}_\alpha$, there exists $\mu_i \in \mathcal{V}_1$ such that $[A_{12}]_{ij} > 0$. Therefore, $\mathcal{V}_1$ spans $\mathcal{Q}$.

□

(a) Example with a single cluster of unmeasured nodes (shown as blue)

(b) Example with three clusters of unmeasured nodes (shown as blue, green, and orange)

Figure 3.1: Examples of clustered network systems, where the measured nodes (shown as black) span the clustering of unmeasured nodes.

Example 3.1. Consider the examples of Figure 3.1 with a single cluster and three clusters of unmeasured nodes, respectively. For Figure 3.1(a), we have $\mathcal{Q} = \mathcal{C}_1 = \mathcal{V}_2$ and $Q^+ = \frac{1}{16}\mathbf{1}_{16}$. Suppose an unweighted digraph, then

$$A_{12} = \begin{bmatrix} 1&1&0&0&0&0&0&0&1&1&0&0&0&0&0&0 \\ 0&0&1&1&0&0&0&0&0&0&1&1&0&0&0&0 \\ 0&0&0&0&1&1&0&0&0&0&0&0&1&1&0&0 \\ 0&0&0&0&0&0&1&1&0&0&0&0&0&0&1&1 \end{bmatrix}. \tag{3.12}$$

Notice that $\mathcal{V}_1$ spans $\mathcal{Q}$ since all the unmeasured nodes have a measured node as an out-neighbor. For Figure 3.1(b), we have $\mathcal{Q} = \{\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3\}$, where $\mathcal{C}_1 = \{\nu_1, \ldots, \nu_8\}$ is shown in blue, $\mathcal{C}_2 = \{\nu_9, \ldots, \nu_{16}\}$ is shown in orange, and $\mathcal{C}_3 = \{\nu_{17}, \ldots, \nu_{23}\}$ is shown in green.

We have $Q^+ = \texttt{diag}\left(\frac{1}{8}\mathbf{1}_8^\intercal, \frac{1}{8}\mathbf{1}_8^\intercal, \frac{1}{7}\mathbf{1}_7^\intercal\right)$ and

$$A_{12} = \begin{bmatrix} 1\,1\,1\,1\,1\,1\,1\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0\,0\,1\,1\,1\,1\,1\,1\,1\,1\,0\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,1\,1\,1\,1\,1\,1 \end{bmatrix}. \tag{3.13}$$

Again, $\mathcal{V}_1$ spans $\mathcal{Q}$. On the other hand, the clustered network system in the example of Figure 2.1 is not average reconstructable since the measured nodes of that system do not span the clustering.

3.2.3 Design of average observer under average reconstructability

Since average reconstructability allows for an arbitrary rate of convergence of the average observer, the design of average observer under average reconstructability is flexible and is given as follows. Choose $V \in \mathbb{R}^{k\times k}$ to be any Hurwitz matrix, then, as in (3.9), we have

$$\begin{aligned} M &= V \\ N &= Q^+B_2 - LB_1 \\ K &= Q^+A_{21} - LA_{11} + ML \\ L &= (Q^+A_{22} - VQ^+)A_{12}^+ \end{aligned} \tag{3.14}$$

which ensures $\|\boldsymbol{\zeta}(t)\| \leq \|\boldsymbol{\zeta}_0\|e^{-\gamma t}$, where the rate of convergence $\gamma > 0$ is determined by the eigenvalue of V with smallest magnitude.

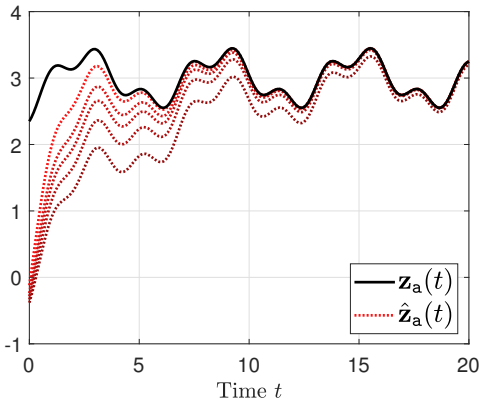


Figure 3.2: Average estimation of the single cluster of unmeasured nodes in the graph of Figure 3.1(a) by the average observer $\boldsymbol{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$.

Example 3.2. Consider a linear flow network, [Walter1999], with the dynamics at each node i given by

$$\dot{x}_i(t) = \sum_{j\in\mathcal{N}_{i\leftarrow\mathcal{V}}} a_{ij}x_j(t) - \sum_{o\in\mathcal{N}_{i\rightarrow\mathcal{V}}} a_{oi}x_i(t) + \sum_{l=1}^{2} b_{il}u_l(t), \tag{3.15}$$

where $a_{ij} = 1$ and $a_{oi} = 1$ for all $i \in \mathcal{V}$, $j \in \mathcal{N}_{i\leftarrow\mathcal{V}}$, and $o \in \mathcal{N}_{i\rightarrow\mathcal{V}}$. Note that $b_{il} = 1$ if $u_l(t)$ is applied on i, and $b_{il} = 0$ otherwise. Let the system be defined on the graph shown in Figure 3.1(a), where the measured nodes are shown as black and the unmeasured nodes

as blue. We consider a single cluster of unmeasured nodes, i.e., $Q = \mathbf{1}_{16}$, and estimate its average state $\mathbf{z_a}(t) = z_1(t)$. Let the input vector be

$$\mathbf{u}(t) = \begin{bmatrix} u_1(t) \\ u_2(t) \end{bmatrix} = \begin{bmatrix} \sin(t) + 5 \\ \sin(3t) \end{bmatrix}$$

and the input matrix $B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}$ with $B_1 = 0_{4\times 2}$ and $B_2 = \begin{bmatrix} \mathbf{1}_8 & \mathbf{0}_8 \\ \mathbf{0}_8 & \mathbf{1}_8 \end{bmatrix}$. The output matrix $C = \begin{bmatrix} I_4 & 0 \end{bmatrix}$. Note that $Q^+A_{22} = -\frac{1}{16}\mathbf{1}_{16}^\intercal = -Q^+$ since the relative outflow centrality of each blue node with respect to the cluster $\mathcal{C}_1$ is -1. Thus, the condition (3.8) is satisfied, where A_{12} is given in (3.12). The design of average observer $\mathbf{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$, for $V = -0.75$, is computed from (3.14) as

$$\begin{aligned} M &= -0.75, \quad N = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \\ K &= \tfrac{3}{256}\mathbf{1}_4^\intercal, \quad L = -\tfrac{1}{64}\mathbf{1}_4^\intercal. \end{aligned}$$

The average state can be estimated asymptotically by the observer $\mathbf{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$, as shown in Figure 3.2, where the solid line shows the average state trajectory of the cluster of unmeasured nodes and the dotted line shows the estimation of the average state at different rates. The rate of convergence is arbitrary and can be set by choosing different values $V = -0.75, -0.5, -0.4, -0.25, -0.1$ from faster to slower, respectively, as shown in the figure.

3.3 Average Observability

The notion of average observability is defined through the projected network system. In general, the notion of 'observability' allows for the reconstruction of current state of the system from the knowledge of its future output and input [Antsaklis2006, Chapter 5]. Therefore, we define average observability as a notion that allows for the reconstruction of current average states of the clusters from the knowledge of future output and input of the projected network system. Precisely, we say that a clustered network system is average observable if, by taking sufficient derivatives of the output and inputs of the projected network system, one can uniquely obtain the average state vector of clusters.

Note that the notions of 'reconstructability' and 'observability' are equivalent for linear time-invariant systems with known inputs [Antsaklis2006, Chapter 5]. However, in the projected network system, we have the average deviation vector $\boldsymbol{\sigma}(t)$ as an unknown input, which makes the two notions of average reconstructability and average observability not equivalent. This shall be clarified by the necessary and sufficient condition of average observability. Finally, similar to average reconstructability, we provide a graph-theoretic interpretation of average observability and present a design of average observer when its necessary and sufficient condition is satisfied.

3.3.1 Necessary and sufficient condition

Given a clustered network system $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ with a set of measured nodes $\mathcal{V}_1$ and a clustering $\mathcal{Q} = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$ of unmeasured nodes $\mathcal{V}_2$, we study the conditions under which the average

state vector $\mathbf{z}_{\mathtt{a}}(t)$ can be uniquely obtained from the output and input of the projected network system $\mathring{\mathbf{\Sigma}}_{\mathcal{V}_1,\mathcal{Q}}$.

We say that $\mathbf{z}_{\mathtt{a}}(t)$ can be uniquely obtained from the projected network system

$$\mathring{\mathbf{\Sigma}}_{\mathcal{V}_1,\mathcal{Q}}: \begin{cases} \dot{\mathbf{z}}(t) &= E\mathbf{z}(t) + F\boldsymbol{\sigma}(t) + G\mathbf{u}(t) \\ \mathbf{0}_k &= Q^{+}\boldsymbol{\sigma}(t) \\ \mathbf{y}(t) &= H\mathbf{z}(t) \end{cases}$$

if the knowledge of future output $\mathbf{y}(\tau) = H\mathbf{z}(t)$ and input $\mathbf{u}(\tau)$, for $\tau \in [t, t+\varepsilon)$ with some 'small' $\varepsilon > 0$, is sufficient to obtain $\mathbf{z}_{\mathtt{a}}(t)$. By a small ε, we mean a small interval after t required to compute k derivatives of the output $\mathbf{y}(t)$ of $\mathring{\mathbf{\Sigma}}_{\mathcal{V}_1,\mathcal{Q}}$, which gives the following system of equations:

$$\mathcal{Z}\mathbf{z}_{\mathtt{a}}(t) + \mathcal{Y}\bar{\mathbf{y}}_k(t) + \mathcal{U}\bar{\mathbf{u}}_{k-1}(t) + \mathcal{S}\bar{\boldsymbol{\sigma}}_{k-1}(t) = 0 \tag{3.16}$$

where

$$\bar{\mathbf{y}}_k(t) = \begin{bmatrix} \mathbf{y}(t) \\ \dot{\mathbf{y}}(t) \\ \vdots \\ \mathbf{y}^{(k)}(t) \end{bmatrix}, \quad \bar{\mathbf{u}}_{k-1}(t) = \begin{bmatrix} \mathbf{u}(t) \\ \dot{\mathbf{u}}(t) \\ \vdots \\ \mathbf{u}^{(k-1)}(t) \end{bmatrix}, \quad \bar{\boldsymbol{\sigma}}_{k-1}(t) = \begin{bmatrix} \boldsymbol{\sigma}(t) \\ \dot{\boldsymbol{\sigma}}(t) \\ \vdots \\ \boldsymbol{\sigma}^{(k-1)}(t) \end{bmatrix}$$

and

$$\mathcal{Z} = \begin{bmatrix} E_{12} \\ E_{12}E_{22} \\ \vdots \\ E_{12}E_{22}^{k-1} \end{bmatrix}$$

$$\mathcal{Y} = \begin{bmatrix} E_{11} & -I_m & 0 & \dots & & & 0 \\ E_{12}E_{21} & E_{11} & -I_m & & & & \vdots \\ E_{12}E_{22}E_{21} & E_{12}E_{21} & E_{11} & -I_m & & \ddots & \\ \vdots & \vdots & & & \ddots & & 0 \\ & & & & & & \\ E_{12}E_{22}^{k-2}E_{21} & E_{12}E_{22}^{k-3}E_{21} & \dots & & E_{12}E_{21} & E_{11} & -I_m \end{bmatrix}$$

$$\mathcal{U} = \begin{bmatrix} G_1 & 0 & \dots & & 0 \\ E_{12}G_2 & G_1 & \ddots & & \vdots \\ E_{12}E_{22}G_2 & E_{12}G_2 & G_1 & & \\ \vdots & & \ddots & \ddots & 0 \\ & & & & \\ E_{12}E_{22}^{k-2}G_2 & \dots & & E_{12}G_2 & G_1 \end{bmatrix}$$

$$\mathcal{S} = \begin{bmatrix} F_1 & 0 & \dots & & 0 \\ E_{12}F_2 & F_1 & \ddots & & \vdots \\ E_{12}E_{22}F_2 & E_{12}F_2 & F_1 & & \\ \vdots & & \ddots & \ddots & 0 \\ & & & & \\ E_{12}E_{22}^{k-2}F_2 & \dots & & E_{12}F_2 & F_1 \end{bmatrix}$$

with the matrices E_{ij}, F_i, G_i defined in (2.10). The average state vector $\mathbf{z}_\mathtt{a}(t)$ can be uniquely obtained from (3.16) if and only if there exists a matrix $\mathcal{X} \in \mathbb{R}^{mk \times mk}$ that satisfies both the following conditions:

(i) $\mathtt{rank}(\mathcal{X}\mathcal{Z}) = k$

(ii) $\mathcal{X}\mathcal{S}\bar{\boldsymbol{\sigma}}_{k-1} = \mathbf{0}$.

The sufficiency of this claim is straightforward because if such a matrix $\mathcal{X}$ exists, then the solution to (3.16) is given by

$$\mathbf{z}_\mathtt{a}(t) = -(\mathcal{X}\mathcal{Z})^+\mathcal{X}(\mathcal{Y}\bar{\mathbf{y}}_k(t) + \mathcal{U}\bar{\mathbf{u}}_{k-1}(t)).$$

To prove necessity, we first assume that (i) holds but (ii) does not hold. Then, $\bar{\boldsymbol{\sigma}}_{k-1}(t)$ cannot be canceled out from (3.16), thus the solution $\mathbf{z}_\mathtt{a}(t)$ does not exist in terms of $\bar{\mathbf{y}}_k(t)$ and $\bar{\mathbf{u}}_{k-1}(t)$. Second, assume that (ii) holds but (i) does not hold, i.e., $\mathcal{X}\mathcal{Z}$ does not have full column rank, then the solution $\mathbf{z}_\mathtt{a}(t)$ exists but is not unique. Therefore, the average state vector $\mathbf{z}_\mathtt{a}(t)$ can be uniquely obtained from $\mathring{\boldsymbol{\Sigma}}_{\mathcal{V}_1,\mathcal{Q}}$ if the solution $\mathbf{z}_\mathtt{a}(t)$ to the equation (3.16) exists, in terms of $\bar{\mathbf{y}}_k(t)$ and $\bar{\mathbf{u}}_{k-1}(t)$, and is unique.

In the following, the notion of average observability is defined in the sense described above.

Definition 3.3. The clustered network system $\boldsymbol{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is said to be *average observable* if—for every $t \in \mathbb{R}_{\geq 0}$, $\mathbf{z}_\mathtt{a}(0) \in \mathbb{R}^k$, and $\boldsymbol{\sigma}(t) \in \mathbb{R}^n$ such that $Q^+\boldsymbol{\sigma} \equiv \mathbf{0}_k$—the average state vector $\mathbf{z}_\mathtt{a}(t)$ can be uniquely obtained from the projected network system $\mathring{\boldsymbol{\Sigma}}_{\mathcal{V}_1,\mathcal{Q}}$ as a solution of (3.16).

For the following theorem, we consider a clustered network system $\boldsymbol{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ with the set of m measured nodes $\mathcal{V}_1$ and the clustering $\mathcal{Q} = \{\mathcal{C}_1, \dots, \mathcal{C}_k\}$ of n unmeasured nodes $\mathcal{V}_2$. Recall that the system matrices of $\boldsymbol{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ are partitioned as in (2.4), and the characteristic matrix Q of $\mathcal{Q}$ and its left pseudo-inverse Q^+ are defined in (2.5) and (2.6), respectively.

Theorem 3.3. The clustered network system $\boldsymbol{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average observable *if and only if*

$$\mathtt{rank}\left(\begin{bmatrix} A_{12} \\ Q^+ \end{bmatrix}\right) = \mathtt{rank}(A_{12}). \tag{3.17}$$

In other words, $\mathtt{ker}(Q^+) \supseteq \mathtt{ker}(A_{12})$.

Proof of sufficiency. Assume (3.17) holds, which means that the rows of Q^+ are in the rowspace of A_{12}. In other words, there exists a matrix $X \in \mathbb{R}^{k \times m}$ such that $XA_{12} = Q^+$ and $X = Q^+A_{12}^\dagger$, where $A_{12}^\dagger = A_{12}^\intercal(A_{12}A_{12}^\intercal)^{-1}$ is the right pseudo-inverse of A_{12}, i.e., $A_{12}A_{12}^\dagger = I_m$. Since $\mathbf{y}(t) = H\mathbf{z}(t) = \mathbf{x}_1(t)$, we consider

$$\dot{\mathbf{y}}(t) = A_{11}\mathbf{y}(t) + A_{12}Q\mathbf{z}_\mathtt{a}(t) + A_{12}\boldsymbol{\sigma}(t) + B_1\mathbf{u}(t)$$

from the projected system $\mathring{\mathbf{\Sigma}}_{\mathcal{V}_1,\mathcal{Q}}$ in (2.10). Then, by multiplying by $X = Q^+ A_{12}^\dagger$ on both sides, we can reconstruct the average state vector

$$\mathbf{z}_\mathtt{a}(t) = X\dot{\mathbf{y}}(t) - XA_{11}\mathbf{y}(t) - XB_1\mathbf{u}(t)$$

where $XA_{12}\boldsymbol{\sigma} = Q^+\boldsymbol{\sigma} \equiv 0$. Thus, if (3.17) is satisfied, we can uniquely obtain $\mathbf{z}_\mathtt{a}(t)$ from the knowledge of $\dot{\mathbf{y}}(t)$, $\mathbf{y}(t)$, and $\mathbf{u}(t)$.

Proof of necessity. Assume $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average observable but (3.17) does not hold. However, notice that, in (3.16), there is a term $F_1\boldsymbol{\sigma}^{(i-1)}(t)$, where $F_1 = A_{12}$, in each expression corresponding to $\mathbf{y}^{(i)}(t) - E_{11}\mathbf{y}^{(i-1)}(t)$, for all $i = 1, \ldots, k$, which cannot be canceled out in order to obtain $\mathbf{z}_\mathtt{a}(t)$. This is because there does not exist any matrix X such that $XA_{12} = Q^+$. Thus, we arrive at a contradiction and, in fact, $\mathbf{\Sigma}$ is not average observable if (3.17) does not hold.

□

From the necessary and sufficient conditions of average observability in (3.17) and average reconstructability in (3.8), we obtain the following corollaries.

Corollary 3.3.1. The clustered network system $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average reconstructable if and only if both the following hold:

(i) $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average observable

(ii) $\ker(Q^+A_{22}) \supseteq \ker(A_{12})$.

Proof. The proof follows directly from (3.8) and (3.17).

□

The above corollary implies that average observability is necessary for average reconstructability, but not vice versa. In other words, convergence of average observer $\mathbf{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$ at an arbitrary rate implies average observability. However, average observability may not imply the convergence of $\mathbf{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$ at an arbitrary rate because the condition in Corollary 3.3.1(ii) may not hold.

If $\ker(Q^+) \supseteq \ker(F)$, where F is given in (2.10), then, and only then, for any arbitrary $W \in \mathbb{R}^{(m+k)\times k}$, there exists a matrix $N \in \mathbb{R}^{(m+k)\times(m+k)}$ such that $NF = WQ^+$. From this, we claim the following equivalence

$$NF = WQ^+ \quad \Leftrightarrow \quad NF\boldsymbol{\sigma} \equiv \mathbf{0}_k \text{ with } \boldsymbol{\sigma}(t) \in \ker(Q^+). \tag{3.18}$$

To prove, assume $NF = WQ^+$, then $NF\boldsymbol{\sigma} = WQ^+\boldsymbol{\sigma} \equiv \mathbf{0}_k$ for $\boldsymbol{\sigma}(t) \in \ker(Q^+)$. In the other direction, assume $NF\boldsymbol{\sigma} \equiv \mathbf{0}_k$, then it holds that $\ker(NF) \supseteq \ker(Q^+)$. Hence, for any arbitrary $W_1 \in \mathbb{R}^{(m+k)\times(m+k)}$, there exists $W_2 \in \mathbb{R}^{(m+k)\times k}$ such that $W_2Q^+ = W_1NF$. Therefore, by choosing $W = W_1^{-1}W_2$, we obtain $WQ^+ = NF$, which completes the proof of (3.18).

Using the above fact, we can multiply the state equation of the projected network system $\mathring{\mathbf{\Sigma}}_{\mathcal{V}_1,\mathcal{Q}}$ by N and obtain a singular projected network system $\mathring{\mathbf{\Sigma}}_N(\mathcal{V}_1, \mathcal{Q})$, where

$$\mathring{\mathbf{\Sigma}}_N(\mathcal{V}_1, \mathcal{Q}) : \begin{cases} N\dot{\mathbf{z}}(t) &= NE\mathbf{z}(t) + NG\mathbf{u}(t) \\ \mathbf{y}(t) &= H\mathbf{z}(t). \end{cases}$$

Thus, another way to obtain the average state vector $\mathbf{z}_\mathtt{a}(t)$ is through the observability of $\mathring{\mathbf{\Sigma}}_N(\mathcal{V}_1, \mathcal{Q})$ under the condition that $NF\boldsymbol{\sigma} \equiv 0$.

Corollary 3.3.2. If the clustered network system $\boldsymbol{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average observable, then there exists a matrix $N \in \mathbb{R}^{(m+k)\times(m+k)}$ such that both the following conditions hold:

(i) $NF\boldsymbol{\sigma} \equiv \mathbf{0}_k$

(ii) $\mathring{\boldsymbol{\Sigma}}_N(\mathcal{V}_1, \mathcal{Q})$ is observable.

Proof. Assume $\boldsymbol{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average observable, which is equivalent to (3.17) by Theorem 3.3. Then, for any arbitrary $W \in \mathbb{R}^{(m+k)\times k}$, there exists a matrix $X \in \mathbb{R}^{(m+k)\times m}$ such that if we choose $N = \begin{bmatrix} X & 0_{(m+k)\times k} \end{bmatrix} \in \mathbb{R}^{(m+k)\times(m+k)}$, we have $NF = XA_{12} = WQ^+$, where F is given in (2.10). Therefore, we have $NF\boldsymbol{\sigma} = WQ^+\boldsymbol{\sigma} \equiv \mathbf{0}_k$, which proves (i).

Second, to prove (ii), note that $\mathring{\boldsymbol{\Sigma}}_N(\mathcal{V}_1, \mathcal{Q})$ is observable if and only if, [Yip1981, Cobb1984, Bejarano2009, Bejarano2011],

$$\mathtt{rank}\left(\begin{bmatrix} sN - NE \\ H \end{bmatrix}\right) = m + k, \quad \forall s \in \mathbb{C}. \tag{3.19}$$

We can write $E = \begin{bmatrix} E_1 & FQ \end{bmatrix}$, where $E_1 = \begin{bmatrix} E_{11} \\ E_{21} \end{bmatrix}$, which gives

$$\begin{aligned} \mathtt{rank}\begin{bmatrix} sN - NE \\ H \end{bmatrix} &= \mathtt{rank}\begin{bmatrix} sX - NE_1 & -W \\ I_m & 0_{m\times k} \end{bmatrix} \\ &= m + \mathtt{rank}(W) \end{aligned}$$

for all $s \in \mathbb{C}$, where $N = \begin{bmatrix} X & 0_{(m+k)\times k} \end{bmatrix}$ and $NE = \begin{bmatrix} NE_1 & W \end{bmatrix}$. By (3.19), we need to prove that $\mathtt{rank}(W) = k$. Since W is arbitrary, we can choose $W = \begin{bmatrix} I_k & 0_{k\times m} \end{bmatrix}^\intercal$, which implies that $\mathtt{rank}(W) = k$ and that $X = Q^+A_{12}^\dagger$. Thus, $\mathring{\boldsymbol{\Sigma}}_N(\mathcal{V}_1, \mathcal{Q})$ is observable. □

3.3.2 Graph-theoretic interpretation of average observability

A graph-theoretic interpretation of average observability is similar to that of average reconstructability. For completeness, however, we present the following corollary.

Corollary 3.3.3. If the clustered network system $\boldsymbol{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average observable, then the set of measured nodes $\mathcal{V}_1$ spans the clustering $\mathcal{Q}$ of unmeasured nodes.

Proof. The proof is same as that of Corollary 3.2.1. □

The condition of Corollary 3.2.1 and 3.3.3 concerns the induced subgraphs $\mathcal{G}_{\mu\alpha}$, for $\alpha = 1, \ldots, k$. This is a necessary condition for average reconstructability and average observability. However, average reconstructability also depends on the graph topology of clusters $\mathcal{C}_1, \ldots, \mathcal{C}_k$ and local damping weights a_{ii} of unmeasured nodes because of the matrix Q^+A_{22} in (3.8). On the other hand, average observability depends solely on the topology of $\mathcal{G}_{\mu\alpha}$. For this reason, the examples illustrated in Figure 3.1 are average observable irrespective of the local damping weights a_{ii} of unmeasured nodes and the topology of clusters.

3.3.3 Design of average observer under average observability

We remarked that average observability may not ensure the exponential convergence of the average observer $\mathbf{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$ at an arbitrary rate. However, in this subsection, we show that average observability allows for an average observer where the average estimation error can be made arbitrarily small by choosing a gain parameter γ to be arbitrarily large.

Theorem 3.4. If the clustered network system $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average observable, then, for any $\gamma > 0$, there exists a design matrix $L = L_\gamma \in \mathbb{R}^{k\times m}$ of the average observer $\mathbf{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$ and increasing positive-valued functions $a(\,\cdot\,)$ and $b(\,\cdot\,)$ such that, for any $r > 0$, $\boldsymbol{\zeta}(0) = \boldsymbol{\zeta}_0 \in \mathbb{R}^k$ with $\|\boldsymbol{\zeta}_0\| \le r$, and $\|\boldsymbol{\sigma}\|_\infty \le \bar{\sigma} < \infty$, the average estimation error $\boldsymbol{\zeta}(t)$ in (3.7) satisfies

$$\|\boldsymbol{\zeta}(t)\| \le a(r)e^{-\gamma t} + b(\bar{\sigma})\frac{1-e^{-\gamma t}}{\gamma}. \tag{3.20}$$

In particular, the matrix $L = L_\gamma$ is given by

$$L_\gamma = (Q^+A_{22}Q + \gamma I_k)Q^+A_{12}^\dagger. \tag{3.21}$$

Proof. If $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average observable, then, by Theorem 3.3, we have $\ker(Q^+) \supseteq \ker(A_{12})$. This implies that the solution to $LA_{12} = VQ^+$, for any $V \in \mathbb{R}^{k\times k}$, is given by $L = VQ^+A_{12}^\dagger$, where $Q^+ = (Q^\intercal Q)^{-1}Q^\intercal$ and $A_{12}^\dagger = A_{12}^\intercal(A_{12}A_{12}^\intercal)^{-1}$. Let $V = Q^+A_{22}Q + \gamma I_k$ for an arbitrary $\gamma > 0$. Then, from (3.5), we have $R_L = Q^+A_{22} - VQ^+$ and $R_LQ = -\gamma I_k$. And the dynamics of average estimation error $\boldsymbol{\zeta}(t)$ in (3.7) can be written as

$$\dot{\boldsymbol{\zeta}}(t) = -\gamma\boldsymbol{\zeta}(t) + Q^+A_{22}\boldsymbol{\sigma}(t)$$

which implies

$$\begin{aligned}
\|\boldsymbol{\zeta}(t)\| &\le \|\boldsymbol{\zeta}(0)\|e^{-\gamma t} + \|\int_0^t e^{-\gamma\tau}Q^+A_{22}\boldsymbol{\sigma}(t-\tau)d\tau\| \\
&\le \|\boldsymbol{\zeta}(0)\|e^{-\gamma t} + \|\int_0^t e^{-\gamma\tau}d\tau\|\|Q^+A_{22}\boldsymbol{\sigma}\|_\infty \\
&= \|\boldsymbol{\zeta}(0)\|e^{-\gamma t} + \frac{1-e^{-\gamma t}}{\gamma}\|Q^+A_{22}\boldsymbol{\sigma}\|_\infty
\end{aligned}$$

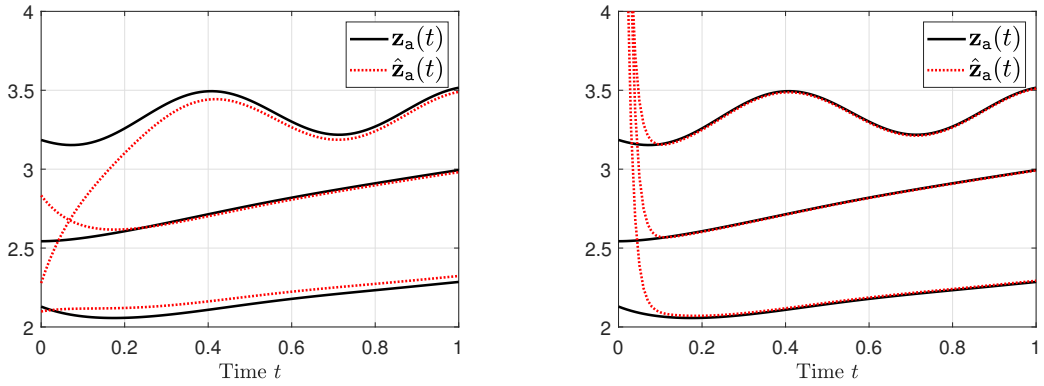
where the first step is due to the triangular inequality, the second step is due to the mean value theorem [Bartle1964], and the third step is obtained by integrating $e^{-\gamma\tau}$ for the given limits. Since $\|\boldsymbol{\sigma}\|_\infty \le \bar{\sigma} < \infty$, we consider $a = \|\boldsymbol{\zeta}_0\|$ and $b = \|Q^+A_{22}\boldsymbol{\sigma}\|_\infty$ to prove (3.20). □

The consequence of (3.20) is that we can consider a high-gain-type design of the average observer $\mathbf{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$ in order to ensure $\lim_{t\to\infty}\|\boldsymbol{\zeta}(t)\| = 0$ with $\gamma \to \infty$. That is, the average estimation error can be made arbitrarily small by choosing γ to be arbitrarily large. The design of $\mathbf{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$, therefore, is given by (3.6) with L_γ given in (3.21).

Example 3.3. Consider a consensus seeking multi-agent system with external inputs

$$\dot{x}_i(t) = \sum_{j\in\mathcal{N}_{i\leftarrow\mathcal{V}}} a_{ij}[x_j(t) - x_i(t)] + \sum_{l=1}^{3} b_{il}u_l(t)$$

where $a_{ij} = 1$ for all $i \in \mathcal{V}$ and $j \in \mathcal{N}_{i\leftarrow\mathcal{V}}$. The scalars $b_{il} = 1$ if $u_l(t)$ is applied on i, and $b_{il} = 0$ otherwise, for $l = 1, 2, 3$. Let the system be defined on the graph of Figure 3.1(b), where the measured nodes are shown as black and the three clusters of unmeasured nodes



(a) $\gamma = 10$ *(b)* $\gamma = 50$

Figure 3.3: Average estimation of three clusters of unmeasured nodes of the graph of Figure 3.1(b).

are shown as blue, green, and orange, respectively. In this example, the state matrix $A = -\mathcal{L}(\mathcal{G})$, where $\mathcal{L}(\mathcal{G})$ is the Laplacian matrix of the graph $\mathcal{G}$ of Figure 3.1(b). The output matrix $C = [\ I_3 \quad 0\ ]$. Let the input matrix $B = [\ I_3 \quad 0\ ]^\intercal$. The input vector is defined as

$$\mathbf{u}(t) = \begin{bmatrix} u_1(t) \\ u_2(t) \\ u_3(t) \end{bmatrix} = \begin{bmatrix} 20\sin(t) + 1 \\ \sin(2t) + 1 \\ \sin(5t) + 1 \end{bmatrix}$$

The characteristic matrix $Q = \mathtt{diag}(\mathbf{1}_8, \mathbf{1}_7, \mathbf{1}_8)$ and A_{12} is given in (3.13). Notice that (3.17) is satisfied and the system is average observable. The design of average observer obtained from (3.21) and (3.6) is given as follows

$$L = \begin{bmatrix} \frac{\gamma}{8} - \frac{11}{64} & \frac{1}{56} & \frac{1}{32} \\ \frac{1}{56} & \frac{\gamma}{7} - \frac{11}{49} & \frac{3}{56} \\ \frac{1}{32} & \frac{3}{56} & \frac{\gamma}{8} - \frac{13}{64} \end{bmatrix}, \quad M = \begin{bmatrix} -\gamma & 0 & 0 \\ 0 & -\gamma & 0 \\ 0 & 0 & -\gamma \end{bmatrix}, \quad N = -L$$

$$K = \begin{bmatrix} \gamma - \gamma\left(\frac{\gamma}{8} - \frac{11}{64}\right) - \frac{3}{8} & \frac{1}{8} - \frac{\gamma}{56} & \frac{1}{4} - \frac{\gamma}{32} \\ \frac{1}{7} - \frac{\gamma}{56} & \gamma - \gamma\left(\frac{\gamma}{7} - \frac{11}{49}\right) - \frac{4}{7} & \frac{3}{7} - \frac{3\gamma}{56} \\ \frac{1}{4} - \frac{\gamma}{32} & \frac{3}{8} - \frac{3\gamma}{56} & \gamma - \gamma\left(\frac{\gamma}{8} - \frac{13}{64}\right) - \frac{5}{8} \end{bmatrix}.$$

Since average observability allows for a high-gain type of average observer, we consider two values of the gain $\gamma = 10$ and $\gamma = 50$ to illustrate (3.20). As shown in Figure 3.3, the larger gain $\gamma = 50$ gives smaller average estimation error asymptotically.

3.4 Average Detectability

The notion of average detectability is defined through the stability of average states of clusters whose dynamics are described by the projected network system. After providing

a necessary and sufficient condition of average detectability, we provide its graph-theoretic interpretation showing that it demands a certain regularity and symmetry in the intra-cluster and inter-cluster graph topologies. Finally, we show that average detectability allows for an open-loop average observer.

3.4.1 Necessary and sufficient condition

In this section, we consider the notion of average detectability, which relates to the exponential stability of the average states of the clusters under the absence of output $\mathbf{y}(t)$ and the input $\mathbf{u}(t)$. Basically, as it will be explained in more detail, we say that a clustered network system $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average detectable if the average state vector $\mathbf{z}_\mathtt{a}(t)$ converges to a zero vector when $\mathbf{y}(t) = 0$ and $\mathbf{u}(t) = 0$ for every $t \in \mathbb{R}_{\geq 0}$ irrespective of the average deviation vector $\boldsymbol{\sigma}(t) \in \mathbb{R}^n$.

Definition 3.4. The clustered network system $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is said to be *average detectable* if—for every $t \in \mathbb{R}_{\geq 0}$, $\mathbf{z}_\mathtt{a}(0) \in \mathbb{R}^k$, and $\boldsymbol{\sigma}(t) \in \mathbb{R}^n$ such that $Q^+\boldsymbol{\sigma} \equiv \mathbf{0}_k$—the zero output $\mathbf{y}(t) = \mathbf{0}_m$ and the zero input $\mathbf{u}(t) = \mathbf{0}_p$ implies that the average state vector $\mathbf{z}_\mathtt{a}(t) \in \mathbb{R}^k$ converges to $\mathbf{0}_k$ asymptotically as time $t \to \infty$. In particular,

$$\mathbf{y} \equiv \mathbf{0}_m \text{ and } \mathbf{u} \equiv \mathbf{0}_p \quad \Rightarrow \quad \lim_{t\to\infty} \|\mathbf{z}_\mathtt{a}(t)\| = 0, \ \forall \boldsymbol{\sigma}(t) \in \mathbb{R}^n \text{ such that } Q^+\boldsymbol{\sigma} \equiv \mathbf{0}_k.$$

Traditionally, for defining the notions of detectability for linear systems, it is assumed that the input $\mathbf{u} \equiv 0$ because it is known and can be subtracted from the solution of $\mathbf{z}_\mathtt{a}(t)$. Then, $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ with a given set of measured nodes $\mathcal{V}_1$ and a clustering $\mathcal{Q} = \{\mathcal{C}_1, \cdots, \mathcal{C}_k\}$ of unmeasured nodes is said to be average detectable if the zero output $\mathbf{y} \equiv 0$ implies that the average state vector $\mathbf{z}_\mathtt{a}(t)$ converges to zero asymptotically as $t \to \infty$.

Another equivalent way of defining average detectability is as follows. From the dynamics of the projected network system $\mathring{\mathbf{\Sigma}}_{\mathcal{V}_1,\mathcal{Q}}$ in equation (2.9), let us write the dynamics of the average state vector

$$\dot{\mathbf{z}}_\mathtt{a}(t) = Q^+A_{22}Q\mathbf{z}_\mathtt{a}(t) + Q^+A_{22}\boldsymbol{\sigma}(t) + Q^+A_{21}\mathbf{y}(t) + Q^+B_2\mathbf{u}(t). \tag{3.22}$$

Here, the average deviation vector $\boldsymbol{\sigma}(t)$ acts as a structured unknown input. By ignoring $\boldsymbol{\sigma}(t)$, we obtain an 'approximated' average state vector $\hat{\mathbf{z}}_\mathtt{a}(t)$ that satisfies

$$\dot{\hat{\mathbf{z}}}_\mathtt{a}(t) = Q^+A_{22}Q\hat{\mathbf{z}}_\mathtt{a}(t) + Q^+A_{21}\mathbf{y}(t) + Q^+B_2\mathbf{u}(t). \tag{3.23}$$

Thus, if we define the average approximation error

$$\tilde{\mathbf{z}}_\mathtt{a}(t) = \mathbf{z}_\mathtt{a}(t) - \hat{\mathbf{z}}_\mathtt{a}(t) \tag{3.24}$$

we obtain

$$\dot{\tilde{\mathbf{z}}}_\mathtt{a}(t) = Q^+A_{22}Q\tilde{\mathbf{z}}_\mathtt{a}(t) + Q^+A_{22}\boldsymbol{\sigma}(t). \tag{3.25}$$

Proposition 3.5. The clustered network system $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average detectable *if and only if*, for every $\tilde{\mathbf{z}}_\mathtt{a}(0) \in \mathbb{R}^k$, the approximation error $\tilde{\mathbf{z}}_\mathtt{a}(t)$ in (3.24) converges to zero $\mathbf{0}_k$ asymptotically as time $t \to \infty$ for all $\boldsymbol{\sigma}(t) \in \mathbb{R}^n$ such that $Q^+\boldsymbol{\sigma} \equiv \mathbf{0}_k$.

Proof of sufficiency. Assume, for every $\tilde{\mathbf{z}}_\mathtt{a}(0) \in \mathbb{R}^k$, $\tilde{\mathbf{z}}_\mathtt{a}(t) \to \mathbf{0}_k$ as $t \to \infty$. If we choose $\tilde{\mathbf{z}}_\mathtt{a}(0) = \mathbf{z}_\mathtt{a}(0)$, then, from (3.22), we have $\mathbf{z}_\mathtt{a}(t) \to \mathbf{0}_k$ as $t \to \infty$ when $\mathbf{y} \equiv \mathbf{0}_m$ and $\mathbf{u} \equiv \mathbf{0}_p$, which proves that $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average detectable.

Proof of necessity. Assume $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average detectable. Then, for every $\mathbf{z}_\mathtt{a}(0) \in \mathbb{R}^k$, $\mathbf{z}_\mathtt{a}(t) \to \mathbf{0}_k$ as $t \to \infty$ when $\mathbf{y} \equiv \mathbf{0}_m$ and $\mathbf{u} \equiv \mathbf{0}_p$. If we choose $\mathbf{z}_\mathtt{a}(0) = \tilde{\mathbf{z}}_\mathtt{a}(0)$, equation (3.22) with "$\mathbf{y} \equiv \mathbf{0}_m$ and $\mathbf{u} \equiv \mathbf{0}_p$" and equation (3.23) are equivalent. This implies that $\tilde{\mathbf{z}}_\mathtt{a}(t) \to \mathbf{0}_k$ as $t \to \infty$.

□

Now that we have defined average detectability from the stability of (3.23), we present the following necessary and sufficient condition that corresponds to the structure of the projected network system $\mathring{\boldsymbol{\Sigma}}_{\mathcal{V}_1,\mathcal{Q}}$.

Theorem 3.6. The clustered network system $\boldsymbol{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average detectable *if and only if* both the following conditions hold:

(i) $\ker(Q^+A_{22}) \supseteq \ker(Q^+)$

(ii) $Q^+A_{22}Q$ is Hurwitz.

Proof of sufficiency. Note that (i) is equivalent to say that, for every vector $\mathbf{v} \in \ker(Q^+)$, it holds $\mathbf{v} \in \ker(Q^+A_{22})$. Then, recall that the columns of $I_n - QQ^+$ form a complete basis of $\ker(Q^+)$ because $Q^+(I_n - QQ^+) = \mathbf{0}_k$ and $\mathtt{nullity}(Q^+) = \mathtt{rank}(I_n - QQ^+) = n - k$. Thus, (i) implies $Q^+A_{22}(I_n - QQ^+) = \mathbf{0}_k$, which gives $Q^+A_{22} = Q^+A_{22}QQ^+$. From (3.23), we obtain the solution trajectory

$$\tilde{\mathbf{z}}_{\mathtt{a}}(t) = \exp(Q^+A_{22}Qt)\tilde{\mathbf{z}}_{\mathtt{a}}(0) + \int_0^t \exp(Q^+A_{22}Q\tau)Q^+A_{22}\boldsymbol{\sigma}(t-\tau)d\tau. \tag{3.26}$$

If (i) holds, then $Q^+A_{22}\boldsymbol{\sigma} = Q^+A_{22}QQ^+\boldsymbol{\sigma} \equiv \mathbf{0}_k$. If (ii) holds, then $\mathtt{eig}(Q^+A_{22}Q) \subset \mathbb{C}_{<0}$ and $\exp(Q^+A_{22}Qt) \to 0_{k\times k}$ as $t \to \infty$. Hence, $\tilde{\mathbf{z}}_{\mathtt{a}}(t) \to 0$ as $t \to \infty$ for all $\tilde{\mathbf{z}}_{\mathtt{a}}(0) \in \mathbb{R}^k$.

Proof of necessity. Assume $\boldsymbol{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average detectable. However, only (ii) holds and (i) does not hold. Then, we have

$$\lim_{t\to\infty} \|\tilde{\mathbf{z}}_{\mathtt{a}}(t)\| = \lim_{t\to\infty} \| \int_0^t \exp(Q^+A_{22}Q\eta)Q^+A_{22}\boldsymbol{\sigma}(t-\eta)d\eta\|$$

which must be 0 for average detectability. Since $\lim_{t\to\infty}\|\boldsymbol{\sigma}(t)\|$ is not necessarily zero and (i) does not hold, therefore we have $\lim_{t\to\infty}\|\tilde{\mathbf{z}}_{\mathtt{a}}(t)\| = 0$ only if $\exp(Q^+A_{22}Qt)Q^+A_{22} = 0_{k\times n}$ for all $t \in \mathbb{R}_{\geq 0}$. This is not possible because matrix exponential $\exp(Q^+A_{22}Qt)$ is always nonsingular for every $t \in \mathbb{R}_{\geq 0}$ (see [Horn2013, Ch. 5, Sec. 5.6, Prob. 43]) and, due to (ii), $Q^+A_{22} \neq 0_{k\times n}$. Second, assume (i) holds but (ii) doesn't hold, then $\exp(Q^+A_{22}Qt) \to \infty$ as $t \to \infty$ and, therefore, $\tilde{\mathbf{z}}_{\mathtt{a}}(t) \to \infty$. This completes the proof. □

In the following corollary, we relate all the three notions of average reconstructability, average observability, and average detectability. Notice that Theorem 3.6(i) is a necessary condition of average detectability.

Corollary 3.6.1. If the clustered network system $\boldsymbol{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average observable and satisfies a necessary condition Theorem 3.6(i) of average detectability, then $\boldsymbol{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average reconstructable.

Proof. If $\boldsymbol{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average observable, then, by Theorem 3.3,

$$\mathtt{rank}\left(\begin{bmatrix} A_{12} \\ Q^+ \end{bmatrix}\right) = \mathtt{rank}(A_{12}).$$

If a necessary condition of average detectability, Theorem 3.6(i), is satisfied, then

$$\mathtt{rank}\left(\begin{bmatrix} Q^+A_{22} \\ Q^+ \end{bmatrix}\right) = \mathtt{rank}(Q^+).$$

The above equalities imply that

$$\mathtt{rank}\left(\begin{bmatrix} A_{12} \\ Q^+A_{22} \\ Q^+ \end{bmatrix}\right) = \mathtt{rank}\left(\begin{bmatrix} A_{12} \\ Q^+ \end{bmatrix}\right) = \mathtt{rank}(A_{12})$$

which concludes the proof by (3.2).

□

In Theorem 3.6, we stated a necessary and sufficient condition of average detectability, which implies that the average approximation error $\tilde{\mathbf{z}}_{\mathtt{a}}(t)$ converges to zero asymptotically as $t \to \infty$. However, irrespective of average detectability, we have $\tilde{\mathbf{z}}_{\mathtt{a}}(t) \to \mathbf{0}_k$ asymptotically as $t \to \infty$ under the following condition.

Theorem 3.7. Assume that $Q^+A_{22}Q$ is Hurwitz. Then, the average approximation error $\tilde{\mathbf{z}}_{\mathtt{a}}(t) \to \mathbf{0}_k$ asymptotically as $t \to \infty$ if $\lim_{t\to\infty} \|\boldsymbol{\sigma}(t)\| = 0$.

Proof. Since $Q^+A_{22}Q$ is assumed to be Hurwitz, therefore, in (3.26), we have

$$\lim_{t\to\infty} \| \exp(Q^+A_{22}Qt)\tilde{\mathbf{z}}_{\mathtt{a}}(0)\| = 0$$

for all $\tilde{\mathbf{z}}_{\mathtt{a}}(0) \in \mathbb{R}^k$. Thus, we consider only the second term on the right hand side of (3.26) denoted as

$$\mathbf{v}(t) = \int_0^t \exp(Q^+A_{22}Q\tau)Q^+A_{22}\boldsymbol{\sigma}(t-\tau)d\tau.$$

By splitting the integral at $t/2$ and then changing the variable τ to η, we obtain

$$\begin{aligned}\mathbf{v}(t) &= \int_0^{t/2} \exp(Q^+A_{22}Q\tau)Q^+A_{22}\boldsymbol{\sigma}(t-\tau)d\tau + \int_{t/2}^t \exp(Q^+A_{22}Q\tau)Q^+A_{22}\boldsymbol{\sigma}(t-\tau)d\tau \\ &= \int_{t/2}^t \exp[Q^+A_{22}Q(t-\eta)]Q^+A_{22}\boldsymbol{\sigma}(\eta)d\eta + \int_{t/2}^t \exp(Q^+A_{22}Q\tau)Q^+A_{22}\boldsymbol{\sigma}(t-\tau)d\tau.\end{aligned}$$

Let $|||\cdot|||$ denote the matrix norm induced by $\|\cdot\|$, then

$$\begin{aligned}\|\mathbf{v}(t)\| &= \| \int_{t/2}^t \exp[Q^+A_{22}Q(t-\eta)]Q^+A_{22}\boldsymbol{\sigma}(\eta)d\eta \\ &\qquad + \int_{t/2}^t \exp(Q^+A_{22}Q\tau)Q^+A_{22}\boldsymbol{\sigma}(t-\tau)d\tau\| \\ &\le \| \int_{t/2}^t \exp[Q^+A_{22}Q(t-\eta)]Q^+A_{22}\boldsymbol{\sigma}(\eta)d\eta\| \\ &\qquad + \| \int_{t/2}^t \exp(Q^+A_{22}Q\tau)Q^+A_{22}\boldsymbol{\sigma}(t-\tau)d\tau\|.\end{aligned}$$

By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\|\mathbf{v}(t)\| \le &\left[\int_{t/2}^t |||\exp[Q^+A_{22}Q(t-\eta)]|||^2 d\eta\right]^{\frac{1}{2}} \left[\int_{t/2}^t \|Q^+A_{22}\boldsymbol{\sigma}(\eta)\|^2 d\eta\right]^{\frac{1}{2}} \\ &+ \left[\int_{t/2}^t |||\exp(Q^+A_{22}Q\tau)|||^2 d\tau\right]^{\frac{1}{2}} \left[\int_{t/2}^t \|Q^+A_{22}\boldsymbol{\sigma}(t-\tau)\|^2 d\tau\right]^{\frac{1}{2}}.\end{aligned}$$

First, notice that

$$\lim_{t\to\infty}\int_{t/2}^{t}\left|\left|\left|\exp(Q^{+}A_{22}Q\tau)\right|\right|\right|^{2}d\tau=0$$

since $Q^{+}A_{22}Q$ is Hurwitz. Second, we have

$$\lim_{t\to\infty}\int_{t/2}^{t}\|Q^{+}A_{22}\boldsymbol{\sigma}(\eta)\|^{2}d\eta=0$$

since $\lim_{t\to\infty}\|\boldsymbol{\sigma}(t)\|=0$. Thus, $\lim_{t\to\infty}\|\mathbf{v}(t)\|=0$, which implies $\lim_{t\to\infty}\|\tilde{\mathbf{z}}_{\mathtt{a}}(t)\|=0$. □

Notice that $\lim_{t\to\infty}\|\boldsymbol{\sigma}(t)\|=0$ means that the states of nodes in each cluster either reach consensus or synchronize. Precisely, we have $\lim_{t\to\infty}\|\boldsymbol{\sigma}(t)\|=0$ if and only if, for every $i,j\in\mathcal{C}_\alpha$ and $\alpha\in\{1,\dots,k\}$, it holds that

$$\lim_{t\to\infty}x_i(t)-x_j(t)=0.$$

Therefore, multi-agent systems that seek consensus [Olfati-Saber2007] or synchronization [Scardovi2009] allow an open-loop average observer (3.23) that converges to the actual average state of clusters.

3.4.2 Graph-theoretic interpretation of average detectability

The necessary and sufficient condition of average detectability provided in Theorem 3.6 depends on the intra-cluster and inter-cluster graph topologies of $\mathcal{Q}=\{\mathcal{C}_1,\dots,\mathcal{C}_k\}$, which are captured by $\mathcal{G}_{\alpha\alpha}=(\mathcal{C}_\alpha,\mathcal{E}_{\alpha\alpha})$ and $\mathcal{G}_{\alpha\beta}=(\mathcal{C}_\alpha,\mathcal{C}_\beta,\mathcal{E}_{\alpha,\beta})$, respectively, for $\alpha,\beta=1,\dots,k$ and $\alpha\neq\beta$, where $\mathcal{E}_{\alpha\alpha}=\mathcal{E}\cap(\mathcal{C}_\alpha\times\mathcal{C}_\alpha)$ and $\mathcal{E}_{\alpha\beta}=\mathcal{E}\cap(\mathcal{C}_\alpha\times\mathcal{C}_\beta)$. Recall the relative outflow centrality of $\nu_i\in\mathcal{V}_2$ with respect to cluster $\mathcal{C}_\alpha$ in (2.7)

$$c_{i\to\mathcal{C}_\alpha}=\begin{cases}a_{ii}+\displaystyle\sum_{j\in\mathcal{C}_\alpha\cap\mathcal{N}_{i\to\mathcal{V}}}a_{ji} & \text{if } i\in\mathcal{C}_\alpha\\[2ex] \displaystyle\sum_{j\in\mathcal{C}_\alpha\cap\mathcal{N}_{i\to\mathcal{V}}}a_{ji} & \text{if } i\notin\mathcal{C}_\alpha.\end{cases}$$

Definition 3.5. The clustering $\mathcal{Q}=\{\mathcal{C}_1,\dots,\mathcal{C}_k\}$ is said to be *equitable* if, for every $\alpha,\beta\in\{1,\dots,k\}$ and $\nu_i\in\mathcal{C}_\alpha$, the relative outflow centrality $c_{i\to\mathcal{C}_\beta}=d_{\alpha\beta}$, where $d_{\alpha\beta}\in\mathbb{R}$.

Theorem 3.8. Assume $Q^{+}A_{22}Q$ is Hurwitz. Then, the clustered network system $\boldsymbol{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average detectable if and only if the clustering $\mathcal{Q}$ is equitable.

Proof. Since $Q^{+}A_{22}Q$ is Hurwitz, by Theorem 3.6, $\mathtt{ker}(Q^{+}A_{22})\supseteq\mathtt{ker}(Q^{+})$ is equivalent to average detectability of $\boldsymbol{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$. Therefore, in this proof, we show that $\mathtt{ker}(Q^{+}A_{22})\supseteq\mathtt{ker}(Q^{+})$ is equivalent to $\mathcal{Q}$ being equitable. Note that $\mathtt{ker}(Q^{+}A_{22})\supseteq\mathtt{ker}(Q^{+})$ is equivalent to $\mathtt{im}((Q^{+}A_{22})^{\intercal})\subseteq\mathtt{im}(Q^{+\intercal})$ by [Campbell2009, Proposition 0.2.1], i.e.,

$$\mathtt{rank}\left(\begin{bmatrix}Q^{+}\\ Q^{+}A_{22}\end{bmatrix}\right)=\mathtt{rank}(Q^{+})=k$$

which, in turn, is equivalent to the existence of a matrix $D\in\mathbb{R}^{k\times k}$ such that $DQ^{+}=Q^{+}A_{22}$. We have

$$Q^{+}A_{22}=\begin{bmatrix}\frac{1}{n_1}c_{1\to\mathcal{C}_1} & \cdots & \frac{1}{n_1}c_{n\to\mathcal{C}_1}\\ \vdots & \ddots & \vdots\\ \frac{1}{n_k}c_{1\to\mathcal{C}_k} & \cdots & \frac{1}{n_k}c_{n\to\mathcal{C}_k}\end{bmatrix}.$$

Without loss of generality, let $\mathcal{Q}$ be such that $\mathcal{C}_1 = \{\nu_1, \dots, \nu_{n_1}\}$, $\mathcal{C}_2 = \{\nu_{n_1+1}, \dots, \nu_{n_2+1}\}$, $\dots$, $\mathcal{C}_k = \{\nu_{n_{k-1}+1}, \dots, \nu_{n_{k-1}+n_k}\}$. Then, the characteristic matrix

$$Q = \mathtt{diag}(\mathbf{1}_{n_1}, \mathbf{1}_{n_2}, \dots, \mathbf{1}_{n_k})$$

and its left-pseudo inverse

$$Q^+ = \begin{bmatrix} \frac{1}{n_1}\mathbf{1}_{n_1}^\intercal & & & \\ & \frac{1}{n_2}\mathbf{1}_{n_2}^\intercal & & \\ & & \ddots & \\ & & & \frac{1}{n_k}\mathbf{1}_{n_k}^\intercal \end{bmatrix}.$$

If the clustering $\mathcal{Q}$ is equitable, then

$$Q^+ A_{22} = \begin{bmatrix} \frac{d_{11}}{n_1}\mathbf{1}_{n_1}^\intercal & \cdots & \frac{d_{1k}}{n_1}\mathbf{1}_{n_k}^\intercal \\ \vdots & \ddots & \vdots \\ \frac{d_{1k}}{n_k}\mathbf{1}_{n_1}^\intercal & \cdots & \frac{d_{kk}}{n_k}\mathbf{1}_{n_k}^\intercal \end{bmatrix}$$

and

$$D = \begin{bmatrix} d_{11} & \cdots & d_{1k} \\ \vdots & \ddots & \vdots \\ d_{k1} & \cdots & d_{kk} \end{bmatrix}$$

implies $DQ^+ = Q^+ A_{22}$, where $d_{\alpha\beta} \in \mathbb{R}$ given in Definition 3.5. In the other direction, let $D \in \mathbb{R}^{k\times k}$ with $[D]_{\alpha\beta} = d_{\alpha\beta}$ be such that $Q^+ A_{22} = DQ^+$, then

$$\begin{bmatrix} \frac{1}{n_1}c_{1\to\mathcal{C}_1} & \cdots & \frac{1}{n_1}c_{n\to\mathcal{C}_1} \\ \vdots & \ddots & \vdots \\ \frac{1}{n_k}c_{1\to\mathcal{C}_k} & \cdots & \frac{1}{n_k}c_{n\to\mathcal{C}_k} \end{bmatrix} = \begin{bmatrix} d_{11} & \cdots & d_{1k} \\ \vdots & \ddots & \vdots \\ d_{k1} & \cdots & d_{kk} \end{bmatrix} \begin{bmatrix} \frac{1}{n_1}\mathbf{1}_{n_1}^\intercal & & & \\ & \frac{1}{n_2}\mathbf{1}_{n_2}^\intercal & & \\ & & \ddots & \\ & & & \frac{1}{n_k}\mathbf{1}_{n_k}^\intercal \end{bmatrix}$$

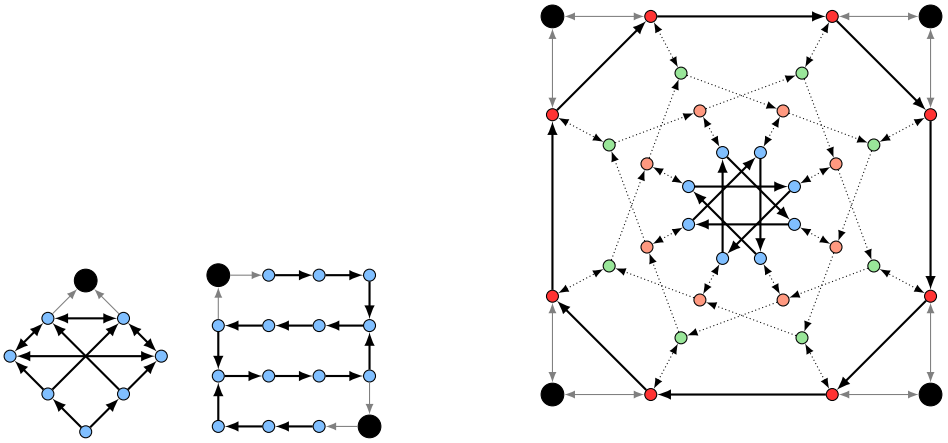
$$= \begin{bmatrix} \frac{d_{11}}{n_1}\mathbf{1}_{n_1}^\intercal & \cdots & \frac{d_{1k}}{n_1}\mathbf{1}_{n_k}^\intercal \\ \vdots & \ddots & \vdots \\ \frac{d_{1k}}{n_k}\mathbf{1}_{n_1}^\intercal & \cdots & \frac{d_{kk}}{n_k}\mathbf{1}_{n_k}^\intercal \end{bmatrix}.$$

Therefore, for every $\alpha \in \{1, \dots, k\}$,

$$\begin{aligned} c_{1\to\mathcal{C}_\alpha} &= \cdots = c_{n_1\to\mathcal{C}_\alpha} &&= d_{1\alpha} \\ \vdots & & \vdots & \vdots \\ c_{(n_{k-1}+1)\to\mathcal{C}_\alpha} &= \cdots = c_{n_k\to\mathcal{C}_\alpha} &&= d_{k\alpha} \end{aligned}$$

which implies that $\mathcal{Q}$ is equitable.

□



(a) Examples with one cluster of unmeasured nodes (shown as blue)

(b) Example with four clusters of unmeasured nodes (shown as red, green, orange, and blue)

Figure 3.4: Examples of clustered network systems with an equitable clustering of unmeasured nodes.

For a clustered network system with a single cluster of unmeasured nodes, we have $\mathcal{Q} = \mathcal{C}_1 = \mathcal{V}_2$. In this case, we only have intra-cluster topology captured by the induced subgraph $\mathcal{G}_{11} = (\mathcal{C}_1, \mathcal{E}_{11})$. Average detectability requires certain regularity of $\mathcal{G}_{11}$. To elaborate a single cluster case, notice that the conditions in Theorem 3.6 boil down to $\mathbf{1}_n^\intercal A_{22} = -\gamma \mathbf{1}_n^\intercal$, where $\mathbf{1}_n = Q$ and $\gamma > 0$. That is, for average detectability of a network system with a single cluster of unmeasured nodes, the induced subgraph $\mathcal{G}_{11} = (\mathcal{C}_1, \mathcal{E}_{11})$ must be regular in a way that the relative outflow centrality $c_{i \to \mathcal{C}_1}$ of every node $i \in \mathcal{C}_1$ must be equal and negative. Such regularity is illustrated by the graphs shown in Figure 3.4(a), where the measured nodes are shown as black and the unmeasured nodes are shown as blue. In this figure, the relative outflow centrality of every unmeasured node with respect to $\mathcal{C}_1$ in the left and right graph is -1 and -2, respectively.

On the other hand, for a clustered network system with multiple clusters of unmeasured nodes, average detectability requires that the relative outflow centrality of any pair of nodes $i, j \in \mathcal{C}_\alpha$ with respect to cluster $\mathcal{C}_\beta$, for every $\alpha, \beta \in \{1, \dots, k\}$, be equal. This is illustrated by an unweighted graph shown in Figure 3.4(b), where the clusters are highlighted with red, green, brown, and blue nodes, respectively, and the measured nodes are shown as black. Consider intra-cluster topology, notice that the clustering is equitable because for all nodes in red, green, brown and blue clusters, we have relative outflow centrality with respect to their own clusters is equal to -2, -1, -2 and -1, respectively. Moreover, considering the inter-cluster topology through the induced bipartite subgraphs $\mathcal{G}_{\alpha\beta}$, we see that the relative outflow centralities of all the nodes in a certain cluster are also equal. For instance, consider the induced bipartite subgraph $\mathcal{G}_{45}$ with directed edges from blue to brown nodes. Each blue node has relative outflow centrality with respect to brown nodes equal to 1. Similarly, in both $\mathcal{G}_{35}$ and $\mathcal{G}_{25}$, the relative outflow centrality of blue nodes with respect to $\mathcal{C}_3$ and $\mathcal{C}_2$ is 0. Therefore, the clustering of unmeasured nodes $\mathcal{Q}$ is equitable.

3.4.3 Design of average observer under average detectability

Under average detectability, we have an average observer $\mathbf{\Omega}_{\mathcal{V}_1, \mathcal{Q}}$ with open-loop design, which is equivalent to (3.23).

Lemma 3.9. The clustered network system $\mathbf{\Sigma}_{\mathcal{V}_1, \mathcal{Q}}$ is average detectable if and only

if, for every $\boldsymbol{\zeta}_0 = \boldsymbol{\zeta}(0) \in \mathbb{R}^k$, there exist constants $a = a(\boldsymbol{\zeta}_0) > 0$ and $\gamma > 0$ such that

$$L = 0_{k\times m} \quad \Rightarrow \quad \|\boldsymbol{\zeta}(t)\| \leq ae^{-\gamma t}.$$

Proof of sufficiency. If, for some $a, \gamma > 0$, $\|\boldsymbol{\zeta}(t)\| \leq ae^{-\gamma t}$, then, by Proposition 3.5, $\boldsymbol{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average detectable.

Proof of necessity. If $\boldsymbol{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average detectable, then $Q^+A_{22}\boldsymbol{\sigma} \equiv \mathbf{0}_k$ and $Q^+A_{22}Q$ is Hurwitz by Theorem 3.6. Then, choosing $L = 0_{k\times m}$ implies that the average estimation error (3.7) is given as

$$\dot{\boldsymbol{\zeta}}(t) = Q^+A_{22}Q\boldsymbol{\zeta}(t) + Q^+A_{22}\boldsymbol{\sigma}(t) = Q^+A_{22}Q\boldsymbol{\zeta}(t)$$

whose solution is given by $\boldsymbol{\zeta}(t) = \exp(Q^+A_{22}Qt)\boldsymbol{\zeta}_0$. Since $Q^+A_{22}Q$ is Hurwitz, there exists a constant $\gamma > 0$ such that $\|\boldsymbol{\zeta}(t)\| \leq \|\boldsymbol{\zeta}_0\|e^{-\gamma t}$. Choosing $a = \|\boldsymbol{\zeta}_0\|$ concludes the proof. □

In light of the above lemma, the design of average observer under average detectability is given by

$$\begin{aligned} M &= Q^+A_{22}Q \\ N &= Q^+B_2 \\ K &= Q^+A_{21} \\ L &= 0_{k\times m}. \end{aligned} \tag{3.27}$$

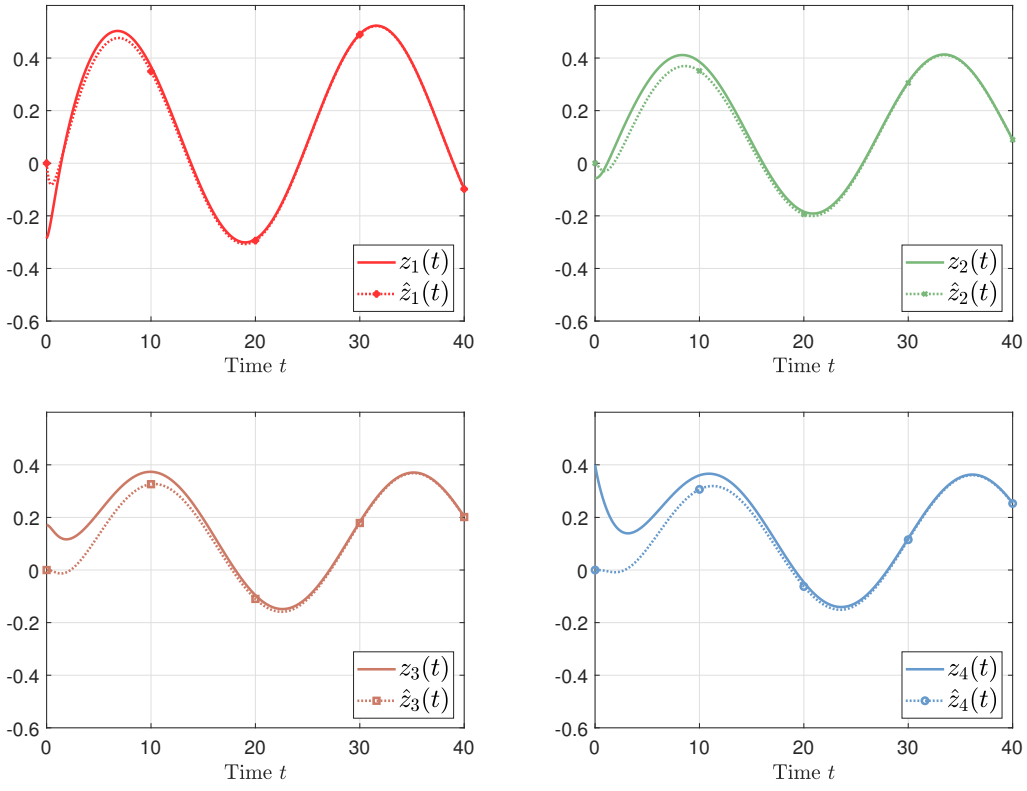


Figure 3.5: Average estimation of clusters of unmeasured nodes of the network shown in Figure 3.4.

Example 3.4. Consider an unweighted graph shown in Figure 3.4(b), where the measured nodes $\mathcal{V}_1 = \{1, 2, 3, 4\}$ are shown as black. The state of each node i evolves according to (3.15). The input is given by

$$u_l(t) = \begin{cases} \sin(0.5t + (l-1)\pi/4) & \text{if } l = 1, \ldots, 4 \\ 0 & \text{otherwise} \end{cases}$$

and the input matrix $B = [\ I_4 \quad 0\ ]^\intercal$. The design of average observer is obtained from (3.27), where $L = 0_{4\times4}$, $N = 0_{4\times4}$,

$$M = \begin{bmatrix} -2 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 1 & -1 \end{bmatrix}, \quad K \begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Note that Theorem 3.6(ii) is satisfied since $M = Q^+ A_{22} Q$ is Hurwitz with $\texttt{eig}(M) = \{-3.5321, -2.3473, -1.0000, -0.1206\}$. Also, Theorem 3.6(i) is satisfied since the clustering is equitable, i.e., $Q^+ A_{22} = Q^+ A_{22} Q Q^+$, where $Q = \texttt{diag}(\mathbf{1}_8, \mathbf{1}_8, \mathbf{1}_8, \mathbf{1}_8)$. Therefore, the average observer $\mathbf{\Omega}_{\mathcal{V}_1, \mathcal{Q}}$ with open-loop design (3.27) converges to the average state of unmeasured clusters as shown in Figure 3.5, where the initial states $\mathbf{x}(0)$ are chosen uniformly in the interval $(-0.5, 0.5)$.

3.5 Remarks on Scale-free Networks and a Scaling Property

3.5.1 Scale-free networks vs. average reconstructability and average observability

Scale-free network structure emerges ubiquitously in real-world large-scale systems such as world wide web [Albert1999], metabolic networks [Jeong2000], epidemic spread [Pastor-Satorras2001b], urban transit system [Wu2004], and many more (see [Barabási2003] and [Barabási2009]). A scale-free network has a property that its degree distribution follows a power law

$$P(d) \sim d^{-\gamma}$$

which states that the fraction of nodes $P(d)$ with degree d in the network decays with d and is proportional to $d^{-\gamma}$, where $\gamma > 0$ is the exponent of decay. Such property implies that scale-free networks have few hub nodes with very large degrees and many exterior nodes with very small degrees. The hubs lie in the tail of power law distribution. For example, the network of Figure 3.1(b) is a scale-free network whose degree distribution is plotted in Figure 3.6 with a power law fitting $P(d) = 0.15 d^{-0.95}$.

Liu et al. [Liu2011] showed that the number of sensors required to render a scale-free network controllable/observable is typically much larger than the requirement for an Erdős-Réyi network. Moreover, considering the hubs of a scale-free network as measured nodes does not usually make the network observable because of dilation [Lin1974, Liu2011] that results in the non-idenfiability of exterior nodes. However, for average observability and average reconstructability, scale-free networks are well-suited and can be made average

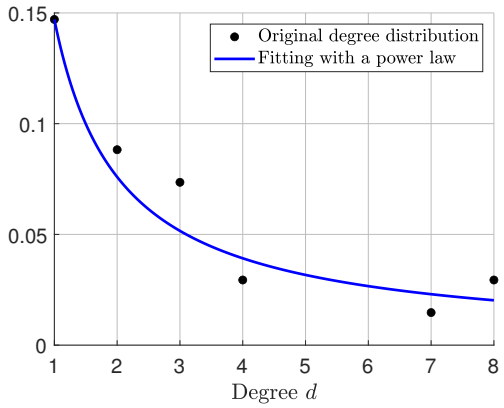


Figure 3.6: Degree distribution of a scale-free network of Figure 3.1(b).

observable and/or average reconstructable by choosing the hubs as measured nodes. This is because the hubs have large degrees and span most of the exterior nodes. Therefore, as depicted in the example of Figure 3.1(b), considering the hubs as measured nodes and the exterior nodes around each hub as a cluster of unmeasured nodes may suffice in scale-free networks to satisfy the condition of average reconstructability and reconstructability in Corollary 3.2.1 and 3.3.3. In general, of course, one may need to also include some exterior nodes as measured nodes in addition to the hubs, nonetheless, the number of measured nodes required for average reconstructability and average observability in scale-free networks is much less as compared to the size of network.

3.5.2 Scaling property vs. average detectability

The necessary and sufficient condition of average detectability requires that the clustering of unmeasured nodes be equitable. However, in many applications, the clustering is a priori specified and fixed, where the task is to estimate and control some aggregated state profile of the clusters [Niazi2020b, Nikitin2021]. In such a case, the system may not be average detectable. In this subsection, we show that as the scale of the system increases, the open-loop estimation error approaches to zero asymptotically every if the system is not average detectable. For simplicity, we assume a single cluster case, where $Q = \mathbf{1}_n$, where the necessary and sufficient condition of average detectability in Theorem 3.6 reduces to

$$\mathbf{1}_n^\intercal A_{22} = -\gamma \mathbf{1}_n^\intercal, \quad \gamma > 0$$

which implies that the relative outflow centrality (2.7) of every node $i \in \mathcal{C}_1$ with respect to $\mathcal{C}_1$ is equal to $-\gamma$, where the minus sign is due to $Q^+ A_{22} Q$ being Hurwitz (Theorem 3.6(ii)). This means that the induced subgraph $\mathcal{G}_{11} = (\mathcal{C}_1, \mathcal{E}_{11})$ formed by cluster $\mathcal{C}_1$ needs to be regular with respect to the relative outflow centralities. If $\mathcal{G}_{11}$ is not completely regular and the average detectability condition is not completely satisfied, then we have

$$\mathbf{1}_n^\intercal A_{22} = -\gamma \mathbf{1}_n^\intercal + \mathbf{s}^\intercal \tag{3.28}$$

where $\mathbf{s} \in \mathbb{R}^n$ is a sparse vector that has non-zero values only at the place of nodes that have different relative outflow centralities.

We would like to study the effect of scale on the average estimation error $\boldsymbol{\zeta}(t)$ when the size of the network is very large, i.e., $n \to \infty$. From (3.27), we have the following design

of average observer $\mathbf{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$

$$\begin{aligned} M &= \tfrac{1}{n}\mathbf{1}_n^\intercal A_{22}\mathbf{1}_n \\ N &= \tfrac{1}{n}\mathbf{1}_n^\intercal B_2 \\ K &= \tfrac{1}{n}\mathbf{1}_n^\intercal A_{21} \\ L &= 0_{1\times 4} \end{aligned}$$

and the average estimation error satisfies

$$\dot{\boldsymbol{\zeta}}(t) = \frac{1}{n}\mathbf{1}_n^\intercal A_{22}\mathbf{1}_n\boldsymbol{\zeta}(t) + \frac{1}{n}\mathbf{1}_n^\intercal A_{22}\boldsymbol{\sigma}(t).$$

From (3.28), we have

$$\begin{aligned} \frac{1}{n}\mathbf{1}_n^\intercal A_{22}\mathbf{1}_n &= \frac{1}{n}(-\gamma\mathbf{1}_n^\intercal + \mathbf{s}^\intercal)\mathbf{1}_n &= -\gamma + \frac{\mathbf{s}^\intercal\mathbf{1}_n}{n} \\ \frac{1}{n}\mathbf{1}_n^\intercal A_{22}\boldsymbol{\sigma}(t) &= \frac{1}{n}(-\gamma\mathbf{1}_n^\intercal + \mathbf{s}^\intercal)\boldsymbol{\sigma}(t) &= \frac{1}{n}\mathbf{s}^\intercal\boldsymbol{\sigma}(t) \end{aligned}$$

which implies

$$\boldsymbol{\zeta}(t) = e^{(-\gamma+\frac{\mathbf{s}^\intercal\mathbf{1}_n}{n})t}\boldsymbol{\zeta}(t) + \frac{1}{n}\int_0^t e^{(-\gamma+\frac{\mathbf{s}^\intercal\mathbf{1}_n}{n})\tau}\mathbf{s}^\intercal\boldsymbol{\sigma}(t-\tau)d\tau$$

Assume n to be sufficiently large such that $\gamma n > \mathbf{s}^\intercal\mathbf{1}_n$. Let $\|\bar{\boldsymbol{\sigma}}\|_\infty \le \bar{\sigma} < \infty$ and $\zeta_\infty = \limsup_{t\to\infty}\|\boldsymbol{\zeta}(t)\|$, then

$$\begin{aligned} \zeta_\infty &\le \limsup_{t\to\infty}\|\frac{1}{n}\int_0^t e^{(-\gamma+\frac{\mathbf{s}^\intercal\mathbf{1}_n}{n})\tau}\mathbf{s}^\intercal\boldsymbol{\sigma}(t-\tau)d\tau\| \\ &\le \frac{1}{n}\mathbf{s}^\intercal\mathbf{1}_n\bar{\sigma}\int_0^\infty e^{(-\gamma+\frac{\mathbf{s}^\intercal\mathbf{1}_n}{n})\tau}d\tau \\ &= \frac{\bar{\sigma}\mathbf{s}^\intercal\mathbf{1}_n}{n\gamma - \mathbf{s}^\intercal\mathbf{1}_n}. \end{aligned}$$

If $\lim_{n\leftarrow\infty}\frac{n}{\mathbf{s}^\intercal\mathbf{1}_n} = \infty$, then we have

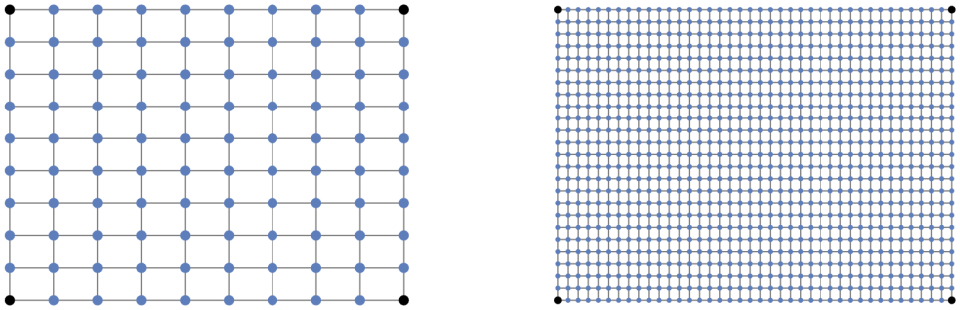
$$\lim_{n\to\infty}\zeta_\infty = 0$$

which means that the average estimation error converges to zero asymptotically as the scale of the system becomes arbitrarily large. We illustrate this by the following example of grid networks.

Example 3.5 (Grid networks). Grid network topology is found in several real-world applications such as urban traffic networks [Gartner2002] and agricultural monitoring [Goh2006]. It also emerges as a result of space-discretization of systems governed by partial differential equations such as asset pricing in finance [Bodeau2000], fluid dynamics [Bungartz2010], topological analysis of indoor spaces [Li2010], and temperature estimation of power modules [Sakhraoui2018]. The main property of a grid graph, also known as lattice, is that it forms a regular tiling, where most of the nodes have degree equal to four.

Consider a spatially-discrete reaction diffusion system [Ishizaki2014] over a grid network with single cluster $\mathcal{Q} = \mathcal{C}_1 = \mathcal{V}_2$ of unmeasured nodes, where the local damping at each node i is given by

$$a_{ii} = -r_i - \sum_{j\neq i} a_{ij}$$



(a) Grid network of size 10×10 *(b) Grid network of size* 25×40

Figure 3.7: Examples of grid networks with four measured nodes (black) at the corners and one cluster of unmeasured nodes (blue).

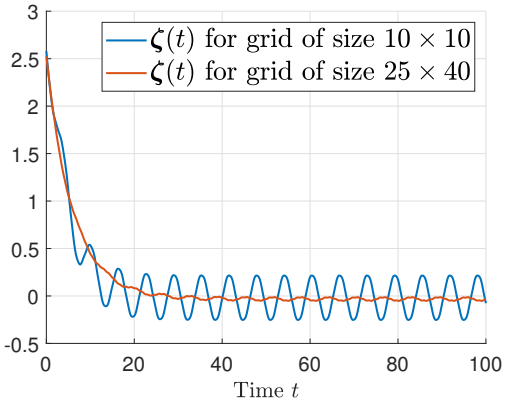


Figure 3.8: Illustration of scaling property for the grid networks of Figure 3.7.

with r_i the reaction rate at i. Suppose the four corner nodes of the grid are measured and the rest of the nodes form a single cluster $\mathcal{C}_1$ of unmeasured nodes as shown in Figure 3.7. For simplicty, assume the grid network is unweighted, i.e., $a_{ij} = 1$, and $r_i = r$ for every $i \in \mathcal{V}$. Although grid networks are quite regular in terms of the internal structure, they do not satisfy the average detectability condition because of the irregularity at the boundary nodes. However, as the scale of the grid becomes larger, the regularity of the inner nodes overcomes the irregularity of the boundary nodes, i.e., the condition $\lim_{n \leftarrow \infty} \frac{n}{\mathbf{s}^\intercal \mathbf{1}_n} = \infty$ is satisfied. Therefore, by the scaling property of open-loop average estimation, the asymptotic value of average estimation error is smaller for the grid of size 25×40 than the grid of size 10×10. This shows that, for instance, smaller granularity of spatial discretization in reaction diffusion systems implies smaller average estimation error as depicted in Figure 3.8.

3.6 Concluding Remarks

We provided a design criteria of a minimum-order average observer for clustered network systems. The notion of average reconstructability of a clustered network system is

defined as the exponential estimation of average states of clusters at an arbitrary rate by the average observer. On the other hand, the notions of average observability and average detectability are defined through the projected network system. However, we also established their relation to the average observer. Average observability allows for asymptotic estimation of average states of clusters when the gain of average observer is arbitrarily large, whereas average detectability allows for exponential open-loop estimation of average states at a fixed rate depending on the eigenvalues of the projected network system. We provided graph-theoretic interpretations through the inter-cluster and intra-cluster graph topologies of the clustered network system of these notions and showed the following relations:

$$\begin{aligned} \text{average reconstructability} &\Rightarrow \text{average observability} \\ \text{average observability AND average detectability} &\Rightarrow \text{average reconstructability.} \end{aligned}$$

Finally, we showed that scale-free networks are suitable candidates for satisfying the necessary condition of average reconstructability and average observability when the hubs are considered as measured nodes. Moreover, under a mild assumption, an important remark on the scaling property of open-loop average estimation showed that the average estimation error converges to zero when the scale of the network system, which is not average detectable, is very large.

The results presented in this chapter point towards several prospects such as sensor location and clustering problems to achieve average reconstructability, average observability, or average detectability. In this thesis, we have studied the clustering problem in Chapter 5, however, the sensor location problem and a combination of sensor location and clustering are reserved for future work. Another prospect is to combine the notions of average reconstructability, average observability, and average detectability in a single clustered network system, where some clusters are chosen because they satisfy the condition of average detectability and other clusters are chosen because they satisfy the condition of average reconstructability. However, such cluster selection can be based only on intra-cluster graph topology, and dealing with the inter-cluster graph topology of the clustered network system is very challenging.

4

Optimal Average Estimation for Clustered Network Systems

This chapter provides an optimal design of the average observer that minimizes the average estimation error when the design criteria of average observer cannot be met. First, in section 4.1, we present a design that minimizes the effect of average deviation on average estimation error. Second, in section 4.2, we provide sufficient conditions for the stability of the average observer. Finally, section 4.3 describes an algorithm to optimally choose the gain of the average observer that minimizes the average estimation error asymptotically. In short, the optimal design is achieved by minimizing the effect of average deviation on average estimation error while keeping the average observer stable. In section 4.4, we illustrate the effectiveness of our methodology on thermal monitoring of a four-room building.

Contents

When it is not possible to asymptotically estimate the average states of clusters in a clustered network system, we resort to an optimal design of the average observer that minimizes the average estimation error. First, we choose a design of the average observer that minimizes the effect of the average deviation vector acting as a structured unknown input in the dynamics of the average estimation error. Then, we perturb the average observer through a gain parameter to stabilize the average observer. The optimal gain parameter is found by solving a convex $\mathcal{H}_2$-optimal average estimation problem through gradient descent or incremental search algorithm. Finally, we show the efficacy of our methodology through the application example of a building thermal system.

4.1 Minimizing the Effect of Average Deviation

Recall a clustered network system $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ with the set of measured nodes $\mathcal{V}_1$ and the clustering $\mathcal{Q} = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$ of unmeasured nodes $\mathcal{V}_2$. Our aim is to estimate the average states of clusters $\mathcal{C}_\alpha$, for $\alpha = 1, \ldots, k$, through the average observer $\mathbf{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$ given in (3.1). We derived the dynamics of average estimation error $\boldsymbol{\zeta}(t) \in \mathbb{R}^k$ in (3.7) as

$$\dot{\boldsymbol{\zeta}}(t) = R_L Q \boldsymbol{\zeta}(t) + R_L \boldsymbol{\sigma}(t) \tag{4.1}$$

where $\boldsymbol{\sigma}(t)$ is the average deviation vector satisfying $Q^+\boldsymbol{\sigma} \equiv \mathbf{0}_k$ and $R_L = Q^+A_{22} - LA_{12}$ as given in (3.5). In the following, we provide a necessary condition of average reconstructability in relation to the average deviation vector.

Lemma 4.1. If the clustered network system $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average reconstructable, then the following equivalent conditions hold:

(i) $R_L\boldsymbol{\sigma} \equiv \mathbf{0}_k$

(ii) $\ker(R_L) \supseteq \ker(Q^+)$

(iii) $R_L = VQ^+$

where, for some Hurwitz $V \in \mathbb{R}^{k\times k}$, the design matrix L of the average observer $\mathbf{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$ is chosen according to (3.9) as $L = (Q^+A_{22} - VQ^+)A_{12}^\dagger$.

Proof. Assume that $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average reconstructable, then, for every $\boldsymbol{\zeta}(0) = \boldsymbol{\zeta}_0 \in \mathbb{R}^k$ and $\gamma > 0$, the average estimation error $\boldsymbol{\zeta}(t)$ satisfies $\|\boldsymbol{\zeta}(t)\| \leq \|\boldsymbol{\zeta}_0\|e^{-\gamma t}\ \forall t \in \mathbb{R}_{\geq 0}$. In other words, $\lim_{t\to\infty}\|\boldsymbol{\zeta}(t)\| = 0$. However, the solution of (4.1)

$$\boldsymbol{\zeta}(t) = \exp(R_LQt)\boldsymbol{\zeta}_0 + \int_0^t \exp[R_LQ(t-\tau)]R_L\boldsymbol{\sigma}(\tau)d\tau$$

where

$$\limsup_{t\to\infty}\|\boldsymbol{\zeta}(t)\| = \limsup_{t\to\infty}\|\int_0^t \exp[R_LQ(t-\tau)]R_L\boldsymbol{\sigma}(\tau)d\tau\|$$

under the assumption that $R_LQ = V$ is Hurwitz. Since matrix exponential is always nonsingular and $\boldsymbol{\sigma}(t)$ is not equal to zero necessarily, the right hand side of the above

equation is equal to zero only if $R_L\boldsymbol{\sigma}(\tau) = 0$ for all $\tau \in [0, \infty)$, thus proving the necessity of (i).

To establish the equivalence (i) $\Leftrightarrow$ (ii), notice that $\boldsymbol{\sigma}(t) \in \ker(Q^+)$ for all t. Thus, if (i) holds, then $\ker(Q^+) \subseteq \ker(R_L)$. In the other direction, if (ii) holds, then, for every vector $\mathbf{v} \in \ker(Q^+)$, we have $R_L\mathbf{v} = \mathbf{0}_k$. Since $\boldsymbol{\sigma}(t) \in \ker(Q^+)$, therefore $R_L\boldsymbol{\sigma}(t) = \mathbf{0}_k$, for every $t \in \mathbb{R}_{\geq 0}$.

To establish the equivalence (ii) $\Leftrightarrow$ (iii), assume (ii). Then, $\mathtt{im}(R_L^\intercal) \subseteq \mathtt{im}(Q^{+\intercal})$, where

$$\begin{aligned} \mathtt{im}(R_L^\intercal) &= \{\mathbf{v}_1 \in \mathbb{R}^n : \mathbf{w}_1^\intercal R_L = \mathbf{v}_1^\intercal, \text{ for } \mathbf{w}_1 \in \mathbb{R}^k\} \\ \mathtt{im}(Q^{+\intercal}) &= \{\mathbf{v}_2 \in \mathbb{R}^n : \mathbf{w}_2^\intercal Q^+ = \mathbf{v}_2^\intercal, \text{ for } \mathbf{w}_2 \in \mathbb{R}^k\}. \end{aligned}$$

That is,

$$\mathtt{rank}\left(\begin{bmatrix} Q^+ \\ R_L \end{bmatrix}\right) = \mathtt{rank}(Q^+)$$

which implies that there exists $V \in \mathbb{R}^{k\times k}$ such that $VQ^+ = R_L$. In the other direction, suppose (iii) holds, then the rows of R_L are linearly dependent on the rows of Q^+. This implies that $\mathtt{im}(R_L^\intercal) \subseteq \mathtt{im}(Q^{+\intercal})$, which is equivalent to (ii).

□

The above lemma shows that average reconstructability cancels the effect of average deviation $\boldsymbol{\sigma}(t)$ from the dynamics of average estimation error (4.1) through the choice of design matrix L. However, when the clustered network system is not average reconstructable, then $R_L\boldsymbol{\sigma}$ may not be equal to zero. Due to the equivalence of Lemma 4.1(i) and (iii), we first find an optimal design matrix L^* such that $\|R_L\boldsymbol{\sigma}(t)\|$ is minimized by minimizing $|||R_L - VQ^+|||$ for a fixed matrix V.

Lemma 4.2. For any $V \in \mathbb{R}^{k\times k}$,

$$L^* = (Q^+A_{22} - VQ^+)A_{12}^\dagger \tag{4.2}$$

is the minimizing solution to

$$\min_{L\in\mathbb{R}^{k\times m}} |||R_L - VQ^+|||.$$

Proof. Consider the equation $R_L = VQ^+$, for some $V \in \mathbb{R}^{k\times k}$, whose least-square solution is given by $L = L^*$ with L^* in (4.2), which is the minimizing solution of $|||R_L - VQ^+|||$, [Campbell2009].

□

The expression for the design matrix L^* in (4.2) is dependent on the matrix V. Now, choosing $L = L^*$, we find optimal matrix V.

Lemma 4.3. Consider L^* in (4.2) and

$$R_V := R_{L^*} = Q^+A_{22}(I_n - A_{12}^\dagger A_{12}) + VQ^+A_{12}^\dagger A_{12}.$$

Then,

$$V^* = Q^+A_{22}Q \tag{4.3}$$

is the minimizing solution to

$$\min_{V\in\mathbb{R}^{k\times k}} |||R_V - VQ^+|||.$$

Proof. Consider the equation $R_V = VQ^+$, then

$$R_V - VQ^+ = (Q^+A_{22} - VQ^+)(I_n - A_{12}^\dagger A_{12})$$

where the ideal solution V that satisfies $R_V - VQ^+ = 0_{k\times n}$ must be such that

$$\texttt{ker}(Q^+A_{22} - VQ^+) \supseteq \texttt{ker}(A_{12})$$

because the columns of $I_n - A_{12}^\dagger A_{12}$ form a complete basis of $\texttt{ker}(A_{12})$. This implies that $Q^+A_{22} - VQ^+ = WA_{12}$ for some $W \in \mathbb{R}^{k\times m}$. However, if the ideal solution does not exist, the minimizing solution is the least-square solution $V = (Q^+A_{22} - WA_{12})Q$, which implies

$$R_V - VQ^+ = (Q^+A_{22}(I_n - QQ^+) + WA_{12}QQ^+)(I_n - A_{12}^\dagger A_{12}).$$

Finally, the minimizing solution to

$$\min_{W\in\mathbb{R}^{k\times m}} \left|\!\left|\!\left| Q^+A_{22}(I_n - QQ^+) + WA_{12}QQ^+ \right|\!\right|\!\right|$$

is $W = Q^+A_{22}(QQ^+ - I_n)QQ^+A_{12}^+ = 0_{k\times m}$ because $(QQ^+ - I_n)Q = 0_{n\times k}$. Thus, $V = V^* = Q^+A_{22}Q$ is the minimizing solution to $\min_{V\in\mathbb{R}^{k\times k}} |||R_V - VQ^+|||$. □

The above lemma shows that the choice of $L = (Q^+A_{22} - VQ^+)A_{12}^\dagger$ with $V = Q^+A_{22}Q$ minimizes the effect of average deviation in (4.1). However, such a choice of V may not ensure the stability of average estimation error or average observer, which is characterized by $M_L = R_LQ$ being Hurwitz. In the next section, we consider a perturbed solution $V = \rho V^*$ and find ρ that stabilizes the average observer while also minimizing the average estimation error.

4.2 On the Stabilizability of Average Observer

We provide a methodology and a sufficient condition for the stabilizability of the average observer or average estimation error. The stabilizability is achieved by making the state matrix $M = R_LQ$ of the average observer Hurwitz.

4.2.1 Preliminary lemmas on the stability of matrices

Recall that a matrix $X \in \mathbb{R}^{n\times n}$ is said to be Hurwitz if its eigenvalues are in the open left-half complex plane, i.e., $\texttt{eig}(X) \subset \mathbb{C}_{<0}$. Moreover, a symmetric matrix $P = P^\intercal \in \mathbb{R}^{n\times n}$ is said to be negative definite if, for every $\mathbf{v} \in \mathbb{R}^n$, we have $\mathbf{v}^\intercal P\mathbf{v} < 0$. It is well-known that $P = P^\intercal \in \mathbb{R}^{n\times n}$ is negative definite if and only if its eigenvalues are in the open left-half complex plane, i.e., $\texttt{eig}(P) \subset \mathbb{C}_{<0}$. In other words, a symmetric Hurwitz matrix is negative definite, and vice versa. A symmetric $P = P^\intercal \in \mathbb{R}^{n\times n}$, on the other hand, is said to be positive definite if, for every $\mathbf{v} \in \mathbb{R}^n$, we have $\mathbf{v}^\intercal P\mathbf{v} > 0$, which is equivalent to have $\texttt{eig}(P) \subset \mathbb{C}_{>0}$. We will use the notation $P < 0$ if P is a negative definite matrix and $P > 0$ if P is a positive definite matrix.

Lemma 4.4 (Lyapunov's theorem)**.** A matrix $X \in \mathbb{R}^{n\times n}$ is Hurwitz *if and only if* there exists a positive definite $P = P^\intercal \in \mathbb{R}^{n\times n}$ such that $PX + X^\intercal P$ is negative definite.

Proof. This is a well-known result attributed to Lyapunov and its proof can be found in [Roger1991, Theorem 2.2.1].

□

The following lemma is also a well-known result, see [Arrow1958, Ostrowski1962, Carlson1968], which is commonly known as S-stability for real matrices and H-stability for complex matrices. This is because it deals with the stability of a matrix that is a product of a Hurwitz matrix and a Symmetric (S) matrix.

Lemma 4.5 (S-stability)**.** If $X + X^\intercal \in \mathbb{R}^{n\times n}$ is negative definite and $S = S^\intercal \in \mathbb{R}^{n\times n}$ is positive definite, then XS is Hurwitz.

Proof. If $X + X^\intercal < 0$ and $S = S^\intercal > 0$, then we have

$$S^\intercal(X + X^\intercal)S = S(X + X^\intercal)S < 0.$$

This is because $S(X+X^\intercal)S$ is congruent to the matrix $X+X^\intercal$ and, according to Sylvester's law of inertia, congruence preserves the inertia of a matrix [Horn2013, Theorem 4.5.8]. In other words, the inertia of $X + X^\intercal \in \mathbb{R}^{n\times n}$ is given by $\iota(X + X^\intercal) = (0, n, 0)$, which means it has $0, n, 0$ eigenvalues in $\mathbb{C}_{>0}, \mathbb{C}_{<0}, \mathbb{C}_{=0}$, respectively, where $\mathbb{C}_{>0}$ denotes the open right-half complex plane, $\mathbb{C}_{<0}$ denotes the open left-half complex plane, and $\mathbb{C}_{=0}$ denotes the imaginary axis of complex plane. Thus, the inertia of $S(X + X^\intercal)$ is also $\iota(S(X + X^\intercal)S) = (0, n, 0)$. Therefore,

$$S(X + X^\intercal)S = S(XS) + (XS)^\intercal S < 0$$

which concludes that XS is Hurwitz by Lyapunov's theorem (Lemma 4.4).

□

For a full-column rank matrix $Q \in \mathbb{R}^{n\times k}$, we say Q^+XQ is the compression of $X \in \mathbb{R}^{n\times n}$ with respect to Q, where $Q^+Q = I_k$. When the context is clear, we simply say "compression of X" instead of "compression of X with respect to Q".

If X is a Hurwitz matrix, then the stability of compression Q^+XQ depends on the matrix Q. In the following, we provide a sufficient condition for the stability of compression of a Hurwitz matrix X for any full-column rank matrix $Q \in \mathbb{R}^{n\times k}$.

Lemma 4.6. Let $X \in \mathbb{R}^{n\times n}$ be a Hurwitz matrix. If $X+X^\intercal$ is negative definite, then, for every $Q \in \mathbb{R}^{n\times k}$ such that $\texttt{rank}(Q) = k$ with $k < n$, the matrix Q^+XQ is Hurwitz.

Proof. The compression of $X+X^\intercal$ by Q is given by $Q^\intercal(X+X^\intercal)Q$. By Cauchy's interlacing theorem, [Bhatia1997, Corollary III.1.5], we have

$$Q^\intercal(X + X^\intercal)Q = Q^\intercal XQ + (Q^\intercal XQ)^\intercal < 0$$

because $X + X^\intercal < 0$ by assumption. Since $(Q^\intercal Q)^{-1} > 0$, therefore, by Lemma 4.5, the matrix $Q^+XQ = (Q^\intercal Q)^{-1}Q^\intercal XQ$ is Hurwitz.

□

For the stabilization of average observer, we first show that the optimal solution $V^* = Q^+A_{22}Q$ minimizing the effect of average deviation vector is Hurwitz for all $Q \in \mathfrak{C}_{n,k}$, where

$$\mathfrak{C}_{n,k} = \{X \in \{0, 1\}^{n\times k} : X\mathbf{1}_k = \mathbf{1}_n\}$$

is the set of characteristic matrices of all clusterings with k clusters of n nodes. In the next subsection, we show the stability of V^* by using the above lemmas.

4.2.2 Sufficient condition for the stability of V^*

Recall the digraph $\mathcal{G} = (\{\mathcal{V}_1, \mathcal{V}_2\}, \mathcal{E})$ describing the structure of the clustered network system $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$, where $\mathcal{V}_1$ is the set of m measured nodes, $\mathcal{V}_2$ is the set of n unmeasured nodes, $\mathcal{E}$ is the set of directed edges, and $\mathcal{Q}$ is the clustering of $\mathcal{V}_2$. Consider $\mathcal{G}_\nu = (\mathcal{V}_2, \mathcal{E}_\nu)$ to be the induced subgraph formed by the unmeasured nodes $\mathcal{V}_2$, where $\mathcal{E}_\nu = \mathcal{E} \cap (\mathcal{V}_2 \times \mathcal{V}_2)$. The off-diagonal entries of the matrix A_{22} constitute the edge configuration of $\mathcal{G}_\nu$. Therefore, the subgraph $\mathcal{G}_\nu$ is weakly connected if and only if the matrix $A_{22} + A_{22}^\intercal$ is irreducible. A matrix is said to be reducible if it can be transformed to a block upper-triangular form by simultaneous row/column permutations. Otherwise, it is said to be irreducible.

The weak connectivity of induced subgraph $\mathcal{G}_\nu$ can also be established by considering an undirected version $\overline{\mathcal{G}}_\nu$ of the subgraph. The edges of $\overline{\mathcal{G}}_\nu$ are obtained by ignoring the directions from the edges of $\mathcal{G}_\nu$. Then, $\mathcal{G}_\nu$ is weakly connected if and only if $\overline{\mathcal{G}}_\nu$ is connected, which is equivalent to having the rank of its Laplacian matrix $\texttt{rank}(\mathcal{L}(\overline{\mathcal{G}}_\nu) = n - 1$.

For every unmeasured node $i \in \mathcal{V}_2$, let

$$s_i = \sum_{j \in \mathcal{N}_{i \leftarrow \mathcal{V}_2}} a_{ij} + \sum_{h \in \mathcal{N}_{i \rightarrow \mathcal{V}_2}} a_{hi}$$

to be the sum of the weights of all edges going into and emerging from i within $\mathcal{G}_\nu$, where $\mathcal{N}_{i \leftarrow \mathcal{V}_2}$ and $\mathcal{N}_{i \rightarrow \mathcal{V}_2}$ are the sets of in-neighbors and out-neighbors of i. That is, the weights of all edges of the node ν_i's in-neighbors and out-neighbors. Finally, recall that all the diagonal entries of A_{22} are non-positive, i.e., $[A_{22}]_{jj} = a_{jj} \leq 0$ for $j = 1, \ldots, n$.

Theorem 4.7. If

(i) the induced subgraph $\mathcal{G}_\nu = (\mathcal{V}_2, \mathcal{E}_\nu)$ is weakly connected and

(ii) for every unmeasured node $i \in \mathcal{V}_2$, we have $s_i \leq 2|a_{ii}|$, and, for at least one $j \in \mathcal{V}_2$, it holds $s_j < 2|a_{jj}|$

then, for any $Q \in \mathfrak{C}_{n,k} \subset \{0,1\}^{n \times k}$, the matrix $V^* = Q^+ A_{22} Q$ is Hurwitz.

Proof. First, if (i) holds, then the symmetric part of A_{22}, $\mathcal{S}(A_{22}) = A_{22} + A_{22}^\intercal$, is irreducible. That is, an undirected graph $\overline{\mathcal{G}}_\nu$ capturing the structure of $\mathcal{S}(A_{22})$ is connected. Thus, the Laplacian matrix of $\overline{\mathcal{G}}_\nu$ defined as

$$[\mathcal{L}(\overline{\mathcal{G}}_\nu)]_{ij} = \begin{cases} s_i, & \text{if } i = j \\ a_{ij} + a_{ji}, & \text{if } i \neq j \end{cases}$$

is of rank $n - 1$ and nullity 1. Since $\mathcal{L}(\overline{\mathcal{G}}_\nu)$ is positive semi-definite, we have, for every $\mathbf{v} \in \mathbb{R}^n$, $\mathbf{v}^\intercal \mathcal{L}(\overline{\mathcal{G}}_\nu)\mathbf{v} \geq 0$. Moreover, $0 \in \texttt{eig}(\mathcal{L}(\overline{\mathcal{G}}_\nu))$ with algebraic multiplicity 1 because $\overline{\mathcal{G}}_\nu$ is connected, therefore we have $\mathbf{v}^\intercal \mathcal{L}(\overline{\mathcal{G}}_\nu)\mathbf{v} = 0$ if and only if $\mathbf{v} = \mathbf{1}_n$, which is in the direction of eigenvector of $\mathcal{L}(\overline{\mathcal{G}}_\nu)$ corresponding to the 0 eigenvalue.

Second, if (ii) holds, then

$$\mathcal{S}(A_{22}) = -\mathcal{L}(\overline{\mathcal{G}}_\nu) - \mathcal{D}$$

where $\mathcal{D} = \texttt{diag}(2|a_{11}| - s_1, \ldots, 2|a_{nn}| - s_n)$ is a diagonal matrix, which is positive semi-definite because, for all $i \in \{1, \ldots, n\}$, we have $s_i \leq 2|a_{ii}|$ and, for at least one $j \in \{1, \ldots, n\}$, we have $s_j < 2|a_{jj}|$. Thus, for every $\mathbf{v} \in \mathbb{R}^n$, we have $\mathbf{v}^\intercal \mathcal{D} \mathbf{v} \geq 0$. However, we know that $\mathbf{1}_n^\intercal \mathcal{D} \mathbf{1}_n > 0$ and, for some $\mathbf{v}_1 \in \mathbb{R}^n$ such that $\mathbf{v}_1^\intercal \mathcal{D} \mathbf{v} = 0$, we have $\mathbf{v}_1^\intercal \mathcal{L}(\overline{\mathcal{G}}_\nu)\mathbf{v}_1 > 0$. Therefore, $\mathcal{L}(\overline{\mathcal{G}}_\nu) + \mathcal{D}$ is positive definite, i.e.,

$$\mathcal{S}(A_{22}) = A_{22} + A_{22}^\intercal = -(\mathcal{L}(\overline{\mathcal{G}}_\nu) + \mathcal{D})$$

is negative definite. Therefore, for any $Q \in \mathfrak{C}_{n,k}$, we have $Q^+A_{22}Q$ Hurwitz by Lemma 4.5 and 4.6.

□

In the next subsection, under the assumption that the sufficient condition of Theorem 4.7 holds, we perturb the matrix V^* in order to stabilize the average observer and provide a sufficient condition of stabilizability.

4.2.3 Stabilizability of average observer

The optimal design matrix $L^* \in \mathbb{R}^{k\times m}$ of the average observer $\mathbf{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$ that minimizes the effect of the average deviation from the average estimation error $\boldsymbol{\zeta}(t)$ is given by $L^* = (Q^+A_{22} - V^*Q^+)A_{12}^\dagger$, where $V^* = Q^+A_{22}Q$. However, such a choice $L = L^*$ may not ensure the stability of the average estimation error $\boldsymbol{\zeta}(t)$, which is to say that the matrix R_LQ may not be Hurwitz. Thus, instead of considering the optimal solution V^* that minimizes the effect of average deviation from $\boldsymbol{\zeta}(t)$, we consider a perturbed solution

$$V = \rho V^*, \text{ where } \rho \in \mathbb{R} \tag{4.4}$$

in order to ensure the stability of $\boldsymbol{\zeta}(t)$ and find optimal design in terms of ρ that minimizes the $\boldsymbol{\zeta}(t)$ asymptotically. Notice that if $V^* = Q^+A_{22}Q$ is Hurwitz, then $V = \rho V^*$ is also Hurwitz for all $\rho \in \mathbb{R}_{>0}$. To elucidate, let $\lambda_1, \dots, \lambda_k$ to be the eigenvalues of V^*, then $\rho\lambda_1, \dots, \rho\lambda_k$ are the eigenvalues of ρV^*.

The perturbed solution $V = \rho V^*$ gives

$$L =: L_\rho = (Q^+A_{22} - \rho Q^+A_{22}QQ^+)A_{12}^\dagger \tag{4.5}$$

and

$$\begin{aligned} R_L =: R_\rho &= Q^+A_{22} - L_\rho A_{12} \\ &= Q^+A_{22}(I_n - A_{12}^\dagger A_{12}) + \rho Q^+A_{22}QQ^+A_{12}^\dagger A_{12}. \end{aligned} \tag{4.6}$$

Thus, the dynamics of the average estimation error is given by

$$\dot{\boldsymbol{\zeta}}(t) = M_\rho \boldsymbol{\zeta}(t) + R_\rho \boldsymbol{\sigma}(t) \tag{4.7}$$

where

$$M_\rho = R_\rho Q = Q^+A_{22}(I_n - A_{12}^\dagger A_{12})Q + \rho Q^+A_{22}QQ^+A_{12}^\dagger A_{12}Q. \tag{4.8}$$

Lemma 4.8. If a matrix $X \in \mathbb{R}^{k\times k}$ is Hurwitz, then, for any matrix $Y \in \mathbb{R}^{k\times k}$, there exists $\phi \in \mathbb{R}$ such that $\rho X + Y$ is Hurwitz for every $\rho > \phi$.

Proof. By Lemma 4.4, X is Hurwitz if and only if there exists a positive definite $P = P^\intercal \in \mathbb{R}^{k\times k}$ such that $PX + X^\intercal P < 0$. For such a $P > 0$, we have

$$P(\rho X + Y) + (\rho X + Y)^\intercal P = \rho(PX + X^\intercal P) + (PY + Y^\intercal P) < 0 \tag{4.9}$$

if, for every $\mathbf{v} \in \mathbb{R}^k$,

$$\rho\mathbf{v}^\intercal(PX + X^\intercal P)\mathbf{v} < -\mathbf{v}^\intercal(PY + Y^\intercal P)\mathbf{v}.$$

Since $PX + X^\intercal P$ is negative definite, we have $\mathbf{v}^\intercal(PX + X^\intercal P)\mathbf{v} < 0$ for every $\mathbf{v} \in \mathbb{R}^k$ and $\mathbf{v} \neq \mathbf{0}_k$. Therefore, dividing both sides of the above inequality by $\mathbf{v}^\intercal(PX + X^\intercal P)\mathbf{v}$ changes the sign of inequality and gives

$$\rho > -\frac{\mathbf{v}^\intercal(PY + Y^\intercal P)\mathbf{v}}{\mathbf{v}^\intercal(PX + X^\intercal P)\mathbf{v}} \tag{4.10}$$

which satisfies (4.9). Let

$$\begin{aligned} \mathbf{v}_1 &= \arg \min_{\mathbf{v}\in\mathbb{R}^k, \|\mathbf{v}\|=1} \quad |\mathbf{v}^\intercal (PX + X^\intercal P)\mathbf{v}| \\ \mathbf{v}_2 &= \arg \max_{\mathbf{v}\in\mathbb{R}^k, \|\mathbf{v}\|=1} \quad \mathbf{v}^\intercal (PY + Y^\intercal P)\mathbf{v}. \end{aligned} \tag{4.11}$$

Then, choose

$$\phi = \frac{\mathbf{v}_2^\intercal (PY + Y^\intercal P)\mathbf{v}_2}{|\mathbf{v}_1^\intercal (PX + X^\intercal P)\mathbf{v}_1|}.$$

From (4.11), we have, for every $\mathbf{v} \in \mathbb{R}^k$,

$$\begin{aligned} |\mathbf{v}_1^\intercal (PX + X^\intercal P)\mathbf{v}_1| &\leq |\mathbf{v}^\intercal (PX + X^\intercal P)\mathbf{v}| \\ \mathbf{v}_2^\intercal (PY + Y^\intercal P)\mathbf{v}_2 &\geq \mathbf{v}^\intercal (PY + Y^\intercal P)\mathbf{v} \end{aligned}$$

which implies

$$\phi \geq \frac{\mathbf{v}^\intercal (PY + Y^\intercal P)\mathbf{v}}{|\mathbf{v}^\intercal (PX + X^\intercal P)\mathbf{v}|}.$$

Therefore, choosing $\rho > \phi$ ensures (4.10), which concludes the proof.

□

The above result implies that given two matrices $X, Y \in \mathbb{R}^{k\times k}$, we can choose ρ such that it satisfies (4.10) to ensure that $\rho X + Y$ is a Hurwitz matrix.

Theorem 4.9. Let Assumption 2.1 hold and assume $V^* = Q^+ A_{22} Q$ to be Hurwitz. If $\mathtt{rank}(A_{12}Q) = k$, then there exists $\phi \in \mathbb{R}$ such that $M_\rho = R_\rho Q$ is Hurwitz for every $\rho > \phi$, where R_ρ is given in (4.6).

Proof. If $\mathtt{rank}(A_{12}Q) = k$ and that Assumption 2.1 holds, i.e., $\mathtt{rank}(A_{12}) = m$, where $A_{12} \in \mathbb{R}^{m\times n}_{\geq 0}$, then

$$\begin{aligned} \mathtt{rank}(A_{12}Q) &= \mathtt{rank}((A_{12}A_{12}^\dagger)^{-\frac{1}{2}} A_{12}Q) \\ &= \mathtt{rank}(Q^\intercal A_{12}^\intercal (A_{12}A_{12}^\dagger)^{-1} A_{12}Q) \\ &= \mathtt{rank}((Q^\intercal Q)^{-1} Q^\intercal A_{12}^\intercal (A_{12}A_{12}^\dagger)^{-1} A_{12}Q) \\ &= \mathtt{rank}(Q^+ A_{12}^\dagger A_{12}Q) \\ &= k \end{aligned}$$

where we used the properties $\mathtt{rank}(X^\intercal X) = \mathtt{rank}(X)$ and $\mathtt{rank}(YX) = \mathtt{rank}(X)$, for some matrix $X \in \mathbb{R}^{a\times b}$ and a non-singular $Y \in \mathbb{R}^{a\times a}$. This implies that

$$S := Q^\intercal A_{12}^\dagger A_{12} Q = [(A_{12}A_{12}^\dagger)^{-\frac{1}{2}} A_{12}Q]^\intercal (A_{12}A_{12}^\dagger)^{-\frac{1}{2}} A_{12}Q$$

is positive definite.

Notice that we can write $M_\rho = R_\rho Q$ in (4.8) as

$$M_\rho = \rho X S + Y$$

where

$$\begin{aligned} X &= Q^+ A_{22} Q (Q^\intercal Q)^{-1} \\ Y &= Q^+ A_{22} (I_n - A_{12}^\dagger A_{12}) Q \end{aligned}$$

Then, notice that $X = V^*(Q^\intercal Q)^{-1}$ is Hurwitz by Lemma 4.5 and it holds that

$$X + X^\intercal = Q^+(A_{22} + A_{22}^\intercal)Q^{+\intercal} < 0.$$

Therefore, again by Lemma 4.5 and the fact that $S > 0$, we have that XS is Hurwitz. Finally, by Lemma 4.8, there exists $\phi \in \mathbb{R}$ such that $M_\rho = \rho XS + Y$ is Hurwitz for every $\rho > \phi$.

□

The above theorem provides a sufficient condition for the stabilizability of the matrix M_ρ, which is the state matrix of the average observer $\mathbf{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$ and characterizes its stabilizability. Notice that the sufficient condition of Theorem 4.9 is contingent on the sufficient condition of Theorem 4.7. If both conditions are satisfied, then the average observer is stabilizable. In the next section, we provide an algorithm to optimally choose ρ that minimizes the average estimation error.

Furthermore, the sufficient condition of Theorem 4.9 also provides an upper bound on the number of clusters k.

Corollary 4.9.1. If $\texttt{rank}(A_{12}Q) = k$, then the number of clusters k is less than or equal to the number of measured nodes m, i.e., $k \leq m$.

Proof. Note that for any matrix $X \in \mathbb{R}^{m\times k}$, we have $\texttt{rank}(X) \leq \min(m, k)$. Therefore, requiring that $A_{12}Q \in \mathbb{R}^{m\times k}$ is a matrix with full-column rank implies that the number of rows m is greater than or equal to the number of columns k of $A_{12}Q$.

□

4.3 $\mathcal{H}_2$-Optimal Average Estimation

In this section, we formulate an $\mathcal{H}_2$-optimal average estimation problem with respect to the gain (or perturbation) parameter ρ. The problem assumes the following:

Assumption 4.1. The clustered network system $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is such that the sufficient conditions of Theorem 4.7 and 4.9 are satisfied.

The above assumption ensures that the average observer $\mathbf{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$ is stabilizable, which suffices for the feasibility of $\mathcal{H}_2$-optimal average estimation problem defined in the following subsection.

4.3.1 Problem definition

Recall the dynamics of average estimation error $\boldsymbol{\zeta}(t)$ given in (4.7), $M_\rho \in \mathbb{R}^{k\times k}$ given in (4.8), and $R_\rho \in \mathbb{R}^{k\times n}$ given in (4.6). Let

$$\mathbf{T}_\rho(s) = (sI_k - M_\rho)^{-1}R_\rho$$

be the transfer matrix from $\boldsymbol{\sigma}$ to $\boldsymbol{\zeta}$, where the impulse response is given by

$$T_\rho(t) = \mathcal{L}^{-1}[\mathbf{T}_\rho(s)] = \exp(M_\rho t)R_\rho$$

with $\mathcal{L}^{-1}$ denoting the inverse Laplace transform. Then, the $\mathcal{H}_2$-norm defined as

$$\|\mathbf{T}_\rho(s)\|_{\mathcal{H}_2} = \sqrt{\texttt{trace}\left(\frac{1}{2\pi}\int_0^\infty \mathbf{T}(\iota\omega)\mathbf{T}(\iota\omega)^* d\omega\right)}$$

can be computed, due to Parseval's theorem, as

$$\begin{aligned}\|\mathbf{T}_\rho(s)\|_{\mathcal{H}_2} &= \sqrt{\mathtt{trace}\left(\int_0^\infty T_\rho(t)T_\rho^\intercal(t)dt\right)} \\ &= \sqrt{\mathtt{trace}\,(W_\rho)}.\end{aligned}$$

where

$$W_\rho := \int_0^\infty \exp(M_\rho t)R_\rho R_\rho^\intercal \exp(M_\rho^\intercal t)dt \tag{4.12}$$

is the controllability gramian of (M_ρ, R_ρ).

Define the cost $\mathcal{J}(\rho) := \|\mathbf{T}_\rho(s)\|_{\mathcal{H}_2}^2 = \mathtt{trace}(W_\rho)$, then $\mathcal{H}_2$-optimal average estimation problem is defined as follows: Find $\rho^* \in \mathbb{R}$ such that

$$\rho^* = \arg\min_{\rho\in\mathbb{R}} \mathcal{J}(\rho) \quad \text{subject to } M_\rho \text{ is Hurwitz.} \tag{4.13}$$

Note that the problem (4.13) is a convex optimization problem with a scalar decision variable, see [Boyd1994, Chapter 3] and [Boyd2004, Chapter 4].

4.3.2 Gradient descent algorithm

The gradient descent algorithm is given by

$$\hat{\rho}_{i+1} = \hat{\rho}_i - \eta\nabla\mathcal{J}(\hat{\rho}_i) \tag{4.14}$$

where $\hat{\rho}_i \in \mathbb{R}_{>0}$, $\eta \in \mathbb{R}_{>0}$ is the step size, and $\nabla\mathcal{J}(\hat{\rho}_i) = \frac{d}{d\rho}\mathcal{J}(\rho)|_{\rho=\hat{\rho}_i}$ is the gradient of $\mathcal{J}(\rho)$ evaluated at $\hat{\rho}_i$. Since (4.13) is a convex optimization problem, the algorithm (4.14) converges to the global minimum given that $\eta > 0$ is small enough.

The initial estimate ρ_0 is chosen such that the matrix $M_{\hat{\rho}_0}$ is Hurwitz because otherwise the cost $\mathcal{J}(\hat{\rho}_0)$ is infinite. Let

$$M_\rho = \rho XS + Y$$

where

$$\begin{aligned}X &= Q^+A_{22}Q^{+\intercal} \\ S &= Q^\intercal A_{12}^\dagger A_{12}Q \\ Y &= Q^+A_{22}(I_n - A_{12}^\dagger A_{12})Q.\end{aligned} \tag{4.15}$$

Under Assumption 4.1, the matrix product XS is Hurwitz and there exists $\phi \in \mathbb{R}$ such that M_ρ is Hurwitz for every $\rho > \phi$. Since XS is Hurwitz, there exists $P > 0$ such that $P(XS) + (XS)^\intercal P < 0$. Let $T \in \mathbb{R}^{k\times k}$ be any arbitrary positive definite matrix, then

$$P = \int_0^\infty \exp((XS)^\intercal t)T\exp((XS)t)dt.$$

satisfies $P(XS) + (XS)^\intercal P = -T$. Define

$$\begin{aligned}\mathbf{v}_1 &= \arg\min_{\mathbf{v}\in\mathbb{R}^k} |\mathbf{v}^\intercal(P(XS) + (XS)^\intercal P)\mathbf{v}| = \arg\min_{\mathbf{v}\in\mathbb{R}^k} \mathbf{v}^\intercal T\mathbf{v} \\ \mathbf{v}_2 &= \arg\max_{\mathbf{v}\in\mathbb{R}^k} \mathbf{v}^\intercal(PY + Y^\intercal P)\mathbf{v}.\end{aligned}$$

Then, consider

$$\hat{\rho}_0 = \frac{\mathbf{v}_2^\intercal(PY + Y^\intercal P)\mathbf{v}_2}{\mathbf{v}_1 T\mathbf{v}_1} + \varepsilon \tag{4.16}$$

for some small $\varepsilon > 0$. This choice of $\hat{\rho}_0$ ensures that $M_{\hat{\rho}_0}$ is Hurwitz in the initial iteration and thus yielding a finite cost $\mathcal{J}(\hat{\rho}_0)$.

The gradient $\nabla\mathcal{J}(\rho)$ is computed as

$$\nabla\mathcal{J}(\rho) = \frac{d}{d\rho}\texttt{trace}(W_\rho) = \texttt{trace}\left(\frac{d}{d\rho}W_\rho\right)$$

where

$$\begin{aligned}
\frac{d}{d\rho}W_\rho &= \frac{d}{d\rho}\int_0^\infty \exp(M_\rho t)R_\rho R_\rho^\intercal \exp(M_\rho^\intercal t)dt \\
&= \int_0^\infty \frac{\partial}{\partial\rho}\left(\exp(M_\rho t)R_\rho R_\rho^\intercal \exp(M_\rho^\intercal t)\right)dt \\
&= \int_0^\infty \Bigg[\left(\frac{\partial}{\partial\rho}\exp(M_\rho t)\right)R_\rho R_\rho^\intercal \exp(M_\rho^\intercal t) + \exp(M_\rho t)R_\rho R_\rho^\intercal\left(\frac{\partial}{\partial\rho}\exp(M_\rho^\intercal t)\right) \\
&\qquad + \exp(M_\rho t)\left(\frac{\partial}{\partial\rho}R_\rho\right)R_\rho^\intercal \exp(M_\rho^\intercal t) + \exp(M_\rho t)R_\rho\left(\frac{\partial}{\partial\rho}R_\rho^\intercal\right)\exp(M_\rho t)\Bigg]dt
\end{aligned}$$

with $\frac{\partial}{\partial\rho}R_\rho = XQ^\intercal A_{12}^\dagger A_{12}$ and $\frac{\partial}{\partial\rho}\exp(M_\rho t)$ computed in the following lemmas.

Lemma 4.10. Let

$$\mathcal{M}_\rho = \begin{bmatrix} M_\rho & \frac{d}{d\rho}M_\rho \\ 0_{k\times k} & M_\rho \end{bmatrix}.$$

Then,

$$\exp(\mathcal{M}_\rho t) = \begin{bmatrix} \exp(M_\rho t) & \frac{\partial}{\partial\rho}\exp(M_\rho t) \\ 0_{k\times k} & \exp(M_\rho t) \end{bmatrix}.$$

Proof. Consider the power series expansion of the matrix exponential

$$\frac{\partial}{\partial\rho}\exp(M_\rho t) = \frac{\partial}{\partial\rho}\left(I_k + M_\rho t + M_\rho^2\frac{t^2}{2!} + M_\rho^3\frac{t^3}{3!} + \dots\right)$$

and note that

$$\begin{aligned}
\frac{d}{d\rho}M_\rho &= XS \\
\frac{d}{d\rho}M_\rho^2 &= XSM_\rho + M_\rho XS \\
\frac{d}{d\rho}M_\rho^3 &= XSM_\rho^2 + M_\rho XSM_\rho + M_\rho^2 XS \\
&\vdots \quad \vdots \quad \vdots \\
\frac{d}{d\rho}M_\rho^\ell &= \sum_{i=0}^{\ell-1} M_\rho^i XSM_\rho^{\ell-1-i}.
\end{aligned}$$

Therefore,

$$\frac{\partial}{\partial\rho}\exp(M_\rho t) = \sum_{\ell=1}^{\infty}\sum_{i=0}^{\ell-1} M_\rho^i XSM_\rho^{\ell-1-i}\frac{t^\ell}{\ell!}. \tag{4.17}$$

Further notice that

$$\begin{aligned}
\mathcal{M}_\rho^2 &= \begin{bmatrix} M_\rho^2 & M_\rho\left(\frac{d}{d\rho}M_\rho\right)+\left(\frac{d}{d\rho}M_\rho\right)M_\rho \\ 0_{k\times k} & M_\rho^2 \end{bmatrix} \\
\mathcal{M}_\rho^3 &= \begin{bmatrix} M_\rho^3 & M_\rho^2\left(\frac{d}{d\rho}M_\rho\right)+M_\rho\left(\frac{d}{d\rho}M_\rho\right)M_\rho+\left(\frac{d}{d\rho}M_\rho\right)M_\rho^2 \\ 0_{k\times k} & M_\rho^3 \end{bmatrix} \\
\vdots \quad &\quad \vdots \qquad\qquad \vdots \\
\mathcal{M}_\rho^\ell &= \begin{bmatrix} M_\rho^\ell & \sum_{i=0}^{\ell-1} M_\rho^i\left(\frac{d}{d\rho}M_\rho\right)M_\rho^{\ell-1-i} \\ 0_{k\times k} & M_\rho^\ell \end{bmatrix}.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\exp(\mathcal{M}_\rho t) &= \begin{bmatrix} \sum_{\ell=0}^{\infty} M_\rho^\ell \frac{t^\ell}{\ell!} & \sum_{\ell=1}^{\infty}\sum_{i=0}^{\ell-1} M_\rho^i\left(\frac{d}{d\rho}M_\rho\right)M_\rho^{\ell-1-i}\frac{t^\ell}{\ell!} \\ 0_{k\times k} & \sum_{\ell=0}^{\infty} M_\rho^\ell\frac{t^\ell}{\ell!} \end{bmatrix} \\
&= \begin{bmatrix} \exp(M_\rho t) & \frac{\partial}{\partial\rho}\exp(M_\rho t) \\ 0_{k\times k} & \exp(M_\rho t) \end{bmatrix}
\end{aligned}$$

by the power series of $\exp(\,\cdot\,)$ and (4.17).

□

The power series expression of (4.17) can be computed analytically using the following lemma, which is inspired by [Dieci2001].

Lemma 4.11. Let $M_\rho = \rho XS + Y$, where X, S, Y are given in (4.15). Then,

$$\frac{\partial \exp(M_\rho t)}{\partial \rho} = \int_0^1 \exp[M_\rho t(1-\tau)]XSt\exp(M_\rho t\tau)d\tau. \tag{4.18}$$

Proof. Consider the power series expansion of the right-hand side expression of (4.18)

$$\begin{aligned}
&\int_0^1 \left(I_k + M_\rho t(1-\tau) + M_\rho^2 t^2\frac{(1-\tau)^2}{2!}+\dots\right)XSt\left(I_k + M_\rho t\tau + M_\rho^2 t^2\frac{\tau^2}{2!}+\dots\right) \\
&\quad= \int_0^1 \Bigg\{XSt + [M_\rho XSt(1-\tau) + XSM_\rho t\tau] \\
&\qquad + \left(M_\rho^2 XSt^2\frac{(1-\tau)^2}{2!} + M_\rho XSM_\rho t^2\tau(1-\tau) + XSM_\rho^2 t^2\frac{\tau^2}{2!}\right)+\dots\Bigg\}d\tau.
\end{aligned}$$

Solving the above integral with respect to τ gives

$$\begin{aligned}
&\Bigg\{XSt\tau + \left[M_\rho XSt\left(\tau-\frac{\tau^2}{2}\right)+XSM_\rho t\frac{\tau^2}{2}\right] \\
&\qquad + \left[M_\rho^2 XSt^2\frac{\frac{\tau^3}{3}-\tau^2+\tau}{2!} + M_\rho XSM_\rho t^2\left(\frac{\tau^2}{2}-\frac{\tau^3}{3}\right)+XSM_\rho^2 t^2\frac{\frac{\tau^3}{3}}{2!}\right]+\dots\Bigg\}\Bigg|_0^1 \\
&\quad= XSt + (M_\rho XS + XSM_\rho)\frac{t^2}{2!} + (M_\rho^2 XS + M_\rho XSM_\rho + XSM_\rho^2)\frac{t^3}{3!}+\dots
\end{aligned}$$

which is equal to (4.17).

□

Using Lemma 4.10 and 4.11, we can compute the gradient $\nabla\mathcal{J}(\rho)$ required in the gradient descent algorithm (4.14).

4.3.3 Incremental search algorithm

The main idea of this algorithm is to initialize $\rho \in \mathbb{R}$ and keep on incrementing with a small $\varepsilon > 0$ to search for the optimal solution. The value of $\varepsilon > 0$ is initialized arbitrarily and then in the algorithm is reduced iteratively in order to achieve the required tolerance level to the actual optimal solution ρ^*. In the algorithm, whenever ρ passes the optimal value, we define a smaller interval around that optimal value, divide the interval into several points, choose ε to be length of these divisions, and search for the optimal solution in this interval. This process is done iteratively until a required tolerance level is achieved.

Algorithm 1 Minimum ϕ such that M_ρ is Hurwitz for $\rho \geq \phi$

Input: Matrices required to compute M_ρ, a small $\varepsilon > 0$, tolerance $\underline{\varepsilon} > 0$, an integer $\eta \geq 2$
Output: Minimum ϕ such that, for $\rho = \phi$, M_ρ is Hurwitz

Initialize ρ to be a negative number such that M_ρ is not Hurwitz
repeat
  Compute M_ρ
  if M_ρ is Hurwitz **then**
    Assign $\phi \leftarrow \rho$, $\rho \leftarrow \rho - \varepsilon$, $\varepsilon \leftarrow \varepsilon/\eta$
  else
    Assign $\rho \leftarrow \rho + \varepsilon$
  end if
until $\eta\varepsilon \leq \underline{\varepsilon}$
return ϕ

Algorithm 2 Incremental search algorithm

Input: Matrices required to compute W_ρ, tolerance $\underline{\varepsilon} > 0$, $\eta \geq 2$, $\varepsilon > 0$, and ϕ
Output: Optimal solution ρ^* to (4.13)

Initialize $\rho = \phi$ and assign $\mathcal{J} \leftarrow \texttt{trace}(W_\rho)$
repeat
  Assign $\rho \leftarrow \rho + \varepsilon$ and $\mathcal{J}_1 \leftarrow \texttt{trace}(W_\rho)$
  if $\mathcal{J}_1 > \mathcal{J}$ **then**
    Assign $\rho^* \leftarrow \rho - \varepsilon$, $\rho \leftarrow \rho - 2\varepsilon$, and $\varepsilon \leftarrow \varepsilon/\eta$
  else
    Assign $\rho^* \leftarrow \rho$ and $\mathcal{J} \leftarrow \mathcal{J}_1$
  end if
until $\eta\varepsilon \leq \underline{\varepsilon}$
return ρ^*.

First, we find the minimum $\phi > 0$ such that, for $\rho = \phi$, we have M_ρ Hurwitz. This is achieved by Algorithm 1. Then, in Algorithm 2, we initialize $\rho = \phi$ and increment by $\varepsilon > 0$ until we pass the optimal solution, which is the global minimum. This is because before the global minimum was reached, the trend of cost $\mathcal{J}(\rho)$ at each iteration is downhill, and when the cost starts increasing, this indicates that ρ has passed the global minimum. At this point, we realize that the solution lies in the interval $[\rho - 2\varepsilon, \rho]$, therefore, we decrement ρ by 2ε, decrease the value of ε by dividing η, and start the search process in the specified interval. This process is repeated until the solution ρ^* is within the specified tolerance $\underline{\varepsilon}$ to the true optimal value.

4.4 Application Example: Thermal Monitoring of Buildings

Residential and commercial buildings play a significant part in the global energy consumption and greenhouse gas emissions. In France, for instance, the residential sector is the second largest source of energy consumption[1], and amounts to 23% of the national greenhouse gas emissions [Derbez2014, Lévy2018]. Within the residential sector of France, the space heating takes a share of 70% of the energy consumption [Hache2017]. Therefore, developing efficient techniques for thermal monitoring and control of residential buildings is one of the crucial forefronts for the fight against global warming. In this section, we aim to develop a thermal monitoring technique based on the $\mathcal{H}_2$-optimal average estimation.

Model-based techniques are considered to be quite effective in thermal monitoring and control of buildings [Oldewurtel2010, Maasoumy2013]. In particular, resistor-capacitor (RC) network models offer an exceptional balance between simplicity and accuracy as evidenced in [Bueno2012, Ramallo-González2013]. However, such models are not tractable because they scale badly with the size of a building and require tremendous amount of computational and sensing equipment. To deal with this issue, a model reduction technique to reduce the dimension of building thermal model is presented in [Deng2010, Deng2014], which provide an aggregated thermal representation of several clusters of building elements. Such a representation, although optimal for model reduction, may not be favorable for thermal monitoring of buildings because the clusters, for instance, may consist of several walls in the building that are not directly linked with each other. Therefore, considering that the clusters of building elements are prespecified, we aim to design an optimal average observer to estimate the mean operative temperature of each cluster to facilitate thermal monitoring and regulation.

4.4.1 Building setup and its RC-network model

We consider a 4-room building setup illustrated in Figure 4.1, which is adopted from [Deng2010, Deng2014] with some changes. Unlike [Deng2010, Deng2014], we do not neglect the internal mass of rooms (furniture, carpet, etc.) and we consider that the outside temperature is an exogenous input to the system. The outside temperature is considered to be an input, and not a state of the system, because it alters the temperature inside the building and not vice versa. Moreover, we suppose that each room is equipped with a heater and assume that the doors of the building are airtight that do not allow heat transfer via convection. The windows are also airtight, however, they can allow the heat transfer through diffusion because they have a low thermal mass as compared to the doors.

There is a duality between heat transfer and electrical phenomenon [Skadron2002], where the temperature difference is analogous to voltage, heat flow to current, thermal resistance to electrical resistance, and thermal mass to electrical capacitance. Therefore, resistor-capacitor (RC) network models are considered to be suitable for the heat conduction. Convection and radiation, on the other hand, can be approximated by a resistor with a nominal empirical resistance value [Mathews1994].

We use the model of [Wang006a], where the building envelope is represented by 3R2C shown in Figure 4.2(c) and the internal mass by 2R2C shown in Figure 4.2(b). The building envelope consists of the walls, ceiling, and floor, and the internal mass consists of the carpet, furniture, and people. The mean air temperature of the room is denoted as $x_{in}(t)$ in the figure. The heater model is shown in Figure 4.2(a), where $q_h(t)$ is a known input and $x_h(t)$ is the temperature at the surface of a heater.

[1] Ministère de l'Écologie, Energy Figures 2019

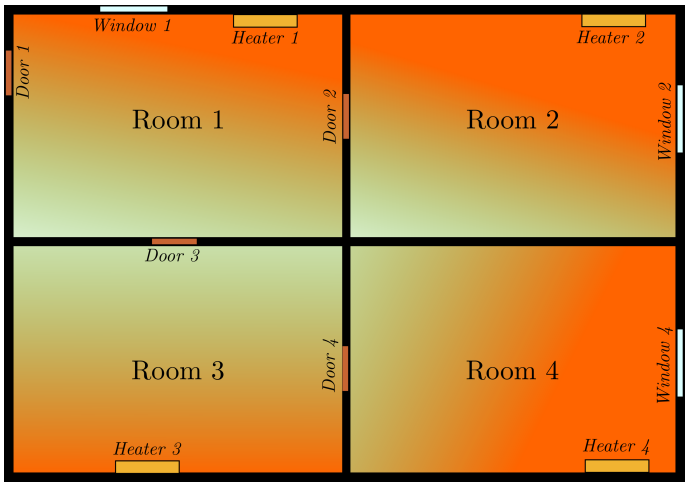
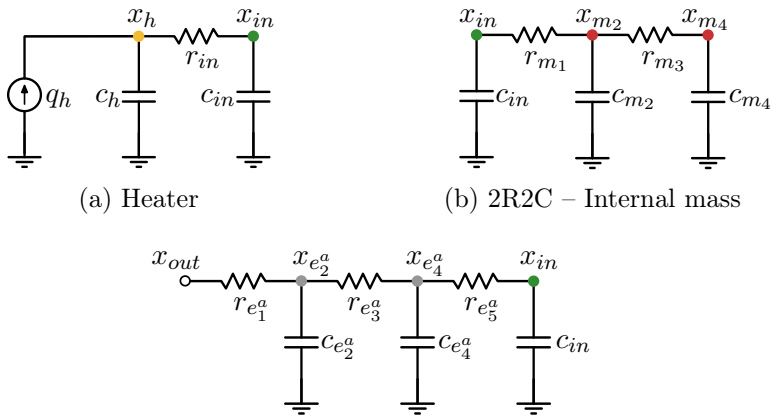


Figure 4.1: A 4-room building setup with heaters.



(a) Heater

(b) 2R2C – Internal mass

(c) 3R2C – Building envelope

Figure 4.2: The elements of the building thermal model.

In Figure 4.2(b), $x_{m_2}(t)$ is the mean surface temperature of the total mass and $x_{m_4}(t)$ is the mean temperature of its core. In Figure 4.2(c), $x_{e_2^a}(t)$ is the mean temperature of the outside surface of envelope-a and $x_{e_4^a}(t)$ is the mean temperature of the inside surface of envelope-a, where a denotes a wall, ceiling, or floor. By employing the Kirchoff's current law, we find that the temperature $x_i(t)$ at node i is governed by

$$c_i \frac{dx_i(t)}{dt} = \sum_{j \in \mathcal{N}_i} \frac{x_j(t) - x_i(t)}{r_{ij}} + \sum_k b_{ik} q_k(t), \tag{4.19}$$

where c_i is the capacitance of node i, $\mathcal{N}_i$ is the set of i's neighboring nodes, r_{ij} is the resistance between i and j, $b_{ik} \in \{0, 1\}$ is a scalar, and $q_k(t)$ is the heater input to i when $b_{ik} = 1$. The outside temperature $x_{out}(t)$ directly influences the room's temperature if it has a window; let r_w be the resistance that the window offers.

The parameter values for the 3R2C and 2R2C models are given in Table 4.1, where the resistance is measured in $\mathrm{m^2KW^{-1}}$ and the capacitance in $\mathrm{MJm^{-2}K^{-1}}$. The temperature is measured in K, which is converted to °C in the figures. We use the parameter values for 3R2C as provided in [Deconinck2016]. The parameter values of 2R2C model are hypothetical because they depend on the type and quantity of internal mass of each

Table 4.1: Parameter values for the building thermal model.

3R2C model	$r_{e_1^a}$	$c_{e_2^a}$	$r_{e_3^a}$	$c_{e_4^a}$	$r_{e_5^a}$
Ceiling	0.3	0.17	4	0.22	0.3
Floor	0.3	0.17	4	0.22	0.3
External walls	0.3	0.17	4	0.22	0.3
Internal walls	0.3	0.22	4	0.22	0.3
2R2C model	r_{m_1}	c_{m_2}	r_{m_3}	c_{m_4}	-
Internal mass	0.16	0.5	3	0.5	-

room; [Wang006b] provides an algorithm to identify these parameters. Finally, the resistance of a window $r_w = 3$ and the resistance from a heater to a room $r_{in} = 0.05$, whereas the capacitance of a room $c_{in} = 0.1$ and of a heater $c_h = 0.5$.

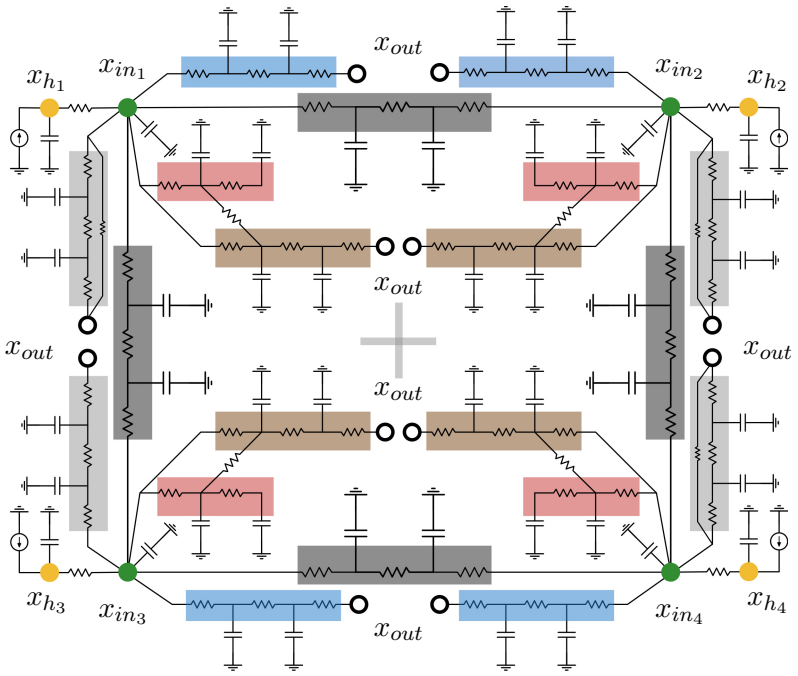


Figure 4.3: RC-network representation of the building setup of Figure 4.1.

The RC-network representation of the building setup is illustrated in Figure 4.3, which has 48 nodes and can be represented by a bidirected graph shown in Figure 4.4. Notice that, in Figure 4.3 and 4.4, there is an edge from the floor to the internal mass. The arrows on some nodes in Figure 4.4 represent the input at those nodes. All the black arrows indicate the influence of the outside temperature $x_{out}(t)$, whereas each yellow arrow indicates a heater input q_{h_p}, for $p = 1, 2, 3, 4$. We assume that the four heater nodes are the *measured* nodes, that is, the temperature evolution on the surface of the heaters is measured by sensors. The remaining nodes in the system are the *unmeasured* nodes.

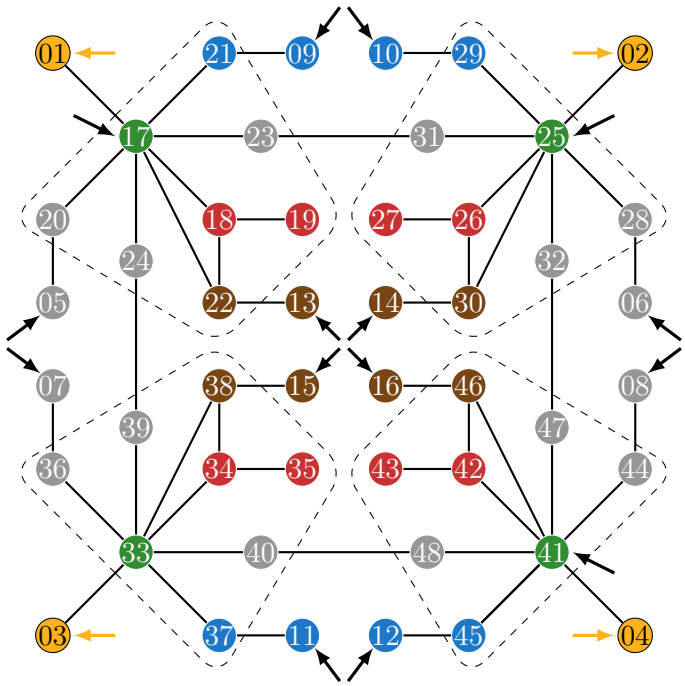


Figure 4.4: Graph representation of the building setup, where the green nodes represent the rooms, the red nodes represent the internal mass, the blue nodes represent the ceiling, the brown nodes represent the floor, the gray nodes represent the walls, and the yellow nodes represent the heaters. The four yellow arrows represent inputs from the four heaters, whereas all the black arrows represent the input $x_{out}(t)$. The clusters of elements corresponding to each room are encircled by a dashed line.

4.4.2 State space representation of the building thermal system

To provide the state-space representation, we index the nodes as follows. The measured nodes are

$$\mathcal{V}_1 = \{1, 2, 3, 4\}$$

which are the heaters' surfaces. The remaining nodes are the unmeasured nodes

$$\mathcal{V}_2 = \mathcal{C}_o \cup \mathcal{C}_{r_1} \cup \mathcal{C}_{r_2} \cup \mathcal{C}_{r_3} \cup \mathcal{C}_{r_4}$$

where the nodes corresponding to the outer building envelope are

$$\mathcal{C}_o = \{5, 6, \ldots, 16\}$$

and the nodes corresponding to the inner elements of the four rooms are

$$\begin{aligned} \mathcal{C}_{r_1} &= \{17, 18, \ldots, 24\} \\ \mathcal{C}_{r_2} &= \{25, 26, \ldots, 32\} \\ \mathcal{C}_{r_3} &= \{33, 34, \ldots, 40\} \\ \mathcal{C}_{r_4} &= \{41, 42, \ldots, 48\} \end{aligned}$$

respectively. Let $i = 1, 2, \ldots, 48$. Then, the temperature of node i at time $t \geq 0$ is denoted as $x_i(t) \in \mathbb{R}$. Let the state vector

$$\mathbf{x}(t) = \begin{bmatrix} x_1(t) & \cdots & x_{48}(t) \end{bmatrix}^\intercal$$

then the state-space representation of the system is

$$\boldsymbol{\Sigma} : \begin{cases} \dot{\mathbf{x}}(t) &= A\mathbf{x}(t) + B\mathbf{u}(t) \\ \mathbf{y}(t) &= C\mathbf{x}(t) \end{cases} \tag{4.20}$$

where

$$\mathbf{u}(t) = \begin{bmatrix} q_{h_1}(t) & q_{h_2}(t) & q_{h_3}(t) & q_{h_4}(t) & x_{out}(t) \end{bmatrix}^\intercal$$

is the input vector and

$$\mathbf{y}(t) = \begin{bmatrix} x_{45}(t) & x_{46}(t) & x_{47}(t) & x_{48}(t) \end{bmatrix}^\intercal$$

is the output vector.

The structure of the system is represented by a bidirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$ shown in Figure 4.4, where $\mathcal{V} = \mathcal{V}_1 \cup \mathcal{V}_2$ is the set of nodes, $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is the set of edges, and

$$\mathcal{W} = \left\{ \frac{1}{c_i r_{ij}} : (i, j) \in \mathcal{E} \right\}$$

is the set of edge weights. The graph is bidirected because the edge weight for $(i, j) \in \mathcal{E}$ is $(c_i r_{ij})^{-1}$, whereas the edge weight for $(j, i) \in \mathcal{E}$ is $(c_j r_{ij})^{-1}$, where c_i is the capacitance of the node i and r_{ij} is the resistance between node i and j; also, $r_{ii} = r_{e_1^a}$, if $i \in \mathcal{V}_o$, and $r_{ii} = r_w$, if $i \in \{13, 21, 37\}$.

The off-diagonal entries of the state matrix A, for $i \neq j$, are given as

$$[A]_{ij} = \begin{cases} (c_i r_{ij})^{-1}, & \text{if } (i, j) \in \mathcal{E} \\ 0, & \text{otherwise} \end{cases}$$

and the diagonal entries are given as

$$[A]_{ii} = \begin{cases} -(r_{ii})^{-1} - \sum_{j \neq i} [A]_{ij}, & \text{if } i \in \mathcal{S}_o \\ -\sum_{j \neq i} [A]_{ij}, & \text{if } i \in \mathcal{V} \setminus \mathcal{S}_o \end{cases}$$

where $\mathcal{S}_o = \mathcal{V}_o \cup \{17, 25, 41\}$ is the set of nodes that are directly influenced by the outside temperature $x_{out}(t)$. The input matrix

$$[B]_{ip} = \begin{cases} (c_i r_{ii})^{-1}, & \text{if } i \in \mathcal{S}_o \text{ and } p = 5 \\ 1, & \text{if } (i,p) \in \{(17,1),(25,2),(33,3),(41,4)\} \\ 0, & \text{otherwise} \end{cases}$$

and the output matrix

$$C = \begin{bmatrix} I_4 & 0_{4\times 44} \end{bmatrix}.$$

4.4.3 Average temperature estimation of building rooms

The average temperature of each room is called the mean operative temperature, which is the mean of the temperature of each element corresponding to the room. In the building setup of Figure 4.1, we have five clusters, where one cluster contains the nodes corresponding to the outer envelope of the building and the four clusters contain the nodes corresponding to each room. The clustering $\mathcal{Q} = \{\mathcal{C}_o, \mathcal{C}_{r_1}, \mathcal{C}_{r_2}, \mathcal{C}_{r_3}, \mathcal{C}_{r_4}\}$, where $\mathcal{C}_o$ is the cluster of nodes representing the elements of outer envelope of the building and $\mathcal{C}_{r_1}, \mathcal{C}_{r_2}, \mathcal{C}_{r_3}, \mathcal{C}_{r_4}$ is the clusters of nodes representing the elements of each room in the building, respectively. The characteristic matrix $Q \in \{0,1\}^{48\times 5}$ of the clustering $\mathcal{Q}$ is given by

$$Q = \mathtt{diag}(\mathbf{1}_{12}, \mathbf{1}_8, \mathbf{1}_8, \mathbf{1}_8, \mathbf{1}_8)$$

where the dimensions of the vectors of ones is due to $|\mathcal{C}_o| = 12$ and $|\mathcal{C}_{r_1}| = \cdots = |\mathcal{C}_{r_4}| = 8$.

The average observer $\mathbf{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$, where $\mathcal{V}_1$ is the set of measured nodes that represent the heaters' surfaces, is given by

$$\mathbf{\Omega}_{\mathcal{V}_1,\mathcal{Q}} := \begin{cases} \dot{\mathbf{w}}(t) &= M\mathbf{w}(t) + K\mathbf{y}(t) + N\mathbf{u}(t) \\ \hat{\mathbf{z}}_{\mathtt{a}}(t) &= \mathbf{w}(t) + L\mathbf{y}(t) \end{cases}$$

where $\hat{\mathbf{z}}_{\mathtt{a}}(t) = \begin{bmatrix} \hat{z}_o(t) & \hat{\mathbf{z}}_r^{\intercal}(t) \end{bmatrix}^{\intercal}$ with $\hat{z}_o(t)$ the estimated average temperature of the outer envelope of the building and $\mathbf{z}_r(t) = \begin{bmatrix} \hat{z}_1(t) & \hat{z}_2(t) & \hat{z}_3(t) & \hat{z}_4(t) \end{bmatrix}^{\intercal}$ the vector of estimated average temperatures of the rooms, and

$$\begin{aligned} M &= (Q^+ A_{22} - L A_{12}) Q \\ N &= Q^+ B_2 - L B_1 \\ K &= Q^+ A_{21} - L A_{22} + ML \\ L &= (Q^+ A_{22} - \rho^* Q^+ A_{22} Q Q^+) A_{12}^{\dagger} \end{aligned}$$

with $\rho^* = 14$ obtained by solving (4.13) by the incremental search algorithm, which is plotted in Figure 4.5. The submatrices $A_{11} \in \mathbb{R}^{4\times 4}$, $A_{12} \in \mathbb{R}_{\geq 0}^{4\times 44}$, $A_{21} \in \mathbb{R}_{\geq 0}^{44\times 4}$, and

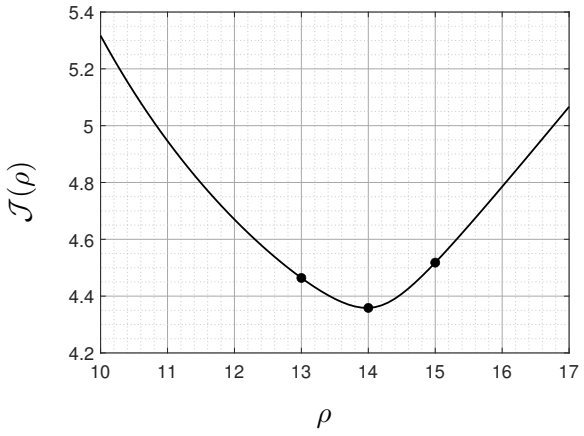


Figure 4.5: The evolution of cost $\mathcal{J}(\rho)$ of (4.13) with respect to ρ.

$A_{22} \in \mathbb{R}^{44\times 44}$ correspond to the partition of state matrix A according to measured and unmeasured nodes as in (2.4).

The initial condition $\mathbf{x}(0) \in \mathbb{R}^{48}$ of the model (4.20), i.e., the vector of initial temperatures of the elements of building, is chosen randomly in the interval $(10, 20)$. The output $\mathbf{y}(t)$ of the system consists of the temperature measurements at the heaters' surfaces. The input $\mathbf{u} = [\ \mathbf{u}_h^\intercal \quad x_{out}\ ]^\intercal$, where $\mathbf{u}_h^\intercal = [\ q_{h_1} \quad q_{h_2} \quad q_{h_3} \quad q_{h_4}\ ]$ is the input of the heaters and x_{out} is the known outside temperature. We suppose $x_{out}(t) = 5\sin(\pi/12\, t - \pi)$.

A simple on/off control policy is used for the heaters by taking a feedback of the estimates of rooms' mean operative temperatures from $\mathbf{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$, which is given by

$$q_{h_j}(\hat{z}_j(t)) = \begin{cases} 50 & \text{if } \hat{z}_j(t) \le 20 \\ 0 & \text{if } \hat{z}_j(t) \ge 22 \\ 50 & \text{if } 20 \le \hat{z}_j(t) < 22 \text{ and } \left.\frac{d\hat{z}_j}{dt}\right|_{\hat{z}_j(t)=20} \ge 0 \\ 0 & \text{if } 20 < \hat{z}_j(t) \le 22 \text{ and } \left.\frac{d\hat{z}_j}{dt}\right|_{\hat{z}_j(t)=22} \le 0 \end{cases}$$

for $j = 1, 2, 3, 4$. That is, the heater j turns on with $q_{h_j}(\hat{z}_j(t)) = 50$ when $\hat{z}_j(t) \le 20$, and it turns off with $q_{h_j}(\hat{z}_j(t)) = 0$ when $\hat{z}_j(t) \ge 22$. Inside the interval $20 < \hat{z}_j(t) < 22$, the control input $q_{h_j}(\hat{z}_j(t))$ retains its value. That is, suppose $\hat{z}_j(t_1) \le 20$, then, for $t > t_1$, the heater turns on with $q_{h_j}(\hat{z}_j(t)) = 50$. When the heater is on, the mean operative temperature of room j starts to rise, and so does its estimate, i.e., $20 < \hat{z}_j(t) < 22$. The heater will stay on until $\hat{z}_j(t) = 22$, where it will be turned off. It will remain off, and the mean operative temperature falls and so does its estimate, until $\hat{z}_j(t)$ touches its lower limit of 20 °C, where the heater will be turned on again.

The plots of average temperatures of the rooms and their estimated trajectories are shown in Figure 4.6. With a simple on/off control policy and an average observer, notice that the average (or mean operative) temperatures of rooms remain inside the thermal comfort range 20-22 °C. This comfort range is nominal but it can be adjusted according to weather, building type, etc. Also, in Figure 4.6, notice that the average temperature of Room-3 reaches inside this range quickly because it doesn't have a window, therefore, it has a smaller influence of the outside temperature.

Let $\mathbf{z}_r(t) = [\ z_1(t) \quad z_2(t) \quad z_3(t) \quad z_4(t)\ ]^\intercal$ and $\hat{\mathbf{z}}_r = [\ \hat{z}_1(t) \quad \hat{z}_2(t) \quad \hat{z}_3(t) \quad \hat{z}_4(t)\ ]^\intercal$.

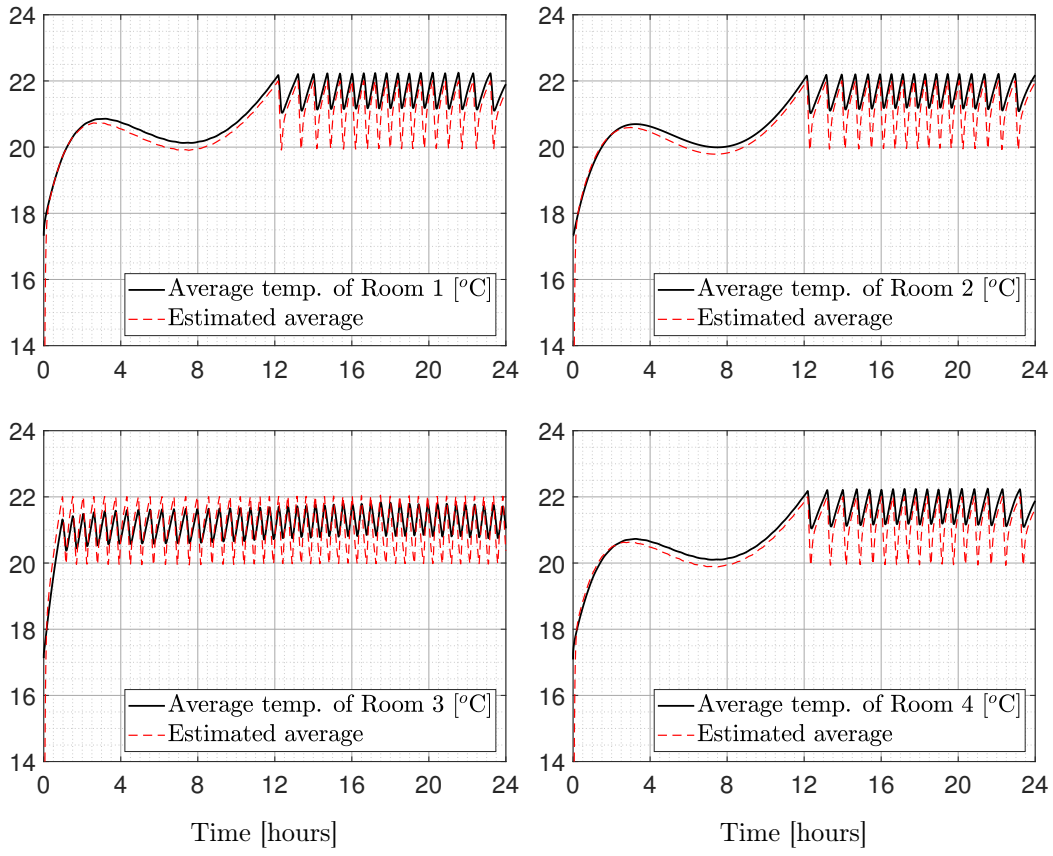


Figure 4.6: Average temperature estimation of the building rooms.

Then, the percentage estimation error $\mathbf{e}_{\%}(t)$ is defined as

$$\mathbf{e}_{\%}(t) = \frac{\|\mathbf{z}_r(t) - \hat{\mathbf{z}}_r(t)\|}{\|\mathbf{z}_r(t)\|}.$$

The performance of $\mathbf{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$ is quite satisfactory as shown in Figure 4.7. For the optimal value $\rho = 14$, the mean percentage error is 2.08%, i.e., around 0.4 °C for the range 20-22 °C, and the maximum percentage error is 4.9%, i.e., around 1 °C for the range 20-22 °C.

In conclusion, the $\mathcal{H}_2$-optimal average observer estimates the mean operative temperatures of rooms in a building with a very small error. The dimension of the proposed observer equals the number of rooms (or clusters) plus one, where the extra 'one' is due to the cluster of nodes representing outer envelope of the building. The problem of error minimization is simplified to a great degree by formulating it with respect to a single parameter ρ, whose optimal value can be found by the gradient descent or the incremental search algorithm. We employed a simple on/off control policy based on the average observer to regulate the mean operative temperatures of rooms. Although it is quite simple, the on/off policy for regulation saves around 25.32% of the energy, which means that the heaters on average remain off 25.32% of the day.

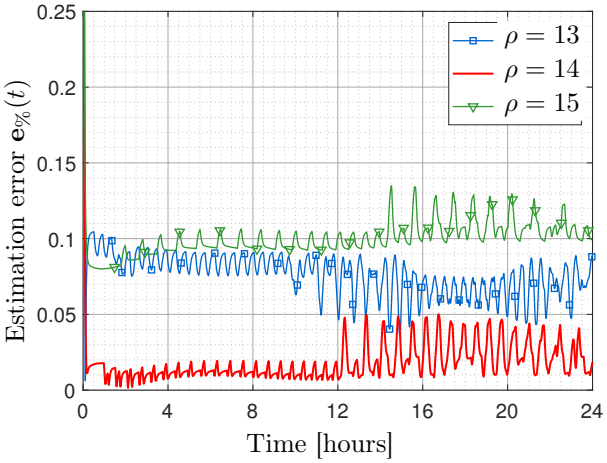


Figure 4.7: Percentage estimation error for optimal $\rho = \rho* = 14$ and non-optimal values of ρ.

4.5 Concluding Remarks

In this chapter, we provided a methodology to minimize the average estimation error when the clustered network system does not meet the design criteria of the average observer. The methodology comprises the minimization of the effect of the average deviation vector on the dynamics of average estimation error and stabilization of the average observer. First, a structure of the design matrix of the average observer is obtained that minimizes the effect of the average deviation vector acting as a structured unknown input. Then, to ensure the stability of the average observer, we perturbed this structure by a single parameter ρ and find its value that achieves stability.

Tools from matrix analysis and algebraic graph theory are employed to establish sufficient conditions for the stabilizability of the average observer. The main assumption in the literature, for instance, [Ishizaki2015], is to have strong connectivity of graph for the stability of the compression of stable matrices. However, we showed that this assumption can be relaxed to weak connectivity of induced subgraph $\mathcal{G}_\nu$ in a network system to ensure that the matrix compression $Q^\intercal A_{22} Q$ is stable. Then, using Lyapunov's theorem and S-stability result, we proved that the average observer can be stabilized if the clustering is such that $A_{12}Q$ is full-column rank. The gain parameter that stabilizes the average observer is the perturbation parameter ρ in the design matrix.

The $\mathcal{H}_2$-optimal average estimation problem is formulated, which is a convex optimization problem that can be solved by a gradient descent algorithm. We provided an analytic expression to compute the gradient of a cost function that is employed in the gradient descent algorithm. Since a single parameter, ρ, is optimized in this problem, we provided an incremental search algorithm, which is much simpler than the gradient descent algorithm. Finally, we showed the efficacy of optimal average estimation methodology for the application of thermal monitoring in buildings.

5

Clustering Algorithms for Large-Scale Network Systems

In this chapter, we study clustering techniques for optimal average estimation, open-loop average estimation, and state variance estimation. Section 5.1 presents conditions for generic stabilizability of average observer for optimal average estimation and considers them as constraints in the clustering algorithm for optimal average estimation. In section 5.2, we introduce the notion of average lumpability and show its relation to average detectability. Then, we propose a clustering algorithm for open-loop average estimation that achieves 'almost' average lumpability by minimizing the distance from the ideal notion. In section 5.3, we show that the state variance can be approximated by obtaining a clustered network system through a K-means type clustering algorithm. The optimal average estimation clustering methodology is illustrated by an application example of an SIS epidemic over a network in section 5.4, whereas the clustering methodologies for open-loop average estimation and state variance estimation are illustrated by numerical examples in their respective sections.

Contents

When the clusters are not pre-specified in a large-scale network system, as was the case in previous chapters, then clustering techniques can be employed to render the network system close to average reconstructability or average detectability. The distance from average reconstructability is minimized by identifying and aggregating clusters that yield a minimum average estimation error. On the other hand, the distance from average detectability is minimized by identifying and aggregating clusters that yield a minimum open-loop average approximation error. In this clustering methodology, constraints pertinent to physical network systems, such as the connectivity of clusters, are also added, and we define average lumpability and show its relation to average detectability and the minimization of open-loop average approximation error.

Finally, another aggregated state profile of network systems, i.e., state variance, which is a nonlinear functional of the state vector and measures the distance of state trajectories from their average mean, is estimated in an approximated sense by employing a K-means type clustering algorithm. The clusters obtained through this clustering algorithm contain nodes whose state trajectories are close to each other, which facilitates the approximation of state variance through the average states of clusters.

5.1 Clustering for Optimal Average Estimation

In this section, we provide a clustering technique to render a network system close to average reconstructability by minimizing the average estimation error. First, we provide a clustering constraint that ensures the stabilizability of the average observer in a generic sense, by which we mean that the average observer can be stabilized through the gain parameter γ with probability one. This notion is defined by changing the rank condition of Theorem 4.9 to a generic rank condition, which is shown to be equivalent to having each cluster contain at least one unmeasured node that is a neighbor of a measured node. Then, we propose a clustering algorithm to achieve a minimum average estimation error.

5.1.1 Preliminaries on the generic rank of matrices

The rank of a matrix $X \in \mathbb{R}^{m\times n}$ denoted as $\texttt{rank}(X)$ is equal to the number of linearly independent rows, or, equivalently, linearly independent columns, of X. In other words, $\texttt{rank}(X)$ is defined as the dimension of the column space (equivalently, the dimension of the row space) of X. However, $\texttt{rank}(X)$ can be computed when X is known. If only the structure (or non-zero pattern) of X is known, i.e., the entries of X that are fixed to be zero and the remaining entries that are arbitrary non-zero real numbers, then a more suitable quantity to consider is the generic rank [Lin1974, Murota1987, Murota2010].

Definition 5.1. The generic rank of $X \in \mathbb{R}^{m\times n}$ denoted as $\texttt{grank}(X)$ is defined as the maximum rank of X among all choices of non-zero entries in the non-zero pattern of X.

In general, we have $\texttt{rank}(X) \leq \texttt{grank}(X)$. However, for a structured matrix $X \in \mathbb{R}^{m\times n}$, we have $\texttt{grank}(X) = \texttt{rank}(X)$ almost always (i.e., with probability 1) except for the

entries of X in some proper algebraic variety, which is of Lebesgue measure zero [Dion2003].

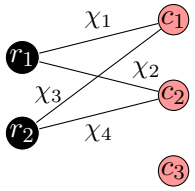
Notice that the structure of X can be represented as a bipartite graph $\mathcal{G}_X = (\mathcal{V}_r, \mathcal{V}_c, \mathcal{E}_X)$, where $\mathcal{V}_r = \{r_1, \ldots, r_m\}$ is the index set of the rows of X, $\mathcal{V}_c = \{c_1, \ldots, c_n\}$ is the index set of the columns of X, and $\mathcal{E}_X \subseteq \mathcal{V}_r \times \mathcal{V}_c$ is the set of edges defined as $(r_i, c_j) \in \mathcal{E}_X$ if $[X]_{r_i c_j} \neq 0$. A *matching* in a bipartite graph $\mathcal{G}_X$ is the set of edges such that no two edges have a vertex in common, whereas a *maximum matching* is a matching with the maximum possible number of edges [Godsil2001].

Lemma 5.1 ([Liu2011]). The generic rank of $X \in \mathbb{R}^{m \times n}$ is equal to the size of maximum matching in the bipartite graph $\mathcal{G}_X$.

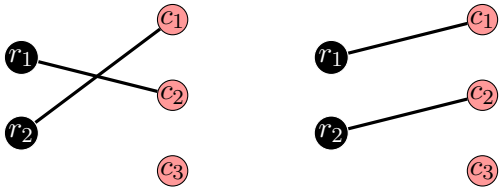
For example, for some non-zero $\chi_1, \chi_2, \chi_3, \chi_4 \in \mathbb{R} \setminus \{0\}$, let

$$X = \begin{bmatrix} \chi_1 & \chi_2 & 0 \\ \chi_3 & \chi_4 & 0 \end{bmatrix}$$

that is represented as a bipartite graph $\mathcal{G}_X$



which has two maximum matchings



where the size of each maximum matching is two. Therefore, in this case, $\texttt{grank}(X) = 2$, which means that the rank of X is 2 for all values of $\chi_1, \chi_2, \chi_3, \chi_4$ except for the case $\chi_1\chi_4 = \chi_2\chi_3$, which is of Lebesgue measure zero.

5.1.2 Clustering constraint for the generic stabilizability of average observer

Recall the network system $\mathbf{\Sigma}$ whose structure is represented by a digraph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \mathcal{V}_1 \cup \mathcal{V}_2$ is the set of nodes with $\mathcal{V}_1$ the set of measured nodes and $\mathcal{V}_2$ the set of unmeasured nodes, and $\mathcal{E}$ is the set of edges. As given in (2.4), the system matrices of $\mathbf{\Sigma}$ are partitioned into block submatrices according to the partition of nodes into measured and unmeasured.

Definition 5.2. A subset of unmeasured nodes $\mathcal{N}_{\mathcal{V}_1 \leftarrow \mathcal{V}_2} \subseteq \mathcal{V}_2$ is said to be the *neighbor set* of measured nodes $\mathcal{V}_1$ with respect to the set of unmeasured nodes $\mathcal{V}_2$ if, for every $j \in \mathcal{N}_{\mathcal{V}_1 \leftarrow \mathcal{V}_2}$, there exists $i \in \mathcal{V}_1$ such that $(i, j) \in \mathcal{E}$.

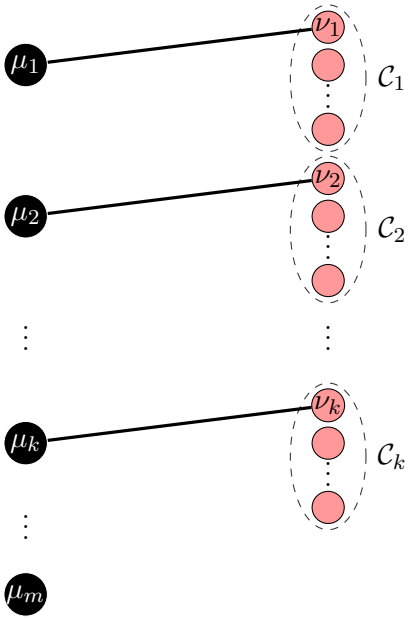
Recall the set of characteristic matrices

$$\mathfrak{C}_{n,k} = \{X \in \{0,1\}^{n \times k} : X\mathbf{1}_k = \mathbf{1}_n\}$$

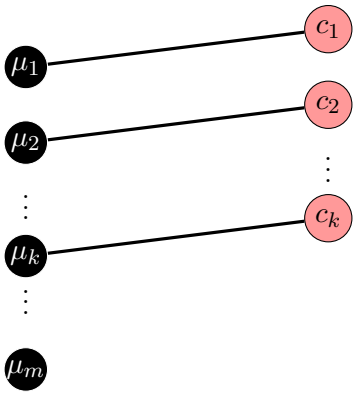
of all clusterings $\mathcal{Q}$ of n nodes with size k.

Theorem 5.2. Let $Q \in \mathfrak{C}_{n,k} \subset \{0,1\}^{n\times k}$ be the characteristic matrix of some clustering $\mathcal{Q}$ of n unmeasured nodes $\mathcal{V}_2$ and let Assumption 2.1 hold. Further, assume $k \leq m$, where m is the number of measured nodes. Then, $\texttt{grank}(A_{12}Q) = k$ *if and only if* the clustering $\mathcal{Q} = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$ is such that, for every $\alpha \in \{1, \ldots, k\}$, $\mathcal{C}_\alpha \cap \mathcal{N}_{\mathcal{V}_1 \leftarrow \mathcal{V}_2} \neq \emptyset$.

Proof of sufficiency. Assume that the clustering $\mathcal{Q} = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$ is such that, for every $\alpha \in \{1, \ldots, k\}$, it holds $\mathcal{C}_\alpha \cap \mathcal{N}_{\mathcal{V}_1 \leftarrow \mathcal{V}_2} \neq \emptyset$. By Assumption 2.1, we have $\texttt{rank}(A_{12}) = m$, where $m \leq n$ since A_{12} has the dimension $m \times n$. Since $\texttt{rank}(A_{12}) \leq \texttt{grank}(A_{12}) \leq m$, therefore $\texttt{grank}(A_{12}) = m$, which implies that a maximum matching of the bipartite graph $\mathcal{G}_{A_{12}}$ is of size m. That is, for all the m measured nodes there are distinct m neighbors in $\mathcal{N}_{\mathcal{V}_1 \leftarrow \mathcal{V}_2}$, which implies $|\mathcal{N}_{\mathcal{V}_1 \leftarrow \mathcal{V}_2}| \geq m$. Since $k \leq m$ and $|\mathcal{N}_{\mathcal{V}_1 \leftarrow \mathcal{V}_2}| \geq m$, we have $|\mathcal{N}_{\mathcal{V}_1 \leftarrow \mathcal{V}_2}| \geq k$. Without loss of generality, let $\nu_1, \ldots, \nu_k \in \mathcal{N}_{\mathcal{V}_1 \leftarrow \mathcal{V}_2}$ to be the k unmeasured nodes that are in clusters $\mathcal{C}_1, \ldots, \mathcal{C}_k$, respectively. Then, a matching of size k of the bipartite graph $\mathcal{G}_{A_{12}}$ is



where $\mu_1, \ldots, \mu_m$ are the m measured nodes representing the rows of A_{12} and the nodes on the right side are the n unmeasured nodes representing the columns of A_{12}. The unmeasured nodes are partitioned into clusters $\mathcal{C}_1, \ldots, \mathcal{C}_k$. The bipartite graph $\mathcal{G}_{A_{12}Q}$ is obtained by aggregating the clusters in $\mathcal{G}_{A_{12}}$. Then, from the matching of size k illustrated above for $\mathcal{G}_{A_{12}}$, we obtain a maximum matching of size k for $\mathcal{G}_{A_{12}Q}$



where the clusters $\mathcal{C}_1, \ldots, \mathcal{C}_k$ are represented as super nodes $c_1, \ldots, c_k$, respectively. Thus, by Lemma 5.1, we have $\texttt{grank}(A_{12}Q) = k$.

Proof of necessity. To prove necessity, assume $\texttt{grank}(A_{12}Q) = k$ and there exists some $\alpha \in \{1, \ldots, k\}$ such that $\mathcal{C}_\alpha \cap \mathcal{N}_{\mathcal{V}_1 \leftarrow \mathcal{V}_2} = \emptyset$. Let $A_{12} = \begin{bmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_n \end{bmatrix}$, where $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^m$ are the columns of $A_{12} \in \mathbb{R}^{m \times n}_{\geq 0}$. Then, we can write

$$A_{12}Q = \begin{bmatrix} \mathbf{p}_1 & \cdots & \mathbf{p}_k \end{bmatrix}$$

where $\mathbf{p}_1, \ldots, \mathbf{p}_k \in \mathbb{R}^m$ are the columns of $A_{12}Q$ with

$$\mathbf{p}_\alpha = \sum_{j \in \mathcal{C}_\alpha} \mathbf{a}_j$$

for $\alpha = 1, \ldots, k$. Since $\texttt{grank}(A_{12}Q) = k$, we have that, for every $\alpha \in \{1, \ldots, k\}$, $\mathbf{p}_\alpha \neq \mathbf{0}_m$. On the other hand, since there exists $\alpha \in \{1, \ldots, k\}$ such that $\mathcal{C}_\alpha \cap \mathcal{N}_{\mathcal{V}_1 \leftarrow \mathcal{V}_2} = \emptyset$, which means the cluster $\mathcal{C}_\alpha$ does not contain any node from the neighbor set $\mathcal{N}_{\mathcal{V}_1 \leftarrow \mathcal{V}_2}$, i.e., $\mathbf{a}_j = 0$ for all $j \in \mathcal{C}_\alpha$, therefore $\mathbf{p}_\alpha = 0$. This implies that $\texttt{grank}(A_{12}Q) < k$, which is a contradiction.

□

If the clustering $\mathcal{Q}$ is not prespecified, then the state matrix $M := M_{\rho,Q}$ of the average observer $\mathbf{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$ depends on the characteristic matrix $Q \in \mathfrak{C}_{n,k}$ of a clustering $\mathcal{Q}$ of unmeasured nodes in addition to the gain parameter $\rho > 0$. From (4.8) and (4.6), we have

$$M_{\rho,Q} = R_{\rho,Q}Q \tag{5.1}$$

where

$$R_{\rho,Q} = Q^+ A_{22}(I_n - A_{12}^\dagger A_{12}) + \rho Q^+ A_{22} Q Q^+ A_{12}^\dagger A_{12}. \tag{5.2}$$

Definition 5.3 (Generic stabilizability of average observer)**.** For a network system $\mathbf{\Sigma}$, the average observer $\mathbf{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$ is said to be generically stabilizable if, for every $Q \in \mathfrak{C}_{n,k}$, it holds that $\texttt{grank}(A_{12}Q) = k$.

In the previous chapter, we stated a sufficient condition for the stabilizability of average observer as the full-column rank of $A_{12}Q$ in Theorem 4.9. The stabilizability of average observer was defined as the existence of some $\phi \in \mathbb{R}$ such that $M_{\rho,Q}$ is Hurwitz for all $\rho > \phi$. However, the generic stabilizability is defined through the generic rank of $A_{12}Q$, which means that there exists $\phi_Q \in \mathbb{R}$ *almost always* such that $M_{\rho,Q}$ is Hurwitz for every $\rho > \phi_Q$. The term 'almost always' indicates that, for any submatrix $A_{12} \in \mathbb{R}^{m \times n}_{\geq 0}$ belonging to a network system $\mathbf{\Sigma}$, the rank $A_{12}Q$ is equal to k with probability one if the condition of Theorem 5.2 is satisfied.

Recall the induced subgraph $\mathcal{G}_\nu = (\mathcal{V}_2, \mathcal{E}_\nu)$ formed by the unmeasured nodes of the network system $\mathbf{\Sigma}$ and that, for every $i \in \mathcal{V}_2$,

$$s_i = \sum_{j \in \mathcal{N}_{i \leftarrow \mathcal{V}_2}} a_{ij} + \sum_{h \in \mathcal{N}_{i \rightarrow \mathcal{V}_2}} a_{hi}$$

is the sum of weighted in-degree and out-degree of i.

Theorem 5.3. Consider a network system $\mathbf{\Sigma}$ in (2.3) with $\mathcal{V}_1$ the set of m measured nodes and $\mathcal{V}_2$ the set of n unmeasured nodes. Then, for any clustering $\mathcal{Q}$ of unmeasured nodes with $k \leq m$ clusters, the average observer $\mathbf{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$ is generically stabilizable if

(i) $\mathcal{G}_\nu = (\mathcal{V}_2, \mathcal{E}_\nu)$ is weakly connected,

(ii) $\forall i \in \mathcal{V}_2$, $s_i \leq 2|a_{ii}|$ and $\exists j \in \mathcal{V}_2$ such that $s_j < 2|a_{jj}|$, and

(iii) $\forall \alpha \in \{1, \ldots, k\}$, $\mathcal{C}_\alpha \cap \mathcal{N}_{\mathcal{V}_1 \leftarrow \mathcal{V}_2} \neq \emptyset$.

Proof. First, assume (i) and (ii), then, for every $Q \in \mathfrak{C}_{n,k}$, $Q^+ A_{22} Q$ is Hurwitz by Theorem 4.7. Second, assume (iii), then the average observer is generically stabilizable by Theorem 4.9 and 5.2.

□

The conditions Theorem 5.3(i) and (ii) correspond to the network system $\mathbf{\Sigma}$, whereas Theorem 5.3(iii) corresponds to the clustering $\mathcal{Q}$. In other words, the first two conditions need to be satisfied by the network system and the third condition needs to be satisfied by the clustering algorithm.

5.1.3 Clustering for $\mathcal{H}_2$-optimal average estimation

We formulate a clustering problem for $\mathcal{H}_2$-optimal average estimation in this subsection. Since they correspond to the network system and not a clustering algorithm, we assume that the conditions of Theorem 5.3(i) and (ii) hold. On the other hand, we consider Theorem 5.3(iii) as a constraint in the clustering algorithm in order to ensure generic stabilizability of average observer.

Problem definition

Consider the dynamics of the average estimation error from (3.7)

$$\dot{\boldsymbol{\zeta}}(t) = M_{\rho,Q} \boldsymbol{\zeta}(t) + R_{\rho,Q} \boldsymbol{\sigma}(t)$$

where $\boldsymbol{\sigma}(t)$ is the average deviation vector defined in (2.11) and $Q \in \mathfrak{C}_{n,k}$ along with $\rho > 0$ is a decision variable. Again, the transfer function from $\boldsymbol{\sigma}$ to $\boldsymbol{\zeta}$ is given by

$$\mathbf{T}_{\rho,Q}(s) = (sI_k - M_{\rho,Q})^{-1} R_{\rho,Q}$$

with its $\mathcal{H}_2$-norm defined as

$$\|\mathbf{T}_{\rho,Q}(s)\|_{\mathcal{H}_2}^2 = \mathtt{trace}(W_{\rho,Q})$$

where

$$W_{\rho,Q} := \int_0^\infty \exp(M_{\rho,Q} t) R_{\rho,Q} \left(\exp(M_{\rho,Q} t) R_{\rho,Q}\right)^\intercal dt \tag{5.3}$$

is the controllability gramian of the pair $(M_{\rho,Q}, R_{\rho,Q})$.

Assumption 5.1. The conditions of Theorem 5.3(i) and (ii) are satisfied. That is, the induced subgraph $\mathcal{G}_\nu = (\mathcal{V}_2, \mathcal{E}_\nu)$ is weakly connected and, $\forall i \in \mathcal{V}_2$, $s_i \leq 2|a_{ii}|$ with at least one $j \in \mathcal{V}_2$ for which $s_j < 2|a_{jj}|$.

Under Assumption 5.1, the clustering problem for $\mathcal{H}_2$-optimal average estimation is formulated as

$$\begin{aligned} \min_{\rho>0, Q \in \mathfrak{C}_{n,k}} \quad & \mathcal{J}(\rho, Q) := \mathtt{trace}(W_{\rho,Q}) \\ \text{subject to} \quad & \begin{cases} M_{\rho,Q} \text{ is Hurwitz} \\ \mathcal{C}_\alpha \cap \mathcal{N}_{\mathcal{V}_1 \leftarrow \mathcal{V}_2} \neq \emptyset, \ \forall \alpha \in \{1, \ldots, k\}. \end{cases} \end{aligned} \tag{5.4}$$

The first constraint in (5.4) is for the stability and the second constraint is for the stabilizability of average observer. Notice that the clustering problems are non-convex, mixed integer optimization problems [Burer2012]. To find the global optimum is NP-hard, therefore, only suboptimal solutions are feasible that converge to a local minimum of the problem (5.4).

Clustering algorithm with optimal average observer gain

Before presenting the clustering algorithm, we need to initialize the k clusters such that the second constraint of (5.4) is satisfied. Notice that if Assumption 5.1 and the second constraint is satisfied, then there exists $\rho \in \mathbb{R}$ such that $M_{\rho,Q}$ is Hurwitz for any $Q \in \mathfrak{C}_{n,k}$ by Theorem 5.3, where such ρ can be obtained by Algorithm 1.

Let $\mathcal{S}_1, \ldots, \mathcal{S}_k$ be k clusters of a subset of $\mathcal{V}_2$ and define

$$\mathcal{N}_{\mathcal{S}_\alpha \leftrightarrow \mathcal{V}_2} = \{j \in \mathcal{V}_2 : (i,j) \in \mathcal{E}_\nu \text{ or } (j,i) \in \mathcal{E}_\nu, \text{ for } i \in \mathcal{S}_\alpha\}$$

to be the set of in-neighbors and out-neighbors of $\mathcal{S}_\alpha$. Then, the selection of initial k clusters is obtained by Algorithm 3.

Algorithm 3 Initialization of k clusters

Input: $\mathcal{G}_\nu = (\mathcal{V}_2, \mathcal{E}_\nu)$ and $\mathcal{N}_{\mathcal{V}_1 \leftarrow \mathcal{V}_2}$
Output: $\mathcal{Q}_0 = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$

Move each $j \in \mathcal{N}_{\mathcal{V}_1 \leftarrow \mathcal{V}_2}$ to either $\mathcal{S}_1, \ldots, \mathcal{S}_k$ such that, $\forall \alpha \in \{1, \ldots, k\}$, $\mathcal{S}_\alpha \neq \emptyset$
repeat
  Assign $\mathcal{S}_1 \leftarrow \mathcal{S}_1 \cup (\mathcal{N}_{\mathcal{S}_1 \leftrightarrow \mathcal{V}_2} \setminus \mathcal{N}_{\mathcal{V}_1 \leftarrow \mathcal{V}_2})$
  for $\alpha = 2, \ldots, k$ **do**
    Assign $\mathcal{S}_\alpha \leftarrow \mathcal{S}_\alpha \cup (\mathcal{N}_{\mathcal{S}_1 \leftrightarrow \mathcal{V}_2} \setminus \mathcal{N}_{\mathcal{V}_1 \leftarrow \mathcal{V}_2} \setminus \mathcal{S}_{\alpha-1})$
  end for
until $\mathcal{S}_1 \cup \cdots \cup \mathcal{S}_k = \mathcal{V}_2$
Assign $\mathcal{C}_1 \leftarrow \mathcal{S}_1, \ldots, \mathcal{C}_k \leftarrow \mathcal{S}_k$
return $\mathcal{Q}_0 = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$

Algorithm 4 Suboptimal clustering $\mathcal{Q}^*$ for fixed ρ

Input: Matrices needed to compute $\mathcal{J}(\rho, Q)$, initial clustering $\mathcal{Q}_0$, initial gain ρ_0, and tolerance $\delta > 0$ (e.g., 10^{-6})
Output: Suboptimal clustering $\mathcal{Q} = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$

Compute $\psi_0 = \mathcal{J}(\rho_0, \mathcal{Q}_0)$ and assign $\mathcal{Q}_1 \leftarrow \mathcal{Q}_0$
repeat
  Assign $\psi_1 \leftarrow \psi_0$
  for $i \in \mathcal{V}_2 \setminus \mathcal{N}_{\mathcal{V}_1 \leftarrow \mathcal{V}_2}$ **do**
    Assign $\mathcal{Q}_2 \leftarrow \mathcal{Q}_1$
    Let β be such that $i \in \mathcal{C}_\beta$
    for $\alpha = 1, \ldots, k$ and $\alpha \neq \beta$ **do**
      Move i to $\mathcal{C}_\alpha$ and update $\mathcal{Q}_2$ accordingly
      Compute $\psi_2 = \mathcal{J}(\rho, Q_2)$
      if $\psi_2 < \psi_0$ **then**
        Assign $\psi_0 \leftarrow \psi_2$ and $\mathcal{Q}_1 \leftarrow \mathcal{Q}_2$
      else
        Move i back to $\mathcal{C}_\beta$ and $\mathcal{Q}_2 \leftarrow \mathcal{Q}_1$
      end if
    end for
  end for
  Assign $\mathcal{Q}^* \leftarrow \mathcal{Q}_1$
until $\psi_1 - \psi_0 < \delta$, i.e., specified tolerance to convergence
return $\mathcal{Q} = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$.

The Algorithm 3 starts by first clustering the neighbor set $\mathcal{N}_{\mathcal{V}_1 \leftarrow \mathcal{V}_2}$ into k nonempty clusters $\mathcal{S}_1, \ldots, \mathcal{S}_k$. This is guaranteed by the assumption of Theorem 5.3 that $k \leq m$ and

by Assumption 2.1 that $\mathtt{rank}(A_{12}) = m$. That is, since the neighbor set $\mathcal{N}_{\mathcal{V}_1 \leftarrow \mathcal{V}_2}$ contains the nodes corresponding to the non-zero columns of A_{12}, we have $|\mathcal{N}_{\mathcal{V}_1 \leftarrow \mathcal{V}_2}| \geq m \geq k$. After clustering $\mathcal{N}_{\mathcal{V}_1 \leftarrow \mathcal{V}_2}$, the second step is to include the remaining nodes $\mathcal{V}_2 \setminus \mathcal{N}_{\mathcal{V}_1 \leftarrow \mathcal{V}_2}$ into the clusters, where a while loop iteratively traverses the graph $\mathcal{G}_\nu$ in a breadth-first search manner to include the immediate neighbors of each cluster into $\mathcal{S}_1, \ldots, \mathcal{S}_k$, respectively. This is done repeatedly until all the unmeasured nodes are clustered, which is guaranteed by the assumption that $\mathcal{G}_\nu$ is weakly connected.

After initialization of k clusters, we obtain the initial optimal gain parameter ρ_0 for the initial clustering $\mathcal{Q}_0$ from Algorithm 1 and 2. Then, by fixing $\rho = \rho_0$, we first obtain a suboptimal clustering $\mathcal{Q}$ from Algorithm 4. Then, for the suboptimal clustering $\mathcal{Q}$, we obtain an optimal ρ from Algorithm 1 and 2. This process is repeated until convergence to a specified tolerance, which is summarized in Algorithm 5.

Finally, note that the clustering algorithm for optimal average estimation is implemented in section 5.4 for an example of SIS epidemics over networks.

Algorithm 5 Suboptimal clustering and optimal average observer

Input: All the inputs of Algorithm 1, 2, 3, and 4; and tolerance $\epsilon > 0$
Output: Suboptimal clustering $\mathcal{Q}^* = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$ and optimal gain ρ^*

Find an initial clustering $\mathcal{Q}_0$ from Algorithm 3
For $\mathcal{Q}_0$, obtain ρ_0 from Algorithm 1 and 2
Assign $\mathcal{Q} \leftarrow \mathcal{Q}_0$ and $\rho \leftarrow \rho_0$
Compute $\psi_1 = \mathcal{J}(\rho, Q)$
repeat
    Assign $\psi_2 \leftarrow \psi_1$
    Find suboptimal clustering $\mathcal{Q}^*$ from Algorithm 4 with $\mathcal{Q}_0 = \mathcal{Q}$ and $\rho_0 = \rho$
    Find optimal gain ρ^* from Algorithm 1 and 2 with characteristic matrix $Q = Q^*$
    Assign $\mathcal{Q} \leftarrow \mathcal{Q}^*$ and $\rho \leftarrow \rho^*$
    Compute $\psi_1 = \mathcal{J}(\rho, \mathcal{Q})$
until $\psi_2 - \psi_1 < \epsilon$
return $\mathcal{Q}^* = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$ and ρ^*

5.2 Clustering for Open-Loop Average Estimation

In this section, we provide a clustering technique to render a network system close to average detectability by minimizing the open-loop average estimation error. We first define a notion of average lumpability, provide its necessary and sufficient condition, and show its relation to average detectability. We show that when $Q^+ A_{22} Q$ is Hurwitz, then average lumpability is equivalent to average detectability. Then, under a constraint of intra-cluster connectivity, we propose a clustering algorithm that minimizes the distance from average lumpability to obtain a minimum open-loop average estimation error and show its efficacy through a simulation example.

5.2.1 Average lumpability and its relation to average detectability

From the dynamics of the projected network system $\mathring{\mathbf{\Sigma}}_{\mathcal{V}_1, \mathcal{Q}}$ in equation (2.9), consider the dynamics of the average state vector

$$\dot{\mathbf{z}}_{\mathtt{a}}(t) = Q^+ A_{22} Q \mathbf{z}_{\mathtt{a}}(t) + Q^+ A_{22} \boldsymbol{\sigma}(t) + Q^+ A_{21} \mathbf{y}(t) + Q^+ B_2 \mathbf{u}(t). \tag{5.5}$$

Here, the average deviation vector acts as a structured unknown input. By ignoring $\boldsymbol{\sigma}(t)$, we obtain an approximated average state vector $\hat{\mathbf{z}}_{\mathtt{a}}(t)$ that satisfies

$$\dot{\hat{\mathbf{z}}}_{\mathtt{a}}(t) = Q^{+}A_{22}Q\hat{\mathbf{z}}_{\mathtt{a}}(t) + Q^{+}A_{21}\mathbf{y}(t) + Q^{+}B_{2}\mathbf{u}(t). \tag{5.6}$$

Definition 5.4. A clustered network system $\boldsymbol{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is said to be *average lumpable* if the following implication holds:

$$\hat{\mathbf{z}}_{\mathtt{a}}(0) = \mathbf{z}_{\mathtt{a}}(0) \quad \Rightarrow \quad \hat{\mathbf{z}}_{\mathtt{a}}(t) = \mathbf{z}_{\mathtt{a}}(t)\ \forall t \in \mathbb{R}_{>0}.$$

The notion of average lumpability states that the inter-cluster and intra-cluster topologies are such that the effect of average deviation vector $\boldsymbol{\sigma}(t)$ is canceled from the dynamics of average state vector $\mathbf{z}_{\mathtt{a}}(t)$.

Theorem 5.4. Consider a clustered network system $\boldsymbol{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ with measured nodes $\mathcal{V}_1$ and a clustering $\mathcal{Q}$ of the unmeasured nodes $\mathcal{V}_2$. Then, the following statements are equivalent:

(i) $\boldsymbol{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average lumpable.

(ii) $\ker(Q^{+}A_{22}) \supseteq \ker(Q^{+})$.

(iii) There exists $V \in \mathbb{R}^{k\times k}$ such that $VQ^{+} = Q^{+}A_{22}$.

Proof. (i) $\Leftrightarrow$ (ii). If $\boldsymbol{\Sigma}$ is average lumpable, then

$$\begin{aligned} \mathbf{z}_{\mathtt{a}}(t) - \hat{\mathbf{z}}_{\mathtt{a}}(t) &= \int_0^t \exp(Q^{+}A_{22}Q\tau)Q^{+}A_{22}\boldsymbol{\sigma}(t-\tau)d\tau \\ &= \mathbf{0}_k \end{aligned} \tag{5.7}$$

where $\hat{\mathbf{z}}_{\mathtt{a}}(0) = \mathbf{z}_{\mathtt{a}}(0)$, which implies that $Q^{+}A_{22}\boldsymbol{\sigma} \equiv \mathbf{0}_k$. Note that $\boldsymbol{\sigma}(t) = I_n - QQ^{+}\mathbf{x}_2(t)$ and the columns of $I_n - QQ^{+}$ form a complete basis of $\ker(Q^{+})$ because $Q^{+}(I_n - QQ^{+}) = 0_{k\times n}$ and $\mathtt{rank}(I_n - QQ^{+}) = \mathtt{nullity}(Q^{+}) = n - k$. Therefore, $\boldsymbol{\sigma}(t) \in \ker(Q^{+})$ and if there exists a matrix X such that $X\boldsymbol{\sigma}(t) = 0$ for all $t \in \mathbb{R}_{\geq 0}$, then $\boldsymbol{\sigma}(t) \in \ker(X)$ and we have an inclusion $\ker(X) \supseteq \ker(Q^{+})$. Hence, $Q^{+}A_{22}\boldsymbol{\sigma}(t) = \mathbf{0}_k$ for all $t \in \mathbb{R}_{\geq 0}$ implies $\ker(Q^{+}A_{22}) \supseteq \ker(Q^{+})$. In the other direction, if $\ker(Q^{+}A_{22}) \supseteq \ker(Q^{+})$, then $Q^{+}A_{22}\boldsymbol{\sigma}(t) = \mathbf{0}_k$ for all $t \in \mathbb{R}_{\geq 0}$ because $\boldsymbol{\sigma}(t) \in \ker(Q^{+})$. Therefore, (5.7) holds and $\boldsymbol{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average lumpable.

(ii) $\Leftrightarrow$ (iii). If $\ker(Q^{+}A_{22}) \supseteq \ker(Q^{+})$, then $\mathtt{im}((Q^{+}A_{22})^{\intercal}) \subseteq \mathtt{im}((Q^{+})^{\intercal})$ because, for any matrix X, it holds that $\mathtt{im}(X^{\intercal}) = \ker(X)^{\perp}$, [Campbell2009, Proposition 0.2.1]. Equivalently, we have

$$\mathtt{rank}\left(\begin{bmatrix} Q^{+}A_{22} \\ Q^{+} \end{bmatrix}\right) = \mathtt{rank}\left(Q^{+}\right) \tag{5.8}$$

which implies that there exists a matrix $V \in \mathbb{R}^{k\times k}$ such that $VQ^{+} = Q^{+}A_{22}$. In words, the rows of $Q^{+}A_{22}$ are linearly dependent on the rows of Q^{+}, and that we can obtain $Q^{+}A_{22}$ by performing k row operations on Q^{+}. In the other direction, if there exists V such that $VQ^{+} = Q^{+}A_{22}$, then $Q^{+}A_{22}$ is linearly dependent on Q^{+} and (5.8) is satisfied. This implies that $\mathtt{im}((Q^{+}A_{22})^{\intercal}) \subseteq \mathtt{im}((Q^{+})^{\intercal})$, which is equivalent to the inclusion $\ker(Q^{+}A_{22}) \supseteq \ker(Q^{+})$.

□

To satisfy Theorem 5.4(iii), the characteristic matrix $Q \in \mathfrak{C}_{n,k}$ must be such that $\mathtt{im}(Q^{+\intercal})$ is a left-invariant subspace of A_{22}, which mirrors the notion of right-invariant subspace [Horn2013]. In other words, the average lumpability of $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is equivalent to $Q^+A_{22} = Q^+A_{22}QQ^+$.

Corollary 5.4.1. The following statements hold:

(i) If $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average detectable, then $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average lumpable.

(ii) If $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average lumpable and $Q^+A_{22}Q$ is Hurwitz, then $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average detectable.

Proof. The proof follows directly from Theorem 3.6 and 5.4. □

The above corollary shows the relation between average lumpability and average detectability. Basically, if $Q^+A_{22}Q$ is Hurwitz, then the two notions are equivalent. Therefore, for open-loop average estimation, we propose a clustering algorithm in the next subsection to obtain a clustered network system that is close to average lumpability. For this clustering algorithm, we suppose that Assumption 5.1 holds to ensure that $Q^+A_{22}Q$ is Hurwitz.

5.2.2 Clustering for open-loop average estimation

Given that $Q^+A_{22}Q$ is Hurwitz, then, by Theorem 3.8 and Corollary 5.4.1, a clustered network system $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is average detectable or average lumpable if and only if the clustering $\mathcal{Q}$ is equitable. In this subsection, we formulate a clustering problem that aims to minimize a distance of a clustered network system from average lumpability under a constraint that all intra-cluster induced subgraphs are weakly connected. This constraint is meaningful in physical network systems such as building thermal systems, urban traffic networks, and sensor networks [Martin2019]. After the problem formulation, we provide a suboptimal clustering algorithm and illustrate it through a simulation example.

Problem definition

Define the average approximation, or open-loop average estimation, error

$$\tilde{\mathbf{z}}_{\mathtt{a}}(t) = \mathbf{z}_{\mathtt{a}}(t) - \hat{\mathbf{z}}_{\mathtt{a}}(t) \tag{5.9}$$

where $\mathbf{z}_{\mathtt{a}}(t)$ satisfies (5.5) and $\hat{\mathbf{z}}_{\mathtt{a}}(t)$ satisfies (5.6). Then,

$$\dot{\tilde{\mathbf{z}}}_{\mathtt{a}}(t) = Q^+A_{22}Q\tilde{\mathbf{z}}_{\mathtt{a}}(t) + Q^+A_{22}\boldsymbol{\sigma}(t)$$

which is independent from the direct influence of the input $\mathbf{u}(t)$. However, notice that the input influences the average deviation $\boldsymbol{\sigma}(t)$, which in turn influences the dynamics of the error $\tilde{\mathbf{z}}_{\mathtt{a}}(t)$. Nonetheless, irrespective of the input $\mathbf{u}(t)$, we exploit the structural property $Q^+\boldsymbol{\sigma} \equiv \mathbf{0}_k$ of the average deviation vector $\boldsymbol{\sigma}(t)$.

The idea is to find the clustering $\mathcal{Q} = \{\mathcal{C}_1, \dots, \mathcal{C}_k\}$ such that the obtained clustered network system $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ is as close as possible to being average lumpable. That is, we minimize the distance from average lumpability of $\mathbf{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$ by finding clustering $\mathcal{Q}$. From Theorem 5.4(iii), we have that average lumpability is equivalent to $Q^+A_{22} = Q^+A_{22}QQ^+$. Therefore, the distance from average lumpability is defined as

$$\left|\left|\left|Q^+A_{22}(I_n - QQ^+)\right|\right|\right|$$

which we aim to minimize with respect to $Q \in \mathfrak{C}_{n,k}$ under the constraint that intra-cluster induced subgraph $\mathcal{G}_{\alpha\alpha} = (\mathcal{C}_\alpha, \mathcal{E}_{\alpha\alpha})$ formed by clusters $\mathcal{C}_\alpha$ are weakly connected,

for $\alpha = 1, \ldots, k$. However, to ensure average detectability, we assume that $Q^+ A_{22} Q$ is Hurwitz, which is ensured by Assumption 5.1 according to Theorem 4.7. Therefore, we assume that Assumption 5.1 holds and formulate the clustering problem as:

$$\begin{aligned} \min_{Q \in \mathfrak{C}_{n,k}} \quad & \mathcal{J}(Q) := |||Q^+ A_{22}(I_n - QQ^+)||| \\ \text{subject to} \quad & \mathcal{G}_{\alpha\alpha} = (\mathcal{C}_\alpha, \mathcal{E}_{\alpha\alpha}) \text{ is weakly connected, } \forall \alpha \in \{1, \ldots, k\}. \end{aligned} \tag{5.10}$$

Again, notice that the clustering problem (5.10) is a non-convex, mixed integer optimization problem [Burer2012]. Similar to the problem (5.4), finding the global optimum is NP-hard, therefore, we resort to a suboptimal solution that is computationally feasible and converges to a local minimum.

Clustering algorithm

Recall the induced subgraphs $\mathcal{G}_{\alpha\alpha} = (\mathcal{C}_\alpha, \mathcal{E}_{\alpha\alpha})$ formed by clusters $\mathcal{C}_\alpha$, for $\alpha = 1, \ldots, k$, where $\mathcal{E}_{\alpha\alpha} = \mathcal{E} \cap (\mathcal{C}_\alpha \cap \mathcal{C}_\alpha)$ is the set of edges between the unmeasured nodes of cluster $\mathcal{C}_\alpha$ in the system's digraph $\mathcal{G}$ with $|\mathcal{C}_\alpha| = n_\alpha$. When dealing with physical network systems, we often require that the clusters $\mathcal{C}_1, \ldots, \mathcal{C}_k$ be chosen such that the corresponding subgraphs $\mathcal{G}_{11}, \ldots, \mathcal{G}_{kk}$ are weakly connected. This constraint arises because one needs to interpret the average state of the clusters in order to monitor physical network systems. In the case of building thermal systems, for example, which is studied in Chapter 4, the average state of each cluster $\mathcal{C}_\alpha$ is the mean operative temperature of thermal elements corresponding to $\mathcal{C}_\alpha$. Thus, if the subgraph $\mathcal{G}_{\alpha\alpha}$ is not weakly connected, then the corresponding mean operative temperature is defined for elements that are far away from each other, which does not make sense in the case of building thermal systems.

Let $\overline{\mathcal{G}}_{\alpha\alpha}$ be the undirected version of $\mathcal{G}_{\alpha\alpha}$, where the edges are assumed to be undirected. The graph $\overline{\mathcal{G}}_{\alpha\alpha}$ has a weighted adjacency matrix $\mathcal{A}(\overline{\mathcal{G}}_{\alpha\alpha}) = \mathcal{A}(\mathcal{G}_{\alpha\alpha}) + \mathcal{A}(\mathcal{G}_{\alpha\alpha})^\intercal$, where $\mathcal{A}(\mathcal{G}_{\alpha\alpha})$ is the weighted adjacency matrix of the directed graph $\mathcal{G}_{\alpha\alpha}$. The Laplacian matrix of $\overline{\mathcal{G}}_{\alpha\alpha}$ is given by

$$\mathcal{L}(\overline{\mathcal{G}}_{\alpha\alpha}) = \mathtt{diag}(\mathcal{A}(\overline{\mathcal{G}}_{\alpha\alpha})\mathbf{1}_{n_\alpha}) - \mathcal{A}(\overline{\mathcal{G}}_{\alpha\alpha}).$$

The undirected version $\overline{\mathcal{G}}_{\alpha\alpha}$ of the original directed subgraph $\mathcal{G}_{\alpha\alpha}$ is considered because of the following properties:

- $\mathcal{G}_{\alpha\alpha}$ is weakly connected if and only if $\overline{\mathcal{G}}_{\alpha\alpha}$ is connected
- $\overline{\mathcal{G}}_{\alpha\alpha}$ is connected if and only if $\mathtt{rank}(\mathcal{L}(\overline{\mathcal{G}}_{\alpha\alpha})) = n_\alpha - 1$.

To ensure that the induced subgraphs $\mathcal{G}_{\alpha\alpha}$, for $\alpha = 1, \ldots, k$, remain weakly connected, we define two rules for the clustering algorithm.

(i) If, for some $i \in \mathcal{V}_2$ and $\alpha \in \{1, \ldots, k\}$, there exists $j \in \mathcal{C}_\alpha$ such that $(i, j) \in \mathcal{E}$ or $(j, i) \in \mathcal{E}$, then i is said to be adjacent to $\mathcal{C}_\alpha$, which is denoted as $i \leftrightarrow \mathcal{C}_\alpha$.

(ii) If $\overline{\mathcal{G}}_{\alpha\alpha}$ is connected and it remains connected after removing i from $\mathcal{C}_\alpha$, then i is said to be removable from $\mathcal{C}_\alpha$, which is denoted as $i \nleftrightarrow \mathcal{C}_\alpha$.

The first rule corresponds to adding nodes to cluster $\mathcal{C}_\alpha$. Due to the connectivity constraint in (5.10), only those nodes can be added to each cluster that are adjacent to that cluster. The second rule corresponds to removing nodes from cluster $\mathcal{C}_\alpha$ and adding them to other clusters. Again, the connectivity constraint demands that only those nodes can be removed from each cluster that do not disconnect the subgraph of that cluster.

Finally, recall the induced subgraph $\mathcal{G}_\nu = (\mathcal{V}_2, \mathcal{E}_\nu)$ formed by the unmeasured nodes $\mathcal{V}_2 = \{\nu_1, \dots, \nu_n\}$. Let $Q_{ij} \in \mathfrak{C}_{n,(n-1)}$ be the characteristic matrix of clustering

$$\mathcal{Q}_{ij} = \{\nu_1, \dots, \{\nu_i, \nu_j\}, \dots, \nu_n\}$$

with $n-1$ clusters of $\mathcal{V}_2$, where except for one cluster $\{\nu_i, \nu_j\}$ all the clusters consist of a single node ν_h, for $h = 1, \dots, n$ and $h \neq i$ or j, and

$$\mathfrak{C}_{n,(n-1)} = \{X \in \{0,1\}^{n\times(n-1)} : X\mathbf{1}_{n-1} = \mathbf{1}_n\}$$

is the set of characteristic matrices of all possible clusterings with $n-1$ clusters of n nodes. We define a weighted adjacency matrix $\mathcal{A}(\overline{\mathcal{G}}_\nu) \in \mathbb{R}^{n\times n}$ of the undirected version $\overline{\mathcal{G}}_\nu$ of the directed induced subgraph $\mathcal{G}_\nu$, whose edge weights are defined as the cost of aggregating ν_i and ν_j in one cluster, for $i, j = 1, \dots, n$ and $i \neq j$. That is,

$$[\mathcal{A}(\overline{\mathcal{G}}_\nu)]_{ij} = [\mathcal{A}(\overline{\mathcal{G}}_\nu)]_{ji} = \begin{cases} \mathcal{J}(Q_{ij}) & \text{if } (i,j) \in \mathcal{E}_\nu \text{ or } (j,i) \in \mathcal{E}_\nu \\ 0 & \text{otherwise} \end{cases} \tag{5.11}$$

where $\mathcal{J}(Q_{ij})$ is the value of cost function defined in (5.10). Finally, define a Laplacian matrix $\mathcal{L}(\overline{\mathcal{G}}_\nu) \in \mathbb{R}^{n\times n}$ as

$$\mathcal{L}(\overline{\mathcal{G}}_\nu) = \texttt{diag}(\mathcal{A}(\overline{\mathcal{G}}_\nu)\mathbf{1}_n) - \mathcal{A}(\overline{\mathcal{G}}_\nu). \tag{5.12}$$

Note that $\texttt{rank}(\mathcal{L}(\overline{\mathcal{G}}_\nu))$ indicates the number of connected components in the undirected graph $\overline{\mathcal{G}}_\nu$, i.e., if $\texttt{rank}(\mathcal{L}(\overline{\mathcal{G}}_\nu)) = n - \ell$, then $\overline{\mathcal{G}}_\nu$ has ℓ connected components, which correspond to weakly connected components of $\mathcal{G}_\nu$.

Algorithm 6 Initialization of k connected clusters

Input: Number of unmeasured nodes n, number of clusters k, and state matrix block A_{22}
Output: Clustering $\mathcal{Q}_0 = \{\mathcal{C}_1^0, \mathcal{C}_2^0, \dots, \mathcal{C}_k^0\}$

Obtain $\mathcal{G}_\nu = (\mathcal{V}_u, \mathcal{E}_\nu)$ and construct $\mathcal{A}(\overline{\mathcal{G}}_\nu)$ as in (5.11)
repeat
  Find an edge $(i_0, j_0) := \arg\max_{(i,j)\in\mathcal{E}_u} [\mathcal{A}(\overline{\mathcal{G}}_\nu)]_{ij}$
  $[\mathcal{A}(\overline{\mathcal{G}}_\nu)]_{i_0 j_0} = [\mathcal{A}(\overline{\mathcal{G}}_\nu)]_{j_0 i_0} = 0$
until $\texttt{rank}(\mathcal{L}(\overline{\mathcal{G}}_\nu)) = n - k$, where $\mathcal{L}(\overline{\mathcal{G}}_\nu)$ is in (5.12)
$c_{\max} := \arg\max_{\alpha\in\{1,\dots,k\}} |\mathcal{C}_\alpha|$ and $c_{\min} := \arg\min_{\alpha\in\{1,\dots,k\}} |\mathcal{C}_\alpha|$
repeat
  for each node $i \in \mathcal{C}_{c_{\max}}$ **do**
    if $i \hookleftarrow \mathcal{C}_{c_{\max}}$ and $i \leftrightarrow \mathcal{C}_{c_{\min}}$ **then**
      Move i from $\mathcal{C}_{c_{\max}}$ to $\mathcal{C}_{c_{\min}}$
    end if
  end for
  $c_{\max} := \arg\max_{\alpha\in\{1,\dots,k\}} |\mathcal{C}_\alpha|$ and $c_{\min} := \arg\min_{\alpha\in\{1,\dots,k\}} |\mathcal{C}_\alpha|$
until ($c_{\max} \leq \frac{n}{k}$ and $c_{\min} > 1$) or the maximum number of iterations
return $\mathcal{Q}_0 = \{\mathcal{C}_1^0, \mathcal{C}_2^0, \dots, \mathcal{C}_k^0\}$.

First, we initialize the k connected clusters, i.e., the clusters whose induced subgraphs $\mathcal{G}_{\alpha\alpha}$ are weakly connected, by employing Algorithm 6. In the first part, the algorithm finds k connected components by removing edges between nodes that yield high cost if put together in a single cluster. This is a heuristic to discard the worst pairs in the clustering problem. The second part of the algorithm balances the size of clusters in order to avoid disparity. This is necessary because too much disparity may result in a poorly initialized

clusters that may yield the final suboptimal clusters to be very different in sizes, which is not reasonable for average estimation in physical network systems. Then, the suboptimal clustering can be obtained by employing Algorithm 7 that takes the initial clustering from Algorithm 6 as an input.

Algorithm 7 Suboptimal clustering for open-loop average estimation

Input: Number of unmeasured nodes n, state matrix block A_{22}, and $\mathcal{Q}_0 = \{\mathcal{C}_1^0, \cdots, \mathcal{C}_k^0\}$
Output: Suboptimal clustering $\mathcal{Q} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k\}$

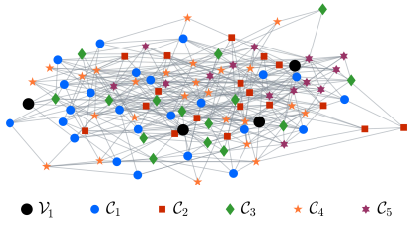
$\psi_0 \leftarrow \mathcal{J}(Q_0)$, where $Q_0 \in \mathfrak{C}_{n,k}$ is the characteristic matrix of $\mathcal{Q}_0$
$\psi^* \leftarrow \psi_0$ and $\mathcal{Q}_1 \leftarrow \mathcal{Q}_0$
repeat
  for each node $i = \nu_1, \nu_2, \cdots, \nu_n$ **do**
    Let β be such that $i \in \mathcal{C}_\beta$
    $\psi \leftarrow \psi^*$ and $\alpha \leftarrow \beta$
    if $|\mathcal{C}_\beta| > 1$ and $i \leftarrow\!\!\!\rightarrow \mathcal{C}_\beta$ **then**
      for $\gamma = 1, 2, \ldots, k$, $\gamma \neq \beta$ and $i \leftrightarrow \mathcal{C}_\gamma$ **do**
        Move node i from $\mathcal{C}_\beta$ into $\mathcal{C}_\gamma$ and update $\mathcal{Q}_1$
        Compute $\psi_1 = \mathcal{J}(Q_1)$, where $Q_1 \in \mathfrak{C}_{n,k}$ is the characteristic matrix of $\mathcal{Q}_1$
        if $\psi_1 < \psi$ **then**
          $\psi \leftarrow \psi_1$ and $\alpha \leftarrow \gamma$
        end if
        Move node i back to the cluster $\mathcal{C}_\beta$
      end for
      Move node i from $\mathcal{C}_\beta$ to $\mathcal{C}_\alpha$ and $\psi^* \leftarrow \psi$
    end if
  end for
until convergence or the maximum number of iterations
return $\mathcal{Q} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k\}$.

Simulation Example

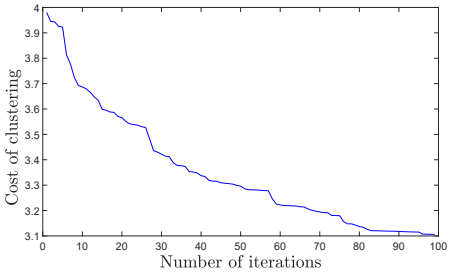
Suppose an undirected random graph $\mathcal{G}$ representing a network system $\boldsymbol{\Sigma}$ as shown in Figure 5.1(a) with 100 nodes. We assume 4 measured nodes $\mathcal{V}_1$ shown as black and find a suboptimal clustering $\mathcal{Q} = \{\mathcal{C}_1, \ldots, \mathcal{C}_5\}$ with 5 clusters for 96 unmeasured nodes. The state matrix $A = -\mathcal{L}(\mathcal{G})$ with $\mathcal{L}(\mathcal{G})$ the Laplacian matrix of $\mathcal{G}$, the input matrix $B \in \{0,1\}^{100\times 4}$ is generated randomly, and the input vector $\mathbf{u}(t) = [\ \sin t \quad \sin 5t \quad \sin 10t \quad \sin 50t\ ]^\intercal$. We initialize the clusters by using Algorithm 6, where the connected subgraphs formed by each cluster are shown in Figure 5.1(c). Then, Algorithm 7 finds a suboptimal clustering as shown in Figure5.1(a), where each cluster forms a connected induced subgraph as shown in Figure5.1(d). The cost minimization with respect to iterations is shown in Figure 5.1(b).

The projected system $\mathring{\boldsymbol{\Sigma}}_{\mathcal{V}_1,\mathcal{Q}}$ is obtained by aggregating the clusters, whose state is given by $\mathbf{z}(t) = [\ \mathbf{y}^\intercal(t) \quad \mathbf{z}_{\mathsf{a}}^\intercal(t)\ ]^\intercal$, where $\mathbf{y}(t)$ is the output of the original clustered network system $\boldsymbol{\Sigma}_{\mathcal{V}_1,\mathcal{Q}}$. Ignoring $\boldsymbol{\sigma}(t)$ in $\mathring{\boldsymbol{\Sigma}}_{\mathcal{V}_1,\mathcal{Q}}$ gives an approximated system whose state is given by $\hat{\mathbf{z}}(t) = [\ \hat{\mathbf{y}}^\intercal(t) \quad \hat{\mathbf{z}}_{\mathsf{a}}^\intercal(t)\ ]^\intercal$, where $\hat{\mathbf{y}}(t) = H\hat{\mathbf{z}}(t)$ and $\hat{\mathbf{z}}_{\mathsf{a}}(t)$ is given in (5.9). The norm of the output of the original network system $\boldsymbol{\Sigma}$, i.e., $\|\mathbf{y}(t)\|$; and the norm of the approximated output with initial clustering, i.e., $\|\hat{\mathbf{y}}^0(t)\|$, and with suboptimal clustering, i.e., $\|\hat{\mathbf{y}}^*(t)\|$, are shown in Figure 5.2(a). Likewise, the norm of states are shown in Figure 5.2(b).

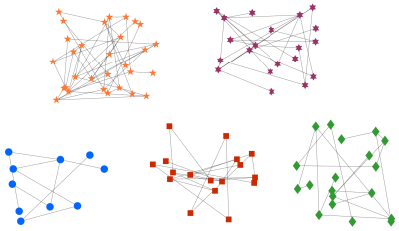
Figure 5.2(c) and (d) show the comparison between the errors for initial clustering and the suboptimal clustering. An interesting thing to note is that suboptimal clustering



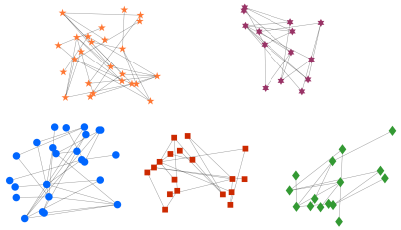
(a) Random network with 100 nodes, $\mathcal{V}_1$ are the measured nodes, and $\mathcal{C}_1, \cdots, \mathcal{C}_5$ are the obtained suboptimal clusters of unmeasured nodes.



(b) Cost minimization at each iteration.

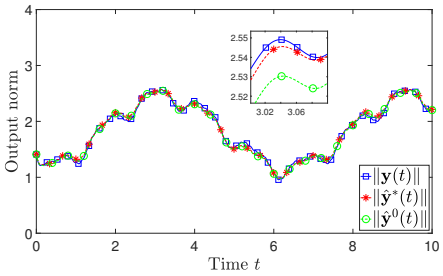


(c) Connected induced subgraphs formed by initial clusters obtained from Algorithm 6.

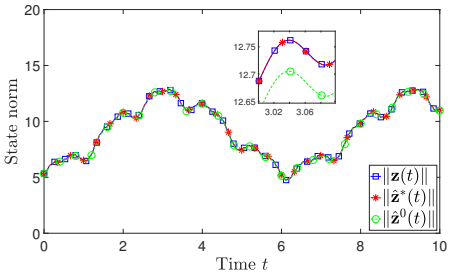


(d) Connected induced subgraphs formed by suboptimal clusters in (a) obtained from Algorithm 7 with initial clusters as in (c).

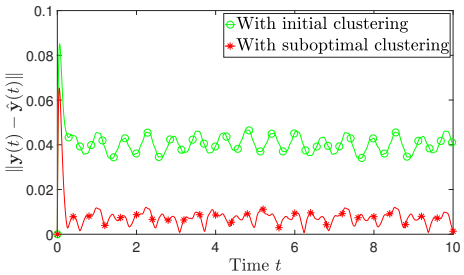
Figure 5.1: Illustration of the suboptimal clustering algorithm for open-loop average estimation



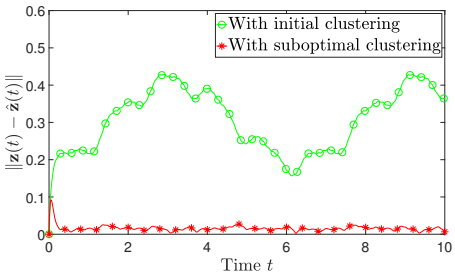
(a) Norm of the network system's output $\mathbf{y}(t)$ and the approximated outputs: $\hat{\mathbf{y}}^*(t)$ with suboptimal clustering and $\hat{\mathbf{y}}^0(t)$ with initial clustering.



(b) Norm of the projected system's state $\mathbf{z}(t)$ and the approximated state: $\hat{\mathbf{z}}^*(t)$ with suboptimal clustering and $\hat{\mathbf{z}}^0(t)$ with initial clustering.



(c) The norm of the output error with initial clustering and final (suboptimal) clustering.



(d) The norm of approximation error with initial clustering and final (suboptimal) clustering.

Figure 5.2: Approximation error minimization by the input-independent clustering algorithm.

reduces the error $\|\mathbf{z}(t) - \hat{\mathbf{z}}(t)\|$ by around 95% and the error $\|\mathbf{y}(t) - \hat{\mathbf{y}}(t)\|$ by around 80%. This suggests that the suboptimal clustering algorithm, on the one hand, is well-suited for control and estimation of the average states of clusters, and, on the other hand, it yields an approximated projected system whose input-output behavior is very similar to the input-output behavior of the original network system.

5.3 Clustering for State Variance Estimation

Large-scale network systems are ubiquitous in modern engineering applications such as traffic networks, building thermal systems, and distributed sensor networks. Complete monitoring of such large-scale systems is usually not possible due to limited computational and sensing resources. Limited computational resources can make the real-time state estimation task infeasible, whereas limited number of sensors may render the network system unobservable. It is reasonable, therefore, to monitor the network system by estimating the aggregated state profiles such as the average and variance. Previously, we have presented several methods to estimate the average states of multiple clusters of unmeasured nodes. In this section, we develop a methodology to estimate state variance of network systems in an approximate sense.

Recall the network system $\mathbf{\Sigma}$ with measured nodes $\mathcal{V}_1$ and unmeasured nodes $\mathcal{V}_2$, where $\mathbf{x}_1(t) \in \mathbb{R}^m$ is the state vector of $\mathcal{V}_1$ and $\mathbf{x}_2(t) \in \mathbb{R}^n$ is the state vector of $\mathcal{V}_2$. The state variance $x_{\mathrm{v}}(t) \in \mathbb{R}_{\geq 0}$ is a nonlinear functional that is defined to be the squared deviation of the states of unmeasured nodes from their average mean. That is,

$$\begin{aligned} x_{\mathrm{v}}(t) &= \frac{1}{n}\sum_{i\in\mathcal{V}_2}\left(x_i(t) - \sum_{j\in\mathcal{V}_2} x_j(t)\right)^2 \\ &= \frac{1}{n}\sum_{i\in\mathcal{V}_2} x_i^2(t) - \left(\frac{1}{n}\sum_{j\in\mathcal{V}_2} x_j(t)\right)^2 \\ &= \frac{1}{n}\mathbf{x}_2^{\mathsf{T}}(t) J_n \mathbf{x}_2(t) \end{aligned} \tag{5.13}$$

where the matrix $J_n = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathsf{T}}$ is symmetric ($J_n = J_n^{\mathsf{T}}$), idempotent ($J_n^2 = J_n$), and positive semi-definite with I_n the identity matrix of size $n \times n$ and $\mathbf{1}_n$ the vector of ones of size $n \times 1$.

5.3.1 Review of functional observers and their limitation

The fundamental concepts of nonlinear functional observers are presented in [Kazantzis2001, Kravaris2011, Kravaris2013, Kravaris2016], which can be employed to estimate the state variance. Given a nonlinear functional $x_{\mathrm{v}}(t)$, a functional observer of order k is a system

$$\begin{aligned} \dot{\mathbf{w}}(t) &= \mathbf{f}(\mathbf{w}(t), \mathbf{y}(t), \mathbf{u}(t)) \\ \hat{x}_{\mathrm{v}}(t) &= h(\mathbf{w}(t), \mathbf{y}(t)) \end{aligned} \tag{5.14}$$

with $\mathbf{f} : \mathbb{R}^k \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}^k$ and $h : \mathbb{R}^k \times \mathbb{R}^m \to \mathbb{R}$ designed such that the error

$$\xi_{\mathrm{v}}(t) = x_{\mathrm{v}}(t) - \hat{x}_{\mathrm{v}}(t)$$

converges to zero asymptotically. Such a functional observer is possible if and only if there exists an invariant manifold $\mathbf{w} = \mathbf{g}(\mathbf{x})$ such that

$$\begin{aligned} \frac{\partial \mathbf{g}(\mathbf{x})}{\partial \mathbf{x}}\big(A\mathbf{x} + B\mathbf{u}\big) &= \mathbf{f}(\mathbf{g}(\mathbf{x}), \mathbf{y}, \mathbf{u}) \\ x_{\mathrm{v}} &= h(\mathbf{g}(\mathbf{x}), \mathbf{y}). \end{aligned} \tag{5.15}$$

For linear systems, we need to find a linear map $\mathbf{g}(\mathbf{x}) = P^{\intercal}\mathbf{x}$, where $P \in \mathbb{R}^{(m+n)\times k}$, in order to satisfy the condition (5.15). This can be stated as follows:

Proposition 5.5. Consider an observer (5.14) with

$$\begin{aligned} \mathbf{f}(\mathbf{w}(t), \mathbf{y}(t), \mathbf{u}(t)) &= M\mathbf{w}(t) + K\mathbf{y}(t) + N\mathbf{u}(t) \\ h(\mathbf{w}(t), \mathbf{y}(t)) &= \mathbf{w}^{\intercal}(t)D\mathbf{w}(t) + \mathbf{y}^{\intercal}(t)L\mathbf{y}(t) \end{aligned} \tag{5.16}$$

where $M \in \mathbb{R}^{k\times k}$ is a Hurwitz matrix, $K \in \mathbb{R}^{k\times m}$, $N \in \mathbb{R}^{k\times p}$, $D \in \mathbb{R}^{k\times k}$, and $L \in \mathbb{R}^{m\times m}$. Then, the error $x_{\mathrm{v}}(t) - \hat{x}_{\mathrm{v}}(t)$ converges to zero asymptotically if and only if there exists a matrix $P \in \mathbb{R}^{(m+n)\times k}$ such that

$$P^{\intercal}A - MP^{\intercal} = KC \tag{5.17a}$$

$$PDP^{\intercal} = T \tag{5.17b}$$

where A is the state matrix of the network system $\mathbf{\Sigma}$ and

$$T = \begin{bmatrix} -L & 0_{m\times n} \\ 0_{n\times m} & \frac{1}{n}J_n \end{bmatrix}.$$

Proof of sufficiency. Define $\mathbf{e}_1(t) = P^{\intercal}\mathbf{x}(t) - \mathbf{w}(t)$, then

$$\dot{\mathbf{e}}_1(t) = M\mathbf{e}_1(t) + (P^{\intercal}A - MP^{\intercal} - KC)\mathbf{x}(t)$$

where we chose $N = P^{\intercal}B$. Moreover, define $\xi_{\mathrm{v}}(t) = x_{\mathrm{v}}(t) - \hat{x}_{\mathrm{v}}(t)$, then

$$\xi_{\mathrm{v}}(t) = \mathbf{x}^{\intercal}(t)(T - PDP^{\intercal})\mathbf{x}(t) + \mathbf{e}_1^{\intercal}(t)D\mathbf{e}_1(t) - 2\mathbf{x}^{\intercal}(t)PD\mathbf{e}_1(t).$$

Assume (5.17a) and (5.17b) hold. Then, (5.17b) implies that $\xi_{\mathrm{v}}(t) = \mathbf{e}_1^{\intercal}(t)D\mathbf{e}_1(t) - 2\mathbf{x}^{\intercal}(t)PD\mathbf{e}_1(t)$ and (5.17a) implies that $\mathbf{e}_1(t) = \exp(Mt)\mathbf{e}_1(0)$. Since M is a Hurwitz matrix, we have $\mathbf{e}_1(t) \to \mathbf{0}_k$ exponentially as $t \to \infty$ for all $\mathbf{e}_1(0) \in \mathbb{R}^k$. Therefore, $\xi_{\mathrm{v}}(t) \to 0$ asymptotically as $t \to \infty$.

Proof of necessity. Assume $\xi_{\mathrm{v}}(t) \to 0$ as $t \to \infty$, then

$$\lim_{t\to\infty} (\mathbf{x}^{\intercal}(t)(T - PDP^{\intercal})\mathbf{x}(t) + \mathbf{e}_1^{\intercal}(t)D\mathbf{e}_1(t) - 2\mathbf{x}^{\intercal}(t)PD\mathbf{e}_1(t)) = 0.$$

In general, when $\lim_{t\to\infty} \mathbf{x}(t) \neq \mathbf{0}_{m+n}$, the above equation implies (5.17b) and

$$\lim_{t\to\infty} \mathbf{e}_1(t) = \mathbf{0}_k$$

which is true only if (5.17a) holds.

□

To design a functional observer of the form (5.16) under the constraint that M is Hurwitz, one has to determine the order k and find $P \in \mathbb{R}^{(n+m)\times k}$ that satisfies (5.17). Finding a minimal order k such that the Sylvester equation (5.17a) is solvable is known to be quite challenging, [Fernando2010a, Rotella2016]. Moreover, in order to solve (5.17b), we see that the order k of the observer must be at least $n-1$. To elucidate this fact, we suppose $P^\intercal = [\, P_1^\intercal \;\; P_2^\intercal \,]$ with $P_1 \in \mathbb{R}^{m\times k}$ and $P_2 \in \mathbb{R}^{n\times k}$, then (5.17b) can be written as

$$\begin{bmatrix} P_1 D P_1^\intercal & P_1 D P_2^\intercal \\ P_2 D P_1^\intercal & P_2 D P_2^\intercal \end{bmatrix} = \begin{bmatrix} -L & 0_{m\times n} \\ 0_{n\times m} & \frac{1}{n} J_n \end{bmatrix}.$$

Apart from $P_1 D P_1^\intercal = -L$, $P_1 D P_2^\intercal = 0_{m\times n}$ and $P_2 D P_1^\intercal = 0_{n\times m}$, we also need to satisfy $P_2 D P_2^\intercal = \frac{1}{n} J_n$, which implies that $\mathtt{rank}(P_2) \geq n-1$ because $\mathtt{rank}(J_n) = n-1$. Hence, it is necessary that $k \geq n-1$, which is a lower bound on the order of functional observer of the form (5.16).

Even if the functional observer is of minimum order, i.e., $k = n-1$, the estimation is still not feasible because n can be very large. Such an observer estimates all but one states of the unmeasured nodes to compute the state variance. This is because $\mathtt{rank}(J_n) = n-1$ and $\mathbf{1}_n^\intercal J_n = 0$, which means that if we estimate $n-1$ elements of the vector $J_n \mathbf{x}_2(t) = \mathbf{x}_2(t) - \mathbf{1}_n x_{\mathtt{a}}(t)$, the n-th element equals the negative sum of the estimated $n-1$ elements. The problem of interest, however, is to estimate the variance $x_{\mathtt{v}}(t)$ without estimating the whole vector $J_n \mathbf{x}_2(t)$, which is not possible due to the limitation on the order of the functional observer. Therefore, instead of the asymptotic estimation, i.e., $x_{\mathtt{v}}(t) - \hat{x}_{\mathtt{v}}(t) \to 0$ as $t \to \infty$, we would like to find an optimal approximate estimation solution, where the order k is chosen according to the available computational capability.

5.3.2 K-means type clustering for state variance estimation

The infeasibility of designing a nonlinear functional observer directs us towards estimation of state variance in an approximate sense. That is, we first approximate the state variance by partitioning $\mathcal{V}_2$ into k clusters such that the states of nodes in each cluster can be approximated by its average state, which is similar to a K-means clustering problem [Abonyi2007]. The approximated state variance is then computed from the average states of the clusters. Then, we employ the $\mathcal{H}_2$-optimal average state observer to estimate the average states of clusters, which gives us an estimated state variance.

Problem definition

Let $k < n$ be the given number of clusters and $\mathcal{Q} = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$ be the clustering of the unmeasured nodes $\mathcal{V}_2 = \{\nu_1, \ldots, \nu_n\}$, where $\mathcal{C}_1 \cup \cdots \cup \mathcal{C}_k = \mathcal{V}_2$ and $\mathcal{C}_\alpha \cap \mathcal{C}_\beta = \emptyset$, for $\alpha, \beta \in \{1, \ldots, k\}$ and $\alpha \neq \beta$. The characteristic matrix $Q \in \mathfrak{C}_{n,k}$ of the clustering $\mathcal{Q}$ is as defined before in (2.5).

The rationale for an approximated state variance is as follows: If $Q \in \mathfrak{C}_{n,k}$ is such that

$$\mathbf{x}_2(t) \approx Q \mathbf{z}_{\mathtt{a}}(t)$$

where $\mathbf{z}_{\mathtt{a}}(t) = Q^+ \mathbf{x}_2(t)$ and $Q^+ = (Q^\intercal Q)^{-1} Q^\intercal$, then, from (5.13),

$$x_{\mathtt{v}}(t) \approx \frac{1}{n} \mathbf{z}_{\mathtt{a}}^\intercal(t) Q^\intercal J_n Q \mathbf{z}_{\mathtt{a}}(t).$$

That is, if a clustering is such that the states of all nodes in a cluster $\mathcal{C}_\alpha$, for $\alpha = 1, \ldots, k$, can be approximated by their average $z_\alpha(t) = \frac{1}{n_\alpha} \sum_{j\in\mathcal{C}_\alpha} x_j(t)$, then the state variance can

be approximated as

$$\begin{aligned} x_{\mathtt{v}}(t) &\approx \frac{1}{n}\mathbf{z}_{\mathtt{a}}^{\intercal}(t)Q^{\intercal}J_nQ\mathbf{z}_{\mathtt{a}}(t) \\ &= \frac{1}{n}\sum_{\alpha=1}^{k} n_\alpha z_\alpha^2(t) - \left(\frac{1}{n}\sum_{\alpha=1}^{k} n_\alpha z_\alpha(t)\right)^2 \end{aligned} \tag{5.18}$$

where $\mathbf{z}_{\mathtt{a}}(t) = [\, z_1(t) \cdots z_k(t) \,]^{\intercal} \in \mathbb{R}^k$ is the average state vector and $n_\alpha = |\mathcal{C}_\alpha|$ with $\sum_{\alpha=1}^{k} n_\alpha = n$.

Recall the average deviation vector $\boldsymbol{\sigma}(t) = \mathbf{x}_2(t) - Q\mathbf{z}_{\mathtt{a}}(t)$ with i-th entry, for $i = 1, \ldots, n$, given by $\sigma_i(t) = x_{\nu_i}(t) - z_\alpha(t)$, where $\nu_i \in \mathcal{C}_\alpha$ and $\alpha \in \{1, \ldots, k\}$. That is, the entries of $\boldsymbol{\sigma}(t)$ are the differences between the states of unmeasured nodes and the average states of the corresponding clusters. We can write

$$\boldsymbol{\sigma}(t) = D_Q\mathbf{x}(t)$$

where

$$D_Q = [\; 0_{n\times m} \quad I_n - QQ^{+} \;].$$

Then, the transfer function from $\mathbf{u}$ to $\boldsymbol{\sigma}$ is given by

$$\mathbf{T}(s) = D_Q(sI - A)^{-1}B$$

with the $\mathcal{H}_2(\tau)$-norm defined as, see [Sinani2019],

$$\|\mathbf{T}\|_{\mathcal{H}_2(\tau)}^2 = \mathtt{trace}(D_Q W_\tau D_Q^{\intercal})$$

where, for some $\tau \in \mathbb{R}_{>0}$,

$$W_\tau = \int_0^\tau \exp(At)BB^{\intercal}\exp(A^{\intercal}t)dt \tag{5.19}$$

is the finite-horizon controllability grammian of the network system $\boldsymbol{\Sigma}$. If the state matrix A is Hurwitz, then the standard $\mathcal{H}_2$-norm can also be considered, which can be computed by using the infinite-horizon controllability grammian [Sinani2019].

The K-means type clustering problem is defined as follows: Find $Q \in \mathfrak{C}_{n,k}$ such that

$$\min_{Q\in\mathfrak{C}_{n,k}} \quad \mathcal{J}(Q) := \mathtt{trace}(D_Q W_\tau D_Q^{\intercal}) \tag{5.20}$$

where $\mathfrak{C}_{n,k} = \{X \in \{0,1\}^{n\times k} : X\mathbf{1}_k = \mathbf{1}_n\}$. The clustering problem (5.20) is a non-convex, mixed-integer NP-hard optimization problem.

Clustering algorithm

In this subsection, we provide a suboptimal clustering algorithm for solving (5.20) in polynomial time. Let $\psi = \mathtt{trace}(D_Q W_\tau D_Q^{\intercal})$ be the cost of (5.20) for some $Q \in \mathfrak{C}_{n,k}$, which is the characteristic matrix of the clustering $\mathcal{Q}$. Similarly, let $\psi_0 = \mathtt{trace}(D_{Q_0} W_\tau D_{Q_0}^{\intercal})$ be the cost of a randomly initialized clustering $\mathcal{Q}_0$.

The suboptimal solution obtain from Algorithm 8 depends on the initial clustering $\mathcal{Q}_0$. Therefore, to obtain a better solution, the Algorithm 9 repeatedly runs Algorithm 8, where at every iteration the clustering is initialized randomly.

Algorithm 8 Suboptimal K-means clustering

Input: Number of unmeasured nodes n, number of clusters k, the controllability grammian W_τ, an initial clustering $\mathcal{Q}_0$, and a stopping criterion $\delta > 0$ (e.g., 10^{-6})
Output: Suboptimal clustering $\mathcal{Q} = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$

Compute $\psi_0 = \mathcal{J}(Q_0)$ and assign $\mathcal{Q}_1 \leftarrow \mathcal{Q}_0$
repeat
  Assign $\psi_1 \leftarrow \psi_0$
  for $i = 1, 2, \cdots, n$ **do**
    Assign $\mathcal{Q}_2 \leftarrow \mathcal{Q}_1$
    Let $\beta \in \{1, \ldots, k\}$ be such that $\nu_i \in \mathcal{C}_\beta$
    if $\mathcal{C}_\beta$ has more than one node, i.e., $|\mathcal{C}_\beta| > 1$, **then**
      for $\alpha = 1, 2, \ldots, k$ and $\alpha \neq \beta$ **do**
        Move ν_i from its cluster to $\mathcal{C}_\alpha$
        Update $\mathcal{Q}_2$ accordingly and compute $\psi_2 = \mathcal{J}(Q_2)$
        if $\psi_2 < \psi_0$ **then**
          Update $\psi_0 \leftarrow \psi_2$ and $\mathcal{Q}_1 \leftarrow \mathcal{Q}_2$.
        end if
      end for
    end if
  end for
until $\psi_1 - \psi_0 < \delta$
Assign $\mathcal{Q} \leftarrow \mathcal{Q}_1$
return $\mathcal{Q} = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$, $\psi = \texttt{trace}(D_Q W_\tau D_Q^\intercal)$.

Algorithm 9 Improving suboptimal K-means clustering

Input: Number of unmeasured nodes n, number of clusters k, the controllability grammian W_τ, an initial clustering $\mathcal{Q}_0$, and a stopping criterion $\delta > 0$ (e.g., 10^{-6}), maximum value of counter $c > 0$
Output: Suboptimal clustering $\mathcal{Q}^* = \{\mathcal{C}_1^*, \ldots, \mathcal{C}_k^*\}$

Assign $a \leftarrow 0$ and $b \leftarrow 0$
repeat
  Compute $\psi_0 = \texttt{trace}(D_{Q_0} W_\tau D_{Q_0}^\intercal)$
  Run Algorithm 8 and store Q and ψ
  Assign $a \leftarrow a + 1$
  if $a = 1$ **then**
    Assign $\psi^* \leftarrow \psi$ and $\mathcal{Q}^* \leftarrow \mathcal{Q}$
    Randomly initialize $\mathcal{Q}_0$ and compute D_{Q_0}
  else
    if $\psi < \psi^*$ **then**
      Assign $\psi^* \leftarrow \psi$ and $\mathcal{Q}^* \leftarrow \mathcal{Q}$
      Randomly initialize $\mathcal{Q}_0$ and compute D_{Q_0}
    else
      Assign $b \leftarrow b + 1$
    end if
  end if
until $b \leq c$
return $\mathcal{Q}^* = \{\mathcal{C}_1^*, \cdots, \mathcal{C}_k^*\}$, $\psi^* = \texttt{trace}(D_{Q^*} W_\tau D_{Q^*}^\intercal)$.

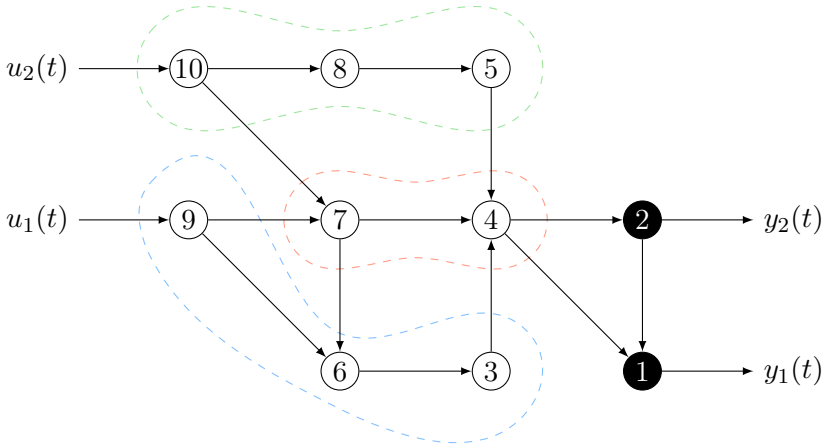
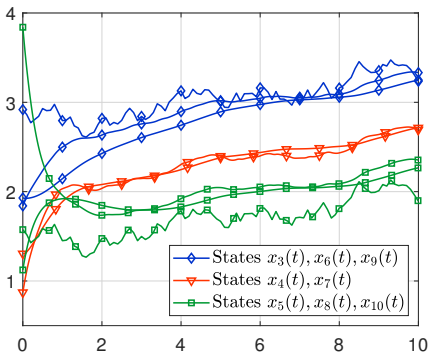


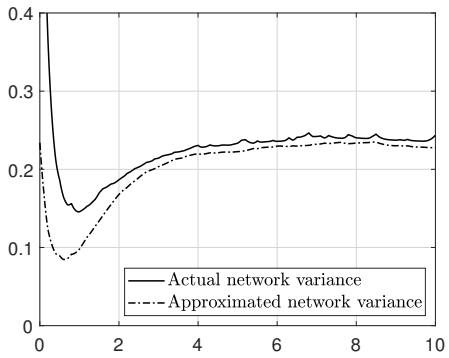
Figure 5.3: Three clusters (enclosed by the dashed lines) identified by Algorithm 9 in the network.

Example 5.1. Consider a network system shown in Figure 5.3, where the input $\mathbf{u} = [\,u_1\ u_2\,]^\intercal$ and the output $\mathbf{y} = [\,y_1\ y_2\,]^\intercal$. The measured nodes $\mathcal{V}_1 = \{1, 2\}$ and the unmeasured nodes $\mathcal{V}_2 = \{3, 4, \ldots, 10\}$. Let the number of clusters be $k = 3$. Then, the clustering $\mathcal{Q}^* = \{\mathcal{C}_1^*, \mathcal{C}_2^*, \mathcal{C}_3^*\}$ is obtained from Algorithm 9, where $\mathcal{C}_1^* = \{3, 6, 9\}$, $\mathcal{C}_2^* = \{4, 7\}$, and $\mathcal{C}_3^* = \{5, 8, 10\}$, which are specified by the dashed lines in Figure 5.3.

The clustering obtained by Algorithm 9 ensures that the state trajectories of each cluster stay closer to each other as time progresses, shown in Figure 5.4(a). For instance, initially the states $x_5(0), x_8(0), x_{10}(0)$ are not close to each other, however, as $t > 1$, we see that their trajectories converge closer to each other. Consequently, the state variance can be approximated as shown in Figure 5.4(b). ⌟



(a) State trajectories of clustered unmeasured nodes of the network system shown in Figure 5.3.



(b) The plots of actual state variance $x_{\mathrm{v}}(t)$ computed by (5.13) from the states of unmeasured nodes and approximated state variance computed by (5.18) from the average mean values of the identified clusters.

Figure 5.4: Approximation of the state variance of a network system.

State Variance Estimation

After obtaining a suboptimal clustering $\mathcal{Q}^*$ from Algorithm 9, we design an $\mathcal{H}_2$-optimal average observer $\mathbf{\Omega}_{\mathcal{V}_1,\mathcal{Q}^*}$ to estimate the average state vector of clusters. Then, the estimated state variance is given by

$$\hat{x}_{\mathtt{v}}(t) = \frac{1}{n}\hat{\mathbf{z}}_{\mathtt{a}}^{\mathsf{T}}(t)Q^{\mathsf{T}}J_nQ\hat{\mathbf{z}}_{\mathtt{a}}(t) \tag{5.21}$$

where $\hat{\mathbf{z}}_{\mathtt{a}}(t) \in \mathbb{R}^k$ is the estimated average state vector by $\mathbf{\Omega}_{\mathcal{V}_1,\mathcal{Q}^*}$. Since the average estimation error $\boldsymbol{\zeta}(t) = \mathbf{z}_{\mathtt{a}}(t) - \hat{\mathbf{z}}_{\mathtt{a}}(t)$, where $\mathbf{z}_{\mathtt{a}}(t) \in \mathbb{R}^k$ is the average state vector of clusters, we have

$$\begin{aligned}\hat{x}_{\mathtt{v}}(t) &= \frac{1}{n}(\mathbf{z}_{\mathtt{a}}(t) - \boldsymbol{\zeta}(t))^{\mathsf{T}}Q^{\mathsf{T}}J_nQ(\mathbf{z}_{\mathtt{a}}(t) - \boldsymbol{\zeta}(t)) \\ &= \frac{1}{n}\left(\mathbf{z}_{\mathtt{a}}^{\mathsf{T}}(t)Q^{\mathsf{T}}J_nQ\mathbf{z}_{\mathtt{a}}(t) + \boldsymbol{\zeta}^{\mathsf{T}}(t)Q^{\mathsf{T}}J_nQ\boldsymbol{\zeta}(t) - 2\mathbf{z}_{\mathtt{a}}(t)Q^{\mathsf{T}}J_nQ\boldsymbol{\zeta}(t)\right)\end{aligned}$$

and since $\mathbf{x}_2(t) = \boldsymbol{\sigma}(t) + Q\mathbf{z}_{\mathtt{a}}(t)$, we have, from (5.13),

$$\begin{aligned}x_{\mathtt{v}}(t) &= \frac{1}{n}(\boldsymbol{\sigma}(t) + Q\mathbf{z}_{\mathtt{a}}(t))^{\mathsf{T}}J_n(\boldsymbol{\sigma}(t) + Q\mathbf{z}_{\mathtt{a}}(t)) \\ &= \frac{1}{n}\left(\boldsymbol{\sigma}^{\mathsf{T}}(t)\boldsymbol{\sigma}(t) + \mathbf{z}_{\mathtt{a}}^{\mathsf{T}}(t)Q^{\mathsf{T}}J_nQ\mathbf{z}_{\mathtt{a}}(t)\right)\end{aligned}$$

where we used the facts that

$$\boldsymbol{\sigma}^{\mathsf{T}}(t)J_n\boldsymbol{\sigma}(t) = \boldsymbol{\sigma}^{\mathsf{T}}(t)\boldsymbol{\sigma}(t) \text{ and } \boldsymbol{\sigma}^{\mathsf{T}}(t)J_nQ\mathbf{z}_{\mathtt{a}}(t) = 0.$$

Therefore, the state variance estimation error $\xi_{\mathtt{v}}(t) := x_{\mathtt{v}}(t) - \hat{x}_{\mathtt{v}}(t)$ is given by

$$\xi_{\mathtt{v}}(t) = \frac{1}{n}\left(\boldsymbol{\sigma}^{\mathsf{T}}(t)\boldsymbol{\sigma}(t) + (2\mathbf{z}_{\mathtt{a}}(t) - \boldsymbol{\zeta}(t))^{\mathsf{T}}Q^{\mathsf{T}}J_nQ\boldsymbol{\zeta}(t)\right). \tag{5.22}$$

The above expression contains two summands. The first summand is the square of the norm of state variance approximation error $\|\boldsymbol{\sigma}(t)\|^2$ and the second is proportional to the average state estimation error $\boldsymbol{\zeta}(t)$. If the optimization problems (5.20) and (4.13) admit a solution that yields a small cost, then the state variance approximation error and the average estimation error will also be small. Consequently, the state variance estimation error will be small.

Simulation Example

As a simulation example, we consider a linear flow network, [Walter1999], where each compartment is a node with a state $x_i(t) \in \mathbb{R}_{\geq 0}$ that represents some physical quantity in i. The nodes $\mathcal{V}$ are connected via an underlying graph $\mathcal{G}$ and the rate of change of node i's state equals the difference between the inflow to i and the outflow from i. The inflow is what i receives from its in-neighbors and the positive external inputs. The outflow is what i gives out to its out-neighbors and the negative external inputs. That is,

$$\dot{x}_i(t) = f_i^{\text{in}}(t) - f_i^{\text{out}}(t),$$

where $f_i^{\text{in}}(t)$ and $f_i^{\text{out}}(t)$ represent the inflow and the outflow, respectively, which are given by

$$\begin{aligned}f_i^{\text{in}}(t) &= \sum_{j\in\mathcal{N}_{i\leftarrow\mathcal{V}}} a_{ij}x_j(t) + b_{il}^{+}u_l^{+}(t), \\ f_i^{\text{out}}(t) &= \sum_{h\in\mathcal{N}_{i\rightarrow\mathcal{V}}} a_{hi}x_i(t) + b_{il}^{-}u_l^{-}(t)\end{aligned}$$

with $\mathcal{N}_{i\leftarrow\mathcal{V}}$ and $\mathcal{N}_{i\rightarrow\mathcal{V}}$ the in-neighbors and the out-neighbors of node i, respectively; $u_l^+(t) \in \mathbb{R}_{\geq 0}$, $u_l^-(t) \in \mathbb{R}_{\leq 0}$ are the positive and the negative inputs, respectively, $b_{il}^+, b_{il}^- \in \{0,1\}$ are the scalars that determine if the input-l is applied at node i, and $a_{ij} \geq 0$ for all $i, j \in \mathcal{V}$ and $i \neq j$.

The state matrix $A = \Lambda - \mathcal{L}(\mathcal{G})$, where the Laplacian matrix $\mathcal{L}(\mathcal{G}) = \mathcal{D}^{\downarrow}(\mathcal{G}) - \mathcal{A}(\mathcal{G})$ with $\mathcal{D}^{\downarrow}(\mathcal{G})$ the weighted in-degree matrix and $\mathcal{A}(\mathcal{G})$ the weighted adjacency matrix of the graph $\mathcal{G}$, and the diagonal matrix $\Lambda = \mathcal{D}^{\uparrow}(\mathcal{G}) + \mathcal{D}^{\downarrow}(\mathcal{G})$ with $\mathcal{D}^{\uparrow}(\mathcal{G})$ the weighted out-degree matrix of $\mathcal{G}$. Then, the state matrix $A = \mathcal{A}(\mathcal{G}) - \mathcal{D}^{\uparrow}(\mathcal{G})$ and $\mathbf{1}_\ell^\intercal A = 0_{1\times\ell}$. There are many applications modeled as above, see [Walter1999], including traffic networks in free flow [Rodriguez-Vega2020, Rodriguez-Vega2021].

We generate a graph $\mathcal{G}$ of 55 nodes by an Erdos-Renyi model with a probability of a directed edge between any pair of nodes equal to 0.15. The number of measured nodes $m = 5$ and the number of unmeasured nodes $n = 50$. We choose the number of clusters to be $k = 5$. We consider the input vector to be $\mathbf{u}(t) = [\, u_1(t) \;\; u_2(t) \,]^\intercal$, where each input $u_l(t) = u_l^+(t) - u_l^-(t)$ with $u_l^+(t), u_l^-(t) \in [0,1]$ representing random, discontinuous signals. The inputs are directly applied to 10% of nodes chosen in a uniformly random way.

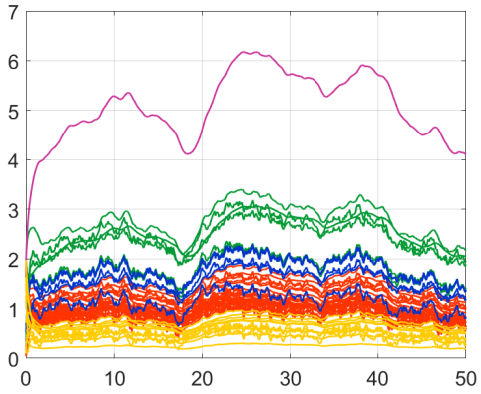


Figure 5.5: The state trajectories $\mathbf{x}_2(t)$ of the 50 unmeasured nodes in the example network system. The colors of the trajectories correspond to the 5 clusters identified by Algorithm 9.

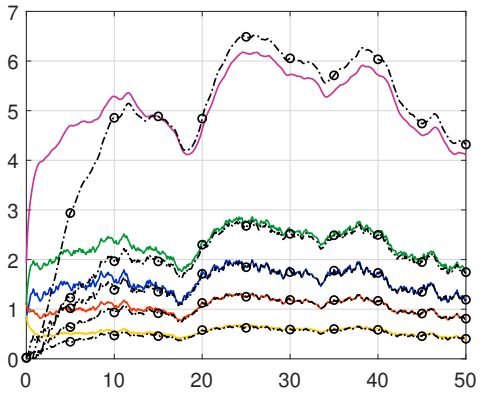


Figure 5.6: Optimal average estimation with the optimal tuning parameter $\rho^* = 1.8879$ obtained from Algorithm 2. The colored solid trajectories show the average states of the clusters $\mathbf{z}_\mathbf{a}(t)$ and black dashed trajectories show the estimated average states $\hat{\mathbf{z}}_\mathbf{a}(t)$.

We compute W_τ for $\tau = 10$ from the expression (5.19) and run Algorithm 9 to obtain suboptimal clustering of unmeasured nodes with $k = 5$ clusters. The state trajectories of

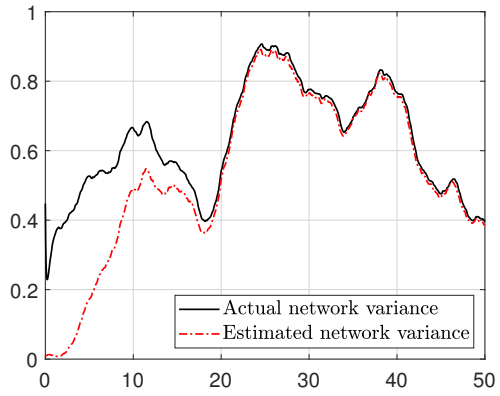


Figure 5.7: The actual state variance $x_{\mathrm{v}}(t)$ plotted with black solid line vs. the estimated state variance $\hat{x}_{\mathrm{v}}(t)$ plotted with red dotted line.

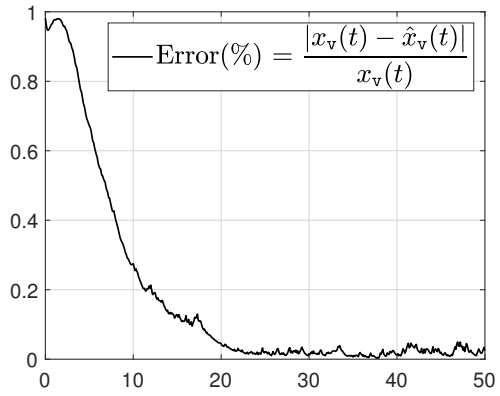


Figure 5.8: Percentage variance estimation error.

the system are shown in Figure 5.5, where the trajectories of the same color correspond to the nodes of the same cluster.

We then design the average observer $\mathbf{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$ and obtain the optimal tuning parameter $\rho^* = 1.8879$ from Algorithm 2, where W_ρ is computed at every iteration using the expression (4.12). The estimation of the average states are shown in Figure 5.6. Notice that the estimation of the four average states shown as green, blue, red, and yellow is very accurate, whereas the estimation of the average state shown as magenta is not very accurate. This can be due to the fact that the cluster represented by magenta has only one node, which is an outlier node since its trajectory is far from the states of other nodes. The average state observer is designed to optimally estimate the average states of a cluster of several nodes, and not the states of individual nodes. Another reason can be the fact that we optimize a single parameter ρ in problem (4.13), which changes the eigenvalues of $M_\rho = R_\rho Q$ with an equal proportion. After obtaining an optimal ρ^*, if the spectrum $\texttt{eig}(M_\rho)$ contains a very small eigenvalue, then the corresponding estimated average state trajectory will not be accurate, as shown in the figure.

We compute the actual state variance from the state trajectories of unmeasured nodes using the expression in (5.13), and the estimated state variance from the estimated average state trajectories using the expression in (5.21). The plot of the actual and the estimated state variance is shown in Figure 5.7 and the percentage state variance estimation error in Figure 5.8. We see that the state variance estimation is very accurate, which is due to the following reasons: (i) The variance approximation is very accurate because of the state trajectories of the identified clusters are very close to each other, and (ii) The average estimation is very accurate because of the optimal tuning parameter ρ^*. From the discussion that follow after the variance estimation error equation (5.22), we conclude that the accuracy of state variance approximation and average estimation results in the accuracy of state variance estimation.

5.4 Application Example: SIS Epidemics over Networks

Modeling and analysis of spreading phenomenon has been a topic of interest not only in mathematical epidemiology but also in computer networks [Pastor-Satorras2001b], wireless communication [Kleinberg2007], statistical physics [Grassberger1983], and social sciences [Boccaletti2006]. This is because, in addition to disease spreading in networks of biological beings, the spreading phenomenon described by the epidemic models also captures virus spreading in computer networks or rumor spreading in social networks. Nonetheless, epidemic models are very crucial in understanding and devising preventive measures and control strategies to mitigate the disease spread as will be discussed in detail in the second part II of this thesis.

The main idea of epidemic models is to consider compartments of different populations, which are divided on the basis of whether they are susceptible, exposed, infected, or recovered. The most common models are (i) susceptible-infected (SI), (ii) susceptible-infected-susceptible (SIS), (iii) susceptible-infected-recovered (SIR), and (iv) susceptible-exposed-infected-recovered (SEIR). The susceptible population in these models is prone to disease and recovered population is immune. In SI model, once the susceptible people are infected, they stay infected and do not recover. In SIS model, the infected people recover but do not attain immunity and become susceptible again. In SIR model, the infected people recover and attain permanent immunity. In SEIR model, the susceptible people do not get infected right away after a contact with infected population,

they first become exposed, then transition to being infected, and then recover with permanent immunity. There is a vast body of literature on epidemic models as evidenced by [Hethcote2000, Brauer2012]. These epidemic models are population models that consider lumped population in each compartment by assuming a homogeneous population structure, where the underlying contact network is assumed to be complete. In this complete structure, all people are connected to each other and thus are equally likely to get infected. These models are simple and fail to capture the inherent network structure embedded in the epidemic spread process, which often results in imprecise estimation of the epidemic situation. The shortcomings posed by the population models are overcome by the networked epidemic models, which are studied in [Pastor-Satorras2001a, Newman2002, Pastor-Satorras2015, Khanafer2016, Nowzari2016, Mei2017, Paré2020]. In this section, we apply our clustering-based optimal average estimation method for a networked SIS epidemic model, where SIS pattern is suitable for infectious diseases in humans such as tuberculosis, meningitis, and gonorrhea [Allen2008, Keeling2011].

5.4.1 SIS epidemic model over networks

We consider a networked metapopulation SIS epidemic model that comprises several groups of population interacting with each other over a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of groups or nodes and $\mathcal{E}$ is the set of edges specifying interactions among the groups. For instance, the nodes of the network may correspond to population of several cities and edges to transportation links between those cities. Denote $x_i(t) \in [0, 1]$ to be the proportion of infected people in node i, then the networked SIS model is given by

$$\dot{x}_i(t) = \beta(1 - x_i(t)) \sum_{j \in \mathcal{N}_{i \leftarrow \mathcal{V}}} a_{ij} x_j(t) - \gamma x_i(t)$$

where $a_{ij} > 0$ is the edge weight of $(i, j) \in \mathcal{E}$, β is the infection rate of i, and γ is the recovery rate of i. Similar to the setup of this thesis, we assume that the proportion of infected people in m nodes are measured, which are $x_1(t), \ldots, x_m(t)$, and the rest of the n nodes are unmeasured.

Let $\mathcal{A} := \mathcal{A}(\mathcal{G})$ be the weighted adjacency matrix of graph $\mathcal{G}$ defined as

$$[\mathcal{A}]_{ij} = \begin{cases} a_{ij} & \text{if } (i, j) \in \mathcal{E} \\ 0 & \text{otherwise.} \end{cases}$$

Then, the vector form of the model is

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \beta(I_{m+n} - \mathtt{diag}(\mathbf{x}(t)))\mathcal{A}\mathbf{x}(t) - \gamma \mathbf{x}(t) \\ \mathbf{y}(t) &= C\mathbf{x}(t) \end{aligned} \tag{5.23}$$

where

$$\begin{aligned} \mathbf{x}(t) &= [\ x_1(t) \quad \ldots \quad x_{m+n}(t)\ ]^\intercal \\ \mathtt{diag}(\mathbf{x}(t)) &= \mathtt{diag}(x_1(t), \ldots, x_{m+n}(t)) \\ \mathbf{y}(t) &= [\ x_1(t) \quad \ldots \quad x_m(t)\ ]^\intercal \\ C &= [\ I_m \quad 0_{m \times n}\ ]. \end{aligned}$$

5.4.2 Simulation results for clustering-based optimal average estimation

Simulation setup

We generate an Erdős-Rényi graph of 100 nodes, where the probability of a directed edge between any pair of nodes is 0.15. We suppose the number of measured nodes $m = 5$, the number of unmeasured nodes $n = 95$, and the number of clusters $k = 4$. The edge weights a_{ij} of the graph are chosen randomly in the interval $(0, 1)$ and the initial condition $\mathbf{x}(0)$ is chosen in the interval $(0, 0.1)^{m+n}$.

The infection rate $\beta = 0.02$ and the recovery rate $\gamma = 0.12$, where $\beta\lambda_{\max}(\mathcal{A})/\gamma = 1.1$, which is greater than one and implies that the networked SIS dynamics will not converge to zero (i.e., disease-free equilibrium) but it will converge to an endemic state, [Mei2017].

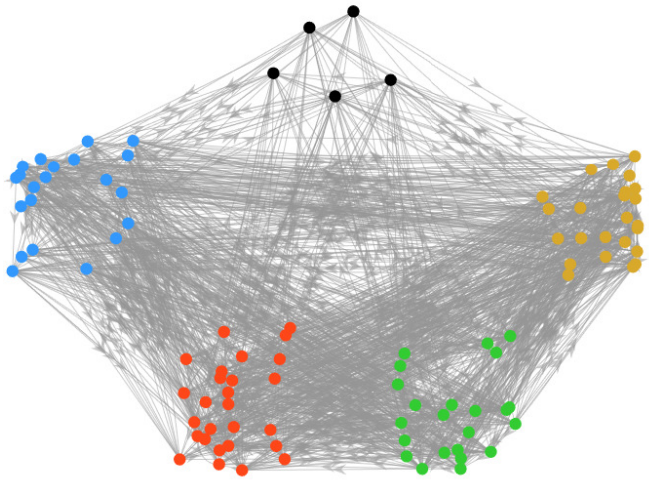


Figure 5.9: Suboptimal clustering obtained by Algorithm 5.

Suboptimal clustering and optimal gain

Linearization around the equilibrium point $\mathbf{x}(t) \approx \mathbf{0}_{m+n}$ yields a linear model

$$\dot{\mathbf{x}}(t) = A\mathbf{x}(t)$$

where $A = \beta\mathcal{A} - \gamma I_{m+n}$, which is partitioned as

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} := \begin{bmatrix} \beta\mathcal{A}_{11} - \gamma I_m & \beta\mathcal{A}_{12} \\ \beta\mathcal{A}_{21} & \beta\mathcal{A}_{22} - \gamma I_n \end{bmatrix}.$$

Using this linear model, we solve Problem (5.4) using Algorithm 5 with $W_{\rho,Q}, M_{\rho,Q}, R_{\rho,Q}$ defined in (5.3), (5.1), and (5.2), respectively. We obtain the optimal gain value $\rho^* = 3.0002$ and the suboptimal clustering $\mathcal{Q}^*$ illustrated in Figure 5.9, where the black nodes are measured nodes and the colored nodes respectively depict the clusters of unmeasured nodes.

The matrices of average observer $\mathbf{\Omega}_{\mathcal{V}_1,\mathcal{Q}^*}$ with suboptimal clustering $\mathcal{Q}^*$ are obtained

as follows

$$M = \begin{bmatrix} -0.1153 & 0.0375 & 0.0373 & 0.0275 \\ 0.0162 & -0.1016 & 0.0230 & 0.0213 \\ 0.0157 & 0.0317 & -0.1030 & 0.0246 \\ 0.0164 & 0.0347 & 0.0332 & -0.0981 \end{bmatrix}$$

$$K = \begin{bmatrix} 0.0018 & 0.0024 & 0.0080 & 0.0066 & 0.0046 \\ 0.0019 & 0.0019 & 0.0057 & 0.0022 & 0.0031 \\ 0.0029 & 0.0038 & 0.0057 & 0.0049 & 0.0043 \\ 0.0024 & 0.0022 & 0.0062 & 0.0042 & 0.0052 \end{bmatrix}$$

$$L = \begin{bmatrix} 0.0761 & 0.0085 & 0.0750 & -0.0679 & -0.0044 \\ -0.0410 & 0.0088 & 0.0242 & 0.0686 & 0.0744 \\ 0.0063 & 0.0015 & 0.1188 & 0.0525 & 0.0080 \\ 0.0503 & 0.0510 & 0.0154 & 0.0249 & 0.0184 \end{bmatrix}$$

where $M = M_{\rho,Q}$ defined in (5.1) with Q being the characteristic matrix of $\mathcal{Q}^*$. Notice that $N = 0$ because there is no input to the system.

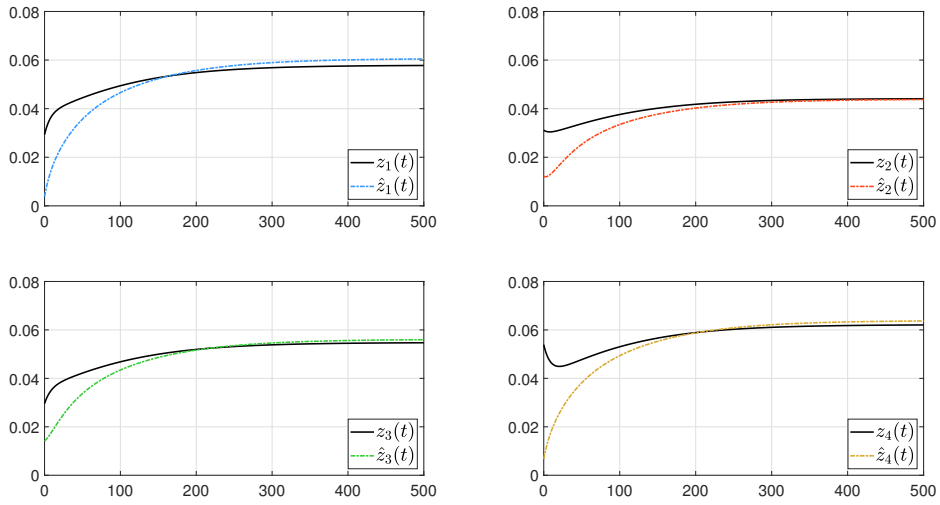


Figure 5.10: Average estimation of four clusters of unmeasured nodes in Figure 5.9.

Simulation results

The estimation of average states of four clusters is illustrated in Figure 5.10, where the solid black lines are the original average states of clusters given by

$$\mathbf{z_a}(t) = Q^{+}\mathbf{x}_2(t) = \begin{bmatrix} z_1(t) & z_2(t) & z_3(t) & z_4(t) \end{bmatrix}^{\intercal}$$

with $\mathbf{x}_2(t)$ the unmeasured state vector of the model (5.23), and the dotted colored lines are the estimated average states of clusters given by the output of the average observer $\mathbf{\Omega}_{\mathcal{V}_1,\mathcal{Q}}$

$$\hat{\mathbf{z}}_\mathbf{a}(t) = \begin{bmatrix} \hat{z}_1(t) & \hat{z}_2(t) & \hat{z}_3(t) & \hat{z}_4(t) \end{bmatrix}^{\intercal}.$$

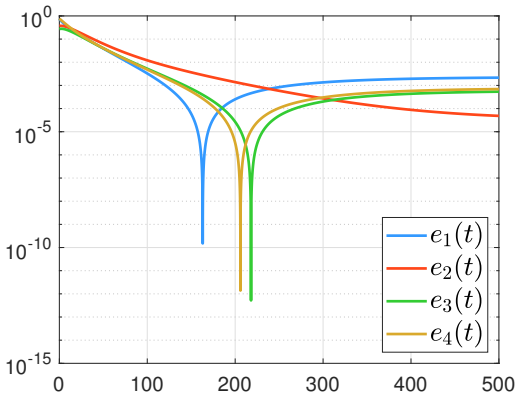


Figure 5.11: Percentage average estimation error.

In Figure 5.11, we illustrate the percentage average estimation errors

$$e_\alpha(t) = \frac{(z_\alpha(t) - \hat{z}_\alpha(t))^2}{(z_\alpha(t))^2}$$

for $\alpha = 1, 2, 3, 4$. The final values of percentage average estimation errors for the four clusters are 0.2160%, 0.0048%, 0.0537%, and 0.0693%, respectively, which are very small and show that the average estimation is reliable.

5.5 Concluding Remarks

We presented clustering algorithms for optimal average estimation, open-loop average estimation, and state variance estimation. We provided a sufficient condition in terms of a clustering constraint for ensuring the generic stabilizability of the average observer. A result on generic rank is employed to ensure the stabilizability because the usual rank condition may not hold for cases that are of Lebesgue measure zero. The clustering algorithm for optimal average estimation considers this sufficient condition of generic stabilizability as a constraint and seeks a local minimum solution. We consider an application example of the SIS epidemic over a large network to show the efficacy of our methodology.

In the clustering methodology of open-loop estimation, we defined a notion of average lumpability and showed the following relation to average detectability:

$$\begin{array}{rcl} \text{average detectability} & \Rightarrow & \text{average lumpability} \\ \text{average lumpability AND } Q^{+}A_{22}Q \text{ Hurwitz} & \Leftrightarrow & \text{average detectability.} \end{array}$$

The clustering algorithm aims to minimize the distance from average lumpability given the number of clusters and the connectivity constraint of clusters. A simulation example showed the effectiveness of the clustering algorithm for open-loop average estimation.

Finally, we showed the limitation of estimating a state variance of a network system by a nonlinear functional observer, where the order of the observer has to be the number of unmeasured nodes minus one. Such an order is not computationally feasible for large-scale network systems. We provided a K-means type clustering methodology to obtain a clustered network system and employ an optimal average observer to approximately estimate the state variance. The state trajectories of nodes in each cluster obtained by the

K-means clustering algorithm are shown to comparatively close to each other, therefore, can be approximated by the average mean. This allows us to approximate the state variance by the average states of clusters. We showed the effectiveness of this approach through a simulation example. The future direction in this regard is to estimate the variance of each cluster in a clustered network system.

Part II

Modeling and Control of Epidemics

6 State of the Art

In this chapter, we provide a literature review of epidemic modeling and control. After providing a general review of different models and control techniques that exist in the recent literature, we focus on works that study policies for testing and urban human mobility to control epidemics. Then, we provide our contributions in relation to the existing literature.

6.1 Literature review

The literature on mathematical epidemiology is very vast. However, we mainly focus on recent papers that appeared following the outbreak of COVID-19. Each model, by and large, is a variant and/or an extension of SIR (susceptible, infected, recovered) and SEIR (susceptible, exposed, infected, recovered) models, which describe the flow of population through three or four mutually exclusive stages of infection, respectively (see [Kermack1927] and [Hethcote2000] for a comprehensive review). These basic models have few parameters that are easy to identify [Massonis2020] and are considered as population models that view the epidemic from the macroscopic perspective. This is in contrast with the approaches that capture the heterogeneity of population structure such as network epidemic models [Khanafer2016, Paré2018a, Paré2018b] or metapopulation epidemic models [Colizza2008, Pastor-Satorras2015, Della Rossa2020], that view the epidemic from the microscopic perspective.

Recent epidemic models, however, are more complex and comprehensive than simple SIR and SEIR. These models include several intermediate stages to accurately portray the dynamics of the epidemic with each focusing on different facets of the epidemic to understand, predict, and control the evolution of the COVID-19 epidemic. For instance, [Lin2020] develops an extension of the SEIR model that incorporates the governmental actions (e.g., preventive measures and restrictions) and the individual behavioral reactions, whereas [Anastassopoulou2020] develops an extension of the SIR model that incorporates the number of deaths due to the epidemic. Another quite interesting model is the one developed in [Giordano2020] that considers an eight-compartment model called SIDARTHE, which includes eight stages of infection: susceptible (S), infected (I),

diagnosed (D), ailing (A), recognized (R), threatened (T), healed (H), and extinct (E). A distinguishing feature of this model is that it differentiates between the infected individuals based on the severity of their symptoms and whether they are diagnosed by a health authority. It is crucial, as also emphasized in [Liu2020, Ducrot2020], to differentiate between diagnosed and undiagnosed individuals because the former are typically isolated and are less likely to spread the infection. Similar models have been adopted and extended to study optimal control policies for the epidemic such as the implementation of social distancing measures [Köhler2020, Morato2020, Perkins2020], lockdown strategies [Casella2020, Olivier2020, Alvarez2020], and heterogeneous policy responses based on age-groups [Acemoglu2020, Brotherhood2020].

In addition to the above NPI strategies, testing and isolating the infected population from the susceptible population is one of the most important strategies to control the epidemic spread. The emphasis of the director-general of WHO on testing was very strong when he gave his message "test, test, test" to all countries. This is because testing is well-known to be a crucial control mechanism for epidemics [Chowell2003]. It limits the spread of disease to the susceptible population by enabling the health authority to detect and isolate the infected people.

In somewhat similar to a resource allocation problem [Nowzari2016, Nowzari2017] in epidemic control, [Pezzutto2020] poses the optimal test allocation as a well-known sensor selection problem in control theory, whereas [Ely2020] poses it as a welfare maximization problem by considering specificity and sensitivity of tests. The main assumption in these papers, however, is the availability of information portfolios of all individuals in a society, which enables the decision-makers to compute the infection probability of individuals and utility loss for each individual in case of decision errors. On the other hand, [Piguillem2020, Berger2020] study the problem of testing policy from an economic perspective, where the goal is to find an optimal testing policy that minimizes the total number of quarantined people to incur minimal cost on the economic activity of a country while also mitigating the epidemic spread. Without such a testing policy, governments usually resort to indiscriminate quarantining of people that burdens the economy of a country without any reason. Therefore, testing allows to identify and isolate the positive cases to allow for case-dependent quarantining. A different aspect of the testing policy is studied in [Charpentier2020], which aims to find an optimal trade-off between testing effort and lockdown intervention under the constraint of limited Intensive Care Units (ICU). However, in all these studies, testing policies lack a control-theoretic perspective even though it is a very crucial control mechanism. Given a current situation of an epidemic, there is no result providing the number of tests needed to be performed per day in order to control the evolution of the epidemic.

Another focal point for controlling an epidemic is urban human mobility. It plays a vital role in the economy of a country, however, when there is an epidemic, it facilitates the spread of disease by allowing contact between infected and susceptible populations. Considering a SIR epidemic model for disease spread, human mobility between different geographic regions has been investigated and modeled in [Sattenspiel1995, Arino2003, Balcan2010, Poletto2013]. In these models, the individuals associated with one region or city can go to and return from other regions. However, these models capture averaged mobility patterns between different cities with large timescales and cannot capture the daily patterns of mobility within an urban environment. To tackle this problem, [Frias-Martinez2011, Pappalardo2015, Pastor-Satorras2015, Nadini2020] study agent-based models of urban human mobility with epidemic spread. These models are powerful tools for computational purposes, however, they rely on the digital footprints of individuals and can lead to privacy violations. A similar line of research in [Song2020] aims to find control policies that restrict mobility to and from regions that are estimated to be of high risk

by employing a reinforcement learning framework and relying on aggregated demand for mobility and regional epidemic statistics. However, a more practical approach is to find optimal capacities and schedules of locations, such as workplaces, schools, and markets, where the nominal density of people is high, in order to mitigate the epidemic.

6.2 Our Contributions

In Chapter 7, we introduce a new model named SIDUR model — susceptible (S), undiagnosed infected (I), diagnosed infected (D), unidentified recovered (U), and identified removed (R) — to study the control of an epidemic through testing. Similar to [Giordano2020, Liu2020, Ducrot2020], we differentiate between the undiagnosed and diagnosed infected population. We assume that the diagnosed infected population are either quarantined and/or hospitalised and only the undiagnosed infected population is responsible for the disease transmission to the susceptible population. The identified removed population consists of people who recover or die after being diagnosed and the unidentified recovered population consists of people who recover without getting diagnosed.

The control input in the SIDUR model is defined as the number of tests performed per day, where the influence of the control is directly linked with the testing specificity. The testing specificity determines the probability of detecting an undiagnosed infected person through a test, which, for instance, can be increased through efficient contact tracing. Notice that COVID-19 can be detected through two types of tests known as type-1 (RT-PCR) and type-2 (serology). In the type-1 test, a swab is inserted into the subject's nose to qualitatively detect nucleic acid from SARS-CoV-2 in the upper and lower respiratory specimens [FDA2020], which enables one to detect whether the subject is currently infected with COVID-19. Type-2 test, on the other hand, is a serum test in order to detect relevant antibodies, which enables one to know whether the subject was infected in the past with COVID-19 and now he/she has recovered. Both types of tests are important in the control of an epidemic. Type-1 tests help to limit the disease spread by the identification of infected individuals and their contact tracing [de Walque2020]. Type-2 tests, on the other hand, are useful in reducing the size of the testable population for type-1 tests [Winter2020] that helps to increase the testing specificity. However, the type-1 test, up to now, is considered to be the only recommended method for the identification and laboratory confirmation of COVID-19 cases according to the WHO [WHO2020]. Moreover, only type-1 tests can provide information in real-time related to describe the outburst of the epidemic, which is the reason that the datasets related to testing only include type-1 tests[1]. Therefore, we assume that the control input in the SIDUR model only accounts for the type-1 (RT-PCR) tests.

We consider the COVID-19 case of France as a benchmark example in Chapter 7. That is, we estimate and validate the model on French COVID-19 data. Then, we propose a testing policy, the so-called best-effort strategy for testing (BEST), for epidemic suppression. The BEST policy provides the minimum number of tests to be performed per day in order to stop the epidemic spread. Thus, BEST is meaningful only during the spreading phase of the disease. We provide an algorithm to compute the number of tests required by BEST policy. Since BEST is a suppression strategy that stops the epidemic growth immediately, it usually requires a lot of tests to be performed per day. However, it requires

[1]Website: Our World in Data: Coronavirus (COVID-19) Testing. (Accessed 30/09/2020)

less number of tests if implemented sooner, which is illustrated for the case of France by plotting the number of tests required by BEST with respect to time.

In Chapter 8, we develop an urban human mobility model that captures the daily mobility patterns and incorporates the process of epidemic spread at each location. Every day a certain number of people go from their residential areas, which are called *origins*, to locations visited daily for work, education, shopping, etc., which are called *destinations*, and return on the same day. The daily mobility patterns are captured by the time-dependent supply and demand gating functions. The *supply gating function* (SGF) of each destination is controlled by its daily destination schedule, which is its opening and closing hours. The *demand gating function* (DGF), on the other hand, is defined on each edge of the mobility network and corresponds to the daily mobility window, which is the time interval during which people utilize that edge to move between origins and destinations. The supply function of each destination controlled by the SGF determines the inflow allowed to that destination and depends on its operating capacity controlled by the capacity control input. The demand function controlled by the DGF determines the outflow from one location to another. The process of urban human mobility is modeled on the network edges that connect different locations through flows and the process of epidemic spread is modeled locally at each location that depends on the number of susceptible and infected people at that location.

We formulate two optimal control problems in Chapter 8 for epidemic mitigation while maximizing the economic activity: (i) optimal capacity control policy and (ii) optimal schedule control policy. These problems aim to find an optimal capacity control input and schedule control input, respectively, that maximizes the economic activity while mitigating the epidemic by keeping the number of active infected cases bounded. The capacity control policy restrict the number of people in destinations of each category by specifying the operating capacities in relation to their nominal capacities. The schedule control inputs, on the other hand, specifies the closing hour of destinations of each category by altering the destination schedules and mobility windows. The effectiveness of these policies is shown numerically for an example of two origins and three destinations.

7

Design of Testing Policy for Epidemic Suppression

In this chapter, we develop an epidemic model that incorporates the testing rate as a control input in section 7.1. After presenting and imputing the data for the COVID-19 case of France in section 7.2, we estimate and validate the model in section 7.3. Then, a suppression strategy, called the best-effort strategy for testing (BEST), is proposed in section 7.4, which provides a lower bound on the testing rate such that the epidemic switches from a spreading to a non-spreading state. To evaluate the BEST policy, we predict its impact on the number of active intensive care unit (ICU) cases and the cumulative number of deaths due to COVID-19 in France.

Contents

Testing for infected cases is one of the most important mechanisms to control an epidemic. It enables the isolation of the detected infected individuals, thereby limiting the disease transmission to the susceptible population. However, despite the significance of testing policies in epidemic control, the literature on this subject lacks a control-theoretic perspective. In this chapter, we develop an epidemic model that incorporates the testing rate as a control input. The proposed model differentiates the undetected infected from the detected infected cases, who are assumed to be removed from the disease spreading process in the population. We consider the COVID-19 case of France to estimate and validate the model. Then, we propose a suppression strategy, the so-called best-effort strategy for testing (BEST), which provides a lower bound on the testing rate such that the epidemic switches from a spreading to a non-spreading state. The BEST policy is evaluated by predicting the impact on the number of active intensive care unit (ICU) cases and the cumulative number of deaths due to COVID-19 in France.

7.1 Formulation of SIDUR Epidemic Model

We develop a five-compartment model with the purpose of evaluating and devising a testing policy for epidemic suppression. We assume that testing allows for diagnosing and isolating the infected people from the population to prevent the transmission of the disease to the susceptible population. The acronym of the proposed model is SIDUR, where the letters correspond to five compartments: susceptible (S), undiagnosed infected (I), diagnosed infected (D), unidentified recovered (U), and identified removed (R). The model is characterized by four parameters and one control input, which is the testing rate.

7.1.1 Model design

Consider a compartmental model SIDUR depicted in Figure 7.1. At any time $t \in \mathbb{R}_{\geq 0}$, where t is measured in days, each compartment is characterized by a single state, which is its population denoted as follows:

$x_{\mathtt{S}}(t)$ Number of susceptible people

$x_{\mathtt{I}}(t)$ Number of undiagnosed infected people

$x_{\mathtt{D}}(t)$ Number of diagnosed infected people

$x_{\mathtt{U}}(t)$ Number of unidentified recovered people

$x_{\mathtt{R}}(t)$ Number of identified removed people.

The susceptible people are prone to the disease and can get infected when they come in contact with the infected people. The undiagnosed infected people are those who are undetected and can infect others, whereas the diagnosed infected people are those who are detected positive with the disease and are isolated. Finally, the unidentified recovered people are those who were infected and then recovered naturally without getting detected, whereas the identified removed people are those who were infected and then recovered or died after getting detected positive with the disease.

Assumption 7.1. We adopt the following assumptions:

(i) The population remains constant during the evolution of the epidemics, i.e.,

$$x_{\mathtt{S}}(t) + x_{\mathtt{I}}(t) + x_{\mathtt{D}}(t) + x_{\mathtt{U}}(t) + x_{\mathtt{R}}(t) = N$$

where N stands for the total population.

(ii) Only the undiagnosed infected population $x_{\mathtt{I}}(t)$ is responsible for the disease transmission to the susceptible population $x_{\mathtt{S}}(t)$.

(iii) All the deaths from epidemic are identified and reported, and are included in the removed population $x_{\mathtt{R}}(t)$ along with the people who recover after being diagnosed.

(iv) The efficiency of the acquired immunity is sustainable enough. That is, the unidentified recovered population $x_{\mathtt{U}}(t)$ and the removed population $x_{\mathtt{R}}(t)$ are not infected again.

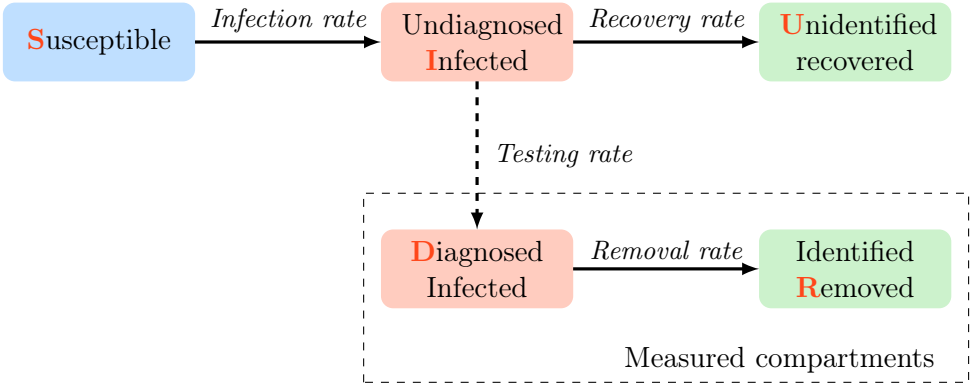


Figure 7.1: Block diagram of SIDUR model.

Based on the above assumptions, the model is given as

$$\dot{x}_{\mathtt{S}}(t) = -\beta(t)x_{\mathtt{S}}(t)\frac{x_{\mathtt{I}}(t)}{N} \tag{7.1a}$$

$$\dot{x}_{\mathtt{I}}(t) = \beta(t)x_{\mathtt{S}}(t)\frac{x_{\mathtt{I}}(t)}{N} - u(t)\frac{x_{\mathtt{I}}(t)}{x_{\mathtt{T}}(t)} - \gamma x_{\mathtt{I}}(t) \tag{7.1b}$$

$$\dot{x}_{\mathtt{D}}(t) = u(t)\frac{x_{\mathtt{I}}(t)}{x_{\mathtt{T}}(t)} - \rho x_{\mathtt{D}}(t) \tag{7.1c}$$

$$\dot{x}_{\mathtt{U}}(t) = \gamma x_{\mathtt{I}}(t) \tag{7.1d}$$

$$\dot{x}_{\mathtt{R}}(t) = \rho x_{\mathtt{D}}(t) \tag{7.1e}$$

where $\beta(t) \geq 0$, $\gamma \geq 0$, and $\rho \geq 0$ are the infection, recovery, and removal rates, respectively, $u(t)$ is the testing rate,

$$\begin{aligned} x_{\mathtt{T}}(t) &= x_{\mathtt{I}}(t) + (1-\theta(t))\left(x_{\mathtt{S}}(t) + x_{\mathtt{U}}(t)\right) \\ &= \theta(t)x_{\mathtt{I}}(t) + (1-\theta(t))\left(N - x_{\mathtt{D}}(t) - x_{\mathtt{R}}(t)\right) \end{aligned} \tag{7.2}$$

is the testable population, and $\theta(t) \in [0,1]$ is the testing specificity parameter.

The recovery rate γ is the inverse of the average recovery time $1/\gamma$ after which an undiagnosed infected person recovers, and the removal rate ρ is the inverse of the average removal time $1/\rho$ after which a diagnosed infected person recovers or dies. The average recovery time is expected to be shorter than the average removal time because the undiagnosed infected population that comprises the undetected asymptomatic cases and cases with mild symptoms usually recover faster than the diagnosed population that comprises mostly the cases with severe symptoms.

The infection rate $\beta := \beta(t)$ is the product of the frequency of contacts among the susceptible and infected populations and the probability of disease transmission after a contact has been made. In fact the parameters β, γ, and ρ are related to the disease biology. However, the value of β can also be partially impacted by a country's government through non-pharmaceutical interventions (NPI) such as social distancing, lockdown, confinement, travel restrictions, and preventive policies (i.e., to maintain a certain distance from other people, to wear a face mask in public spheres, to wash/sanitize hands more often, etc.). The value of β is expected to be smaller when NPI's are implemented in comparison to the case when no NPI is implemented. Depending on the time intervals during which different policies concerning NPIs are implemented, we consider the infection rate β to be piecewise constant.

The testing specificity parameter $\theta := \theta(t)$, unlike other parameters, is solely dependent on the testing policy of the public health authority. Given that the testing rate is constant, the value of θ will be larger when the tests are allocated efficiently through contact tracing than the value of θ when the tests are performed randomly. However, there are other factors that can also influence θ, for example, if only the people with severe symptoms are tested, then the probability $x_{\mathtt{I}}/x_{\mathtt{T}}$ of detecting an infected person from the testable population is equal to one, i.e., the testing specificity parameter $\theta = 1$. This is to indicate that the larger value of θ doesn't necessarily imply the efficiency of testing policy, rather it signifies only the specificity of tests. Depending on the time intervals during which different testing policies are implemented, we consider the testing specificity parameter θ to be piecewise constant.

7.1.2 Control input and testable population

We consider the testing rate $u(t)$ to be the control input of the model, which is the number of tests performed per day. The testing rate allows to detect a proportion of the testable population $x_{\mathtt{T}}(t)$, which is a sample from the total population N, in order to diagnose the infected people at time t. From (7.2), it is obvious that the infected population $x_{\mathtt{I}}(t) \leq x_{\mathtt{T}}(t)$ at any given time t. Thus, given the testing specificity parameter $\theta \in [0, 1]$, the probability of detecting an infected person per test in a homogeneous population structure is given by $x_{\mathtt{I}}(t)/x_{\mathtt{T}}(t)$.

The testing specificity parameter θ allows for the adjustment of the testable population to accommodate for the detection rate of tests. In most countries, at the beginning of an epidemic outbreak, the number of available tests are limited. Thus, the available tests are usually utilized to confirm the symptomatic infected cases or to diagnose certain people such as medical care agents, politicians, athletes, etc. In such a case, the testable population is close to the infected population and the value of θ increases to approximately one. Once the capacity of testing is increased, the size of the testable population is also increased that can include, for example, contacts of diagnosed people, the whole population of a city where a cluster is identified, travellers, etc. As a consequence, the value of θ decreases.

7.1.3 Outflows from the model compartments

The SIDUR model is described by the one-way transfer of population between compartments, where an outflow from one compartment is the inflow to the other compartments. It suffices, therefore, to describe only the outflows from the compartments for describing the dynamics of the model.

Infection transmission In the beginning of the epidemic, most of the population is in the susceptible compartment (S) with the exclusion of those who are initially infected and/or diagnosed. Some of the susceptible people in S may get infected and leave this compartment when they come in contact with an infected person. The rate of the outflow from this compartment is according to the infection transmission rate, which depends on the product of the number of susceptible and infected populations, and is given as

$$\beta x_{\mathtt{S}}(t)\frac{x_{\mathtt{I}}(t)}{N}$$

where β is the infection rate. The term $x_{\mathtt{I}}(t)/N$ is the proportion of undetected infected population at any time t in a homogeneous population structure. Note that in light of Assumption 7.1(ii), diagnosed population $x_{\mathtt{D}}(t)$ does not participate in the infection transmission because they are either quarantined and/or hospitalized, i.e., they are temporarily isolated from the population. Finally, by Assumption 7.1(iv), there is no inflow to the susceptible compartment.

Detection The outflow from the infected compartment (I) is either due to detection (i.e., transfer to the diagnosed compartment (D)) or recovery without detection (i.e., transfer to the unidentified recovered compartment (U)). The first outflow is due to the testing rate $u(t)$. Since the probability of detecting an infected person from a testable population by a single test is $x_{\mathtt{I}}(t)/x_{\mathtt{T}}(t)$, therefore we have

$$u(t)\frac{x_{\mathtt{I}}(t)}{x_{\mathtt{T}}(t)}$$

as the rate of detection of the infected population in I compartment.

Recovery The second outflow $\gamma x_{\mathtt{I}}(t)$ from the I compartment consists of those people who recover naturally with a recovery rate γ without getting diagnosed, where $1/\gamma$ is the average recovery period. The unidentified recovered compartment (U) accumulates the infected people who recover naturally without being detected.

Removal The diagnosed compartment (D) admits $u(t)x_{\mathtt{I}}(t)/x_{\mathtt{T}}(t)$ as an inflow, whereas the outflow is $\rho x_{\mathtt{D}}(t)$ with ρ being the removal rate. That is, $1/\rho$ is the average time period after which a typical diagnosed person either recovers or dies. The removed compartment (R) accumulates the diagnosed people who die or recover with a removal rate ρ.

7.1.4 Output signals from the model

The data reported by the health authorities are chosen as the measured outputs of the SIDUR model. They are five measurements, related directly to the states are as follows:

- Cumulative number of diagnosed people

$$y_1(t) = x_{\mathtt{D}}(t) + x_{\mathtt{R}}(t) \tag{7.3}$$

- Cumulative number of removed people

$$y_2(t) = x_{\mathtt{R}}(t) \tag{7.4}$$

- Number of positively tested people (or positive test results) per day

$$y_3(t) = u(t)\frac{x_{\mathtt{I}}(t)}{x_{\mathtt{T}}(t)} \tag{7.5}$$

- Number of active Intensive Care Unit (ICU) cases (or the number of ICU beds currently occupied by the diagnosed infected):

$$B(t) := y_4(t) = g(A(t-\psi)) \tag{7.6}$$

  where $A(t) = x_{\mathtt{I}}(t) + x_{\mathtt{D}}(t)$ is the number of active infected cases, ψ is the average time period a typical ICU case takes from getting infected to being admitted to ICU, and the function g is to be chosen to fit the data $\overline{B}(k) := \overline{y}_4(k)$.

- Cumulative number of deaths due to the epidemic (or extinct cases):

$$E(t) := y_5(t) = h(I(t-\phi)) \tag{7.7}$$

  where $I(t) = N - x_{\mathtt{S}}(t)$ is the cumulative number of infected cases, ϕ is the average time period a typical extinct case takes from getting infected to death, and the function h is to be chosen to fit the data $\overline{E}(k) := \overline{y}_5(k)$.

The model outputs (7.3), (7.4), and (7.5) are fitted with the data outputs $\overline{y}_1(t)$, $\overline{y}_2(t)$, and $\overline{y}_3(t)$, respectively, in order to estimate the model parameters $\beta, \theta, \gamma, \rho$ in section 7.3 for the COVID-19 case of France. Note that these model outputs are related to each other. Since the number of diagnosed infected people at any time t can be obtained as $x_{\mathtt{D}}(t) = y_1(t) - y_2(t)$, which is also known as the number of active diagnosed cases, we obtain the following relation between $y_1(t)$ and $y_2(t)$ from (7.1e)

$$\dot{y}_2(t) = \rho(y_1(t) - y_2(t)). \tag{7.8}$$

On the other hand, the number of positive test results per day $y_3(t)$ is related to the cumulative number of diagnosed cases $y_1(t)$ by the following relation

$$y_3(t) = \dot{x}_{\mathtt{D}}(t) + \dot{x}_{\mathtt{R}}(t) = \dot{y}_1(t). \tag{7.9}$$

The cumulative number of diagnosed people $y_1(t)$ can be obtained by integrating the daily number of positive test results as

$$y_1(t) - y_1(0) = \int_0^t y_3(\eta)d\eta. \tag{7.10}$$

These output relations (7.8), (7.9), (7.10) are used to infer the missing data from the available data in section 7.2.

Second, the model outputs (7.6) and (7.7) depend on the functions g and h, which are assumed to be polynomials that fit the available COVID-19 data of France on the number of active ICU cases and the cumulative number of deaths, respectively, in section 7.3. These model outputs will be used as performance outputs to evaluate the testing policy proposed for epidemic suppression in section 7.4.

7.1.5 Basic and effective reproduction numbers

An important quantity to assess the epidemic potential of a disease is the *basic reproduction number* R_0, which is defined as the expected number of secondary infected cases produced by a single infected person in a completely susceptible population [Hethcote2000]. If $R_0 > 1$, then each generation of infected cases produces more secondary cases in the next generation and the disease has a potential of becoming an epidemic. If $R_0 < 1$, then each generation of infected cases produces less secondary cases in the next generation and the disease will eventually die out. It is worth noticing, however, that the definition of R_0 assumes that the people around a primary infected case are all susceptible. This suggests that determining R_0 is important only at the onset of an epidemic. However, in the later stages, more people get infected and not all people around an infected person are necessarily susceptible. As more people get infected, the conditions favoring the disease to propagate change and the number of susceptible people that an infected person infects is actually less than that what R_0 predicts. Thus, a more suitable quantity during the later stages of the epidemic is the *effective reproduction number* R_t, which takes into account the proportion of susceptible people in the total population [Rothman2008].

For the SIDUR model, the effective reproduction number R_t is the ratio of the inflow and the outflow of the undiagnosed infected compartment (I). In other words, R_t is the ratio of the number of newly infected people and the number of newly diagnosed and recovered people at time t. If $R_t < 1$, this means that more people are being diagnosed and recovered than the people being infected at time t, which implies that $x_{\mathtt{I}}(t)$ will decrease. If $R_t > 1$, this means that more people are being infected than the people being diagnosed and recovered at time t, which implies that $x_{\mathtt{I}}(t)$ will increase. Notice that the definition of the basic reproduction number R_0 is same as R_t for $t = 0$.

To derive the expression of the effective reproduction number R_t, we consider the model equation (7.1b), where the undiagnosed infected population $x_{\mathtt{I}}(t)$ satisfies

$$\dot{x}_{\mathtt{I}}(t) = \beta x_{\mathtt{S}}(t)\frac{x_{\mathtt{I}}(t)}{N} - u(t)\frac{x_{\mathtt{I}}(t)}{x_{\mathtt{T}}(t)} - \gamma x_{\mathtt{I}}(t).$$

The positive rate (inflow) $\beta x_{\mathtt{S}}(t)x_{\mathtt{I}}(t)/N$ tells how many new infections will be generated in the next moment, and the negative rates (outflows) $u(t)x_{\mathtt{I}}(t)/x_{\mathtt{T}}(t)$ and $\gamma x_{\mathtt{I}}(t)$ tell how many infected people will be diagnosed or recovered in the next moment, respectively. Therefore, the effective reproduction number is given by

$$R_t = \frac{\beta x_{\mathtt{S}}(t)\frac{x_{\mathtt{I}}(t)}{N}}{u(t)\frac{x_{\mathtt{I}}(t)}{x_{\mathtt{T}}(t)} + \gamma x_{\mathtt{I}}(t)} = \frac{\beta}{\frac{u(t)}{x_{\mathtt{T}}(t)} + \gamma}\frac{x_{\mathtt{S}}(t)}{N}. \tag{7.11}$$

To derive the expression of the basic reproduction number R_0, we consider the expression of R_t at $t = 0$, which corresponds to the onset of the epidemic. Furthermore, for $t = 0$, one can assume few infected cases so that $x_{\mathtt{S}}(0) \approx N$ and $x_{\mathtt{T}}(0) \approx (1-\theta)N$. Under these approximations, the basic reproduction number is given by

$$R_0 = \frac{\beta}{\frac{u(0)}{(1-\theta)N} + \gamma}. \tag{7.12}$$

This expression can also be obtained by following the methodology of [Van den Driessche2008]. Notice that the basic reproduction number R_0 depends on the initial testing policy $u(0)$. This indicates that it is possible to suppress the epidemic in the beginning by having an intensive testing policy, which can be seen, for example, in the case of South Korea [Oh2020]. In general, however, we have $u(0) \approx 0$ that implies $R_0 \approx \beta/\gamma$, which is same as the basic reproduction number of SIR epidemic model.

7.2 Data for COVID-19 case of France

The data related to COVID-19 in France is collected from the French government's platform for publicly available data[1] for the time period of January 24 to July 01, 2020. In particular, we use datasets provided by the French Ministry of Social Affairs and Health (Ministère des Solidarités et de la Santé (MSS)) and the French Public Health Agency (Santé Publique France (SPF)). From MSS, we obtain the data about different categories of people affected by COVID-19, i.e., diagnosed, hospitalized, recovered from hospitals, and dead. From SPF, we obtain the data for the number of PCR tests performed and positive test results obtained per day.

The data obtained from both sources is incomplete in several aspects. For instance, the data for the number of recovered people does not record those who were not hospitalized and recovered from their homes after being diagnosed. These people are not hospitalized because of mild symptoms of the disease, but are quarantined in their homes for a certain number of days. Only those who are hospitalized after being diagnosed are recorded as recovered when they are discharged from the hospitals. On the other hand, the data for COVID-19 PCR tests is also incomplete. To illustrate this, we consider three intervals of time: (1) January 24 to March 09, 2020, (2) March 10 to May 12, 2020, and (3) May 13 to July 01, 2020. There is no data available for the tests during the first interval. During the second interval, the testing data is collected only from the medical laboratories and not from the hospitals. However, we have reliable data only during the third interval which is collected both from the medical laboratories and the hospitals. Therefore, the data obtained from the above sources can be considered as a raw data which needs to be imputed. In what follows, we first present the raw data and then detail the procedure for data imputation. The imputed data is obtained by making reasonable assumptions in order to infer the missing data from the raw data.

7.2.1 Raw data

This subsection illustrates the data obtained from MSS and SPF without any modification.

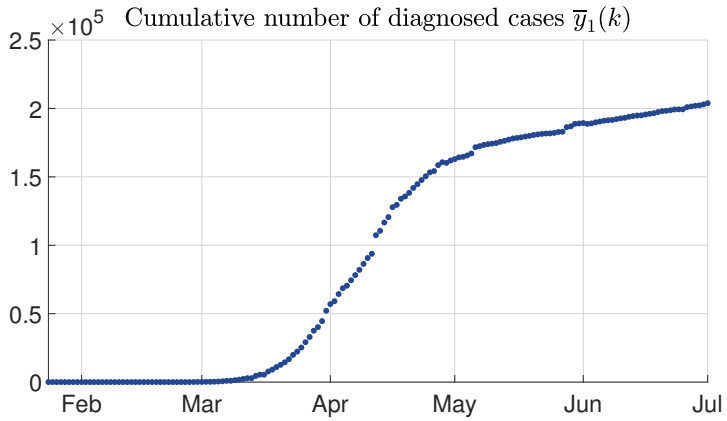


Figure 7.2: Cumulative number of diagnosed cases $\overline{y}_1(k)$ from January 24 to July 01, 2020. Source: MSS.

Cumulative number of diagnosed cases We denote the data for the cumulative number of diagnosed cases by $\overline{y}_1(k)$, which is illustrated in Figure 7.2 and corresponds to

[1] Open platform for French public data (data.gouv.fr)

the model output $y_1(t)$ in (7.3). It is also known as the total "confirmed" cases. This is a cumulative data for all the cases diagnosed with the disease through RT-PCR tests[2]. Thus, it includes both the active cases (those who are either admitted to the hospitals and/or quarantined) and the inactive cases (those who either recovered or died after being diagnosed). That is, $\overline{y}_1(k)$ corresponds to the sum of people in the diagnosed (D) and removed (R) compartments of the SIDUR model (Figure 7.1) on a given day k, as given in (7.3).

There is also an additional data for the diagnosed cases from French retirement homes (EHPAD). However, the French government database[3] and several other international databases[4] [5] do not add the diagnosed cases from EHPAD to the cumulative number of diagnosed (confirmed) cases. That is, the data for the cumulative number of diagnosed cases is considered to be inclusive of the diagnosed cases from EHPAD. However, in all the above databases, the data on cumulative number of deaths is collected separately from both the hospitals and EHPAD.

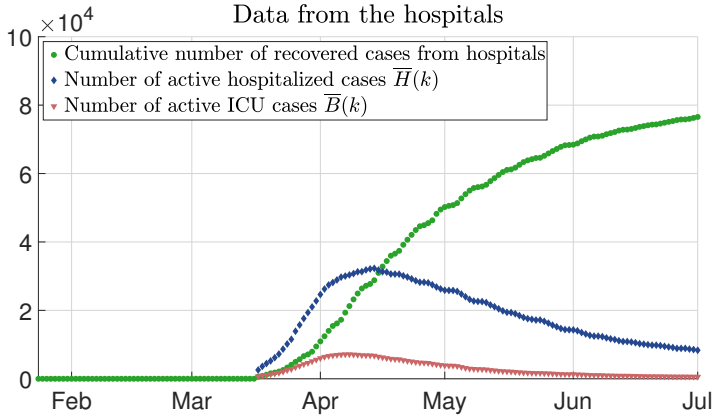


Figure 7.3: Total number of recovered cases who returned home after hospitalization from January 24 to July 01, 2020. The number of active COVID-19 hospitalized cases $\overline{H}(k)$ and ICU cases $\overline{B}(k)$ from March 17 to July 01, 2020. Source: MSS.

Number of active hospitalized and ICU cases The data on the number of active hospitalized cases is denoted as $\overline{H}(k)$ and is illustrated in Figure 7.3 along with the number of active ICU cases denoted as $\overline{B}(k)$. This data corresponds to the number of people who are admitted to the hospitals and/or ICU on a given day. That is, it is not a cumulative data. Moreover, it doesn't include those who were diagnosed but not hospitalized. That is, this data corresponds to a certain proportion of people in the diagnosed compartment (D) of the SIDUR model. This data is available from March 17, 2020, onward.

Cumulative number of recovered cases from hospitals This data is also illustrated in Figure 7.3. It corresponds to people who, after recovering from the disease, were discharged from the hospitals. Obviously, prior to recovering, they were diagnosed with the disease and hospitalized due to having severe symptoms.

[2] Santé Publique France
[3] Open platform for French public data (data.gouv.fr)
[4] European Centre for Disease Prevention and Control
[5] Worldometers.info

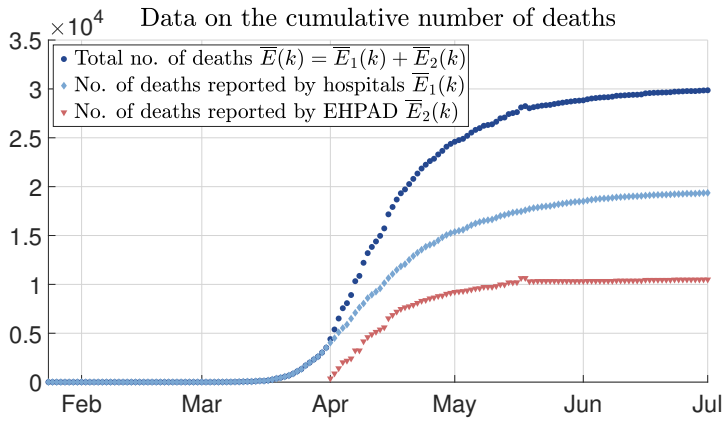


Figure 7.4: Total number of deaths from COVID-19 reported (a) by hospitals from January 24 to July 01, 2020, and (b) by retirement homes (EHPAD) from April 01 to July 01, 2020. Source: MSS.

Cumulative number of deaths As mentioned before, the data on the cumulative number of diagnosed cases is considered to be inclusive of the diagnosed cases from the French retirement homes (EHPAD). However, the case for data on the cumulative number of deaths is different. Those who died at the hospitals and those who died in the retirement homes (EHPAD) are considered to be distinct in the databases. Thus, the cumulative number of deaths is the sum of both data, which are illustrated in Figure 7.4.

Number of tests and positive tests per day We have two types of data related to RT-PCR tests for COVID-19. The first type of data is collected by SPF on the number of tests performed and positive test results per day from March 10 to May 26, 2020. However, this data is collected only from the central sampling laboratories: Eurofins Biomnis and Cerba. Figure 7.5(a) (blue) illustrates the number of tests performed per day and Figure 7.5(b) (blue) illustrates the number of positive test results per day.

The second type of data was made available after the deployment of a new information screening system (SI-DEP) by the SPF. This data is available from May 13, 2020, onward. It is collected from both the laboratories and the hospitals. However, the data reported by SI-DEP is the number of 'tested people' per day instead of the number of 'tests performed' per day. SI-DEP guarantees that only one test is counted per person. In the case of, for instance, multiple negative test results for a certain person, SI-DEP considers only the first date on which the PCR test was performed. Later, if that person gets a positive test result, then only this new result is reported in the data and the previous data is erased. Figure 7.5(a) (red) illustrates the number of tested people per day and Figure 7.5(b) (red) illustrates the number of positively tested people per day.

7.2.2 Imputed data

In the raw data, we only have the data for those who recover or die in the hospitals after being diagnosed with the disease. However, the removed compartment of the SIDUR model also comprises the diagnosed cases who were not hospitalized but were quarantined in their homes. There is no data that records the recovery of these people. Moreover, the data on PCR tests is also incomplete; there is no data from January 24 to March 09, 2020, and the data from March 10 to May 12, 2020, doesn't include the tests performed in the hospitals. Therefore, in order to infer the missing data, we impute the raw data by making reasonable assumptions.

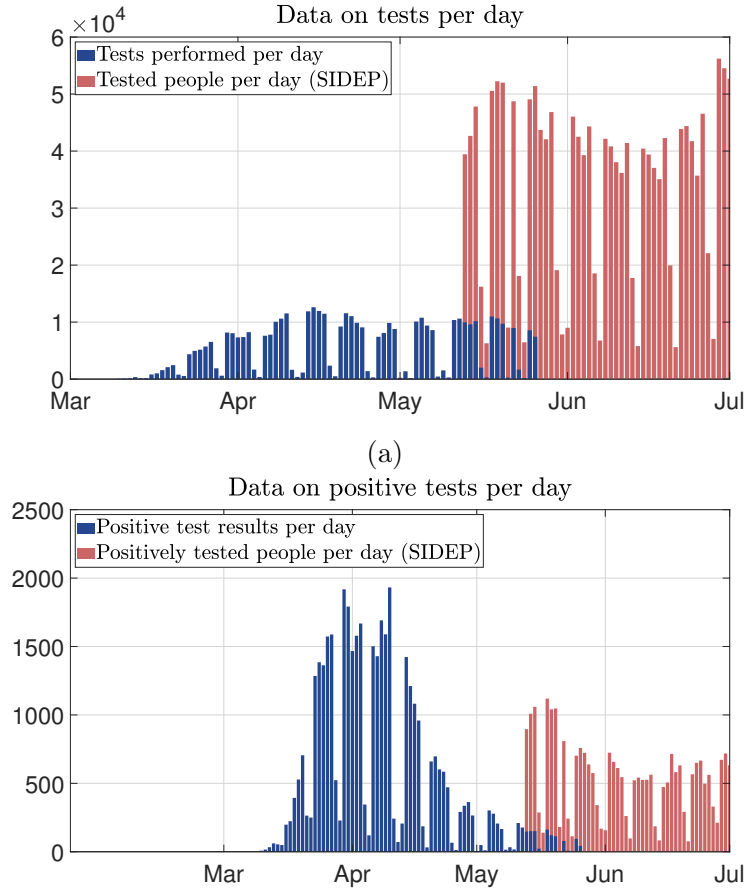


Figure 7.5: Data on the PCR tests: (a) The number of tests performed per day from March 10 to May 26 and the number of tested people per day from May 13 to July 01; (b) The number of positive test results per day from March 10 to May 26 and the number of positively tested people per day from May 13 to July 01. Source: SPF.

Cumulative number of removed cases

From the data on the total number of recovered people from hospitals shown in Figure 7.3, we see that 76,540 people have recovered from the hospitals as of July 01, 2020. If we subtract this number and the total number of deaths (Figure 7.4), i.e., 29,860, from the total number of diagnosed cases (Figure 7.2), i.e., 165,700, we obtain $165,700-76,540-29,860 = 59,300$ people. Further subtracting the currently hospitalized cases (Figure 7.3), i.e., 8336 as of July 01, we obtain $59,300 - 8336 = 50,964$ people, who have an unknown status. These people were diagnosed but were not hospitalized; they were quarantined in their homes, and it is not known whether they have recovered or are still infected. There is no data that provides a correct answer for how many people among them have recovered and how many of them are still infected. Therefore, using the relevant raw data, we infer the cumulative number of removed cases $\overline{y}_2(k)$ by estimating the number of diagnosed cases who recovered from home.

We use the following notations for simplicity and brevity:

$\overline{D}_h(k)$	Total number of diagnosed who are hospitalized
$\overline{D}_q(k)$	Total number of diagnosed who are quarantined at home
$\overline{R}_h(k)$	Total number of recovered or dead from a hospital
$\overline{R}_q(k)$	Total number of recovered from quarantine at home

where, by Assumption 7.1(iii), all deaths are reported and, hence, not included in $\overline{R}_q(k)$.

By definition, we have

$$\begin{aligned} \overline{y}_1(k) &= \overline{D}_h(k) + \overline{D}_q(k) \\ \overline{y}_2(k) &= \overline{R}_h(k) + \overline{R}_q(k) \end{aligned} \tag{7.13}$$

where k is from January 24 to July 01, 2020. Note that $\overline{y}_1(k)$ is illustrated in Figure 7.2 and

$$\overline{D}_h(k) = \overline{R}_h(k) + \overline{H}(k),$$

where $\overline{R}_h(k)$ is the sum of the total number of recovered cases from hospitals (Figure 7.3) and the total number of deaths (Figure 7.4), and $\overline{H}$ is the number of active hospitalized cases (Figure 7.3). Thus, we can compute the total diagnosed cases who were not hospitalized as

$$\overline{D}_q(k) = \overline{y}_1(k) - \overline{D}_h(k).$$

Since there is no data for the diagnosed people who recovered from home, therefore $\overline{R}_q(k)$ is unknown. Thus, we assume the following:

$$\frac{\overline{R}_q(k)}{\overline{D}_q(k)} = \frac{\overline{R}_h(k)}{\overline{D}_h(k)}. \tag{7.14}$$

That is, the ratio of the diagnosed cases who recovered in quarantine at homes to the total diagnosed cases who were quarantined at homes is equal to the ratio of the diagnosed cases who recovered or died at hospitals to the total diagnosed cases who were hospitalized. In other words, we assume that the removal rate of people who were not hospitalized is equal to the removal rate of people who were hospitalized. Thus, from (7.13) and (7.14), we obtain

$$\overline{y}_2(k) = \overline{R}_h(k)\left(1 + \frac{\overline{D}_q(k)}{\overline{D}_h(k)}\right) = \frac{\overline{R}_h(k)}{\overline{D}_h(k)}\overline{y}_1(k).$$

which corresponds to the data on cumulative number of removed cases $y_2(t)$ defined in (7.4).

Combining two types of testing data

From Figure 7.5, we see that the first type of testing data, which is available from March 10 to May 26, 2020, considers the number of tests performed and positive test results per day. On the other hand, the second type of testing data, which is available from May 13, 2020, onward, considers the number of tested people and positively tested people per day. However, no person is usually tested more than once in a single day. Therefore, we assume that *the number of tested people per day is same as the number of tests performed per day.* Similarly, *the number of positively tested people per day is same as the number of positive test results per day.* Note that if a person is tested more than once but on different days, then this assumption is not violated.

We consider three time intervals: (i) January 24–March 09, when the data on PCR tests in France is not available; (ii) March 10–May 12, when the data is available but

incomplete; (iii) May 13–July 01, when the complete data is available. Let $\overline{u}(k)$ and $\overline{y}_3(k)$ denote the number of tests performed and the number of positive test results per day, respectively, for the entire interval January 24 to July 01, 2020. Let $\overline{u}'(k), \overline{u}''(k), \overline{u}'''(k)$ and $\overline{y}_3'(k), \overline{y}_3''(k), \overline{y}_3'''(k)$ be the number of tests performed and the number of positive test results obtained for the first, second, and third time intervals, respectively. Since the data in the third interval is reliable, we do not make any imputations for $\overline{u}'''(k)$ and $\overline{y}_3'''(k)$. For the other two intervals, we make reasonable assumptions to complete the data.

(i) *January 24–March 09:* This time interval corresponds to the beginning of the epidemic in France and the data for tests performed and positive test results per day for this interval is $\overline{u}'(k)$ and $\overline{y}_3'(k)$, respectively, which is not available and is inferred from the other data. Our key observation is that only those people were tested during this time interval who showed symptoms. Therefore, we assume that *the number of tests performed per day is approximately equal to the number of positive test results obtained per day during January 24 and March 09.* Moreover, *the number of positive test results obtained per day is equal to the number of diagnosed cases that day.* Thus, from the output relation (7.9), we have

$$\overline{u}'(k) \approx \overline{y}_3'(k)$$

and

$$\overline{y}_3'(k) = \overline{y}_1(k+1) - \overline{y}_1(k)$$

where $k = 0, 1, \ldots, 46$ are the days from January 24 to March 09.

(ii) *March 10–May 12:* In the second interval, we have the data on PCR tests that is reported only by the laboratories and not by the hospitals. During this interval, we compute the data as follows: $\overline{u}''(k)$ is same as the data (Figure 7.5) and $\overline{y}_3''(k) = \overline{y}_1(k+1) - \overline{y}_1(k)$, where $k = 47, \ldots, 110$ are the days from March 10 to May 12.

(iii) *May 13–July 01:* In this third interval, we have a reliable data $\overline{u}'''(k)$ and $\overline{y}_3'''(k)$ from SPF as shown in Figure 7.5, where $k = 111, \ldots, 160$ are the days from May 13 to July 01.

Based on the above data imputations, we obtain the number of tests performed per day

$$\overline{u}(k) = \begin{cases} \overline{u}'(k) & \text{if } k \in \{0, 1, \ldots, 46\} \\ \overline{u}''(k) & \text{if } k \in \{47, \ldots, 110\} \\ \overline{u}'''(k) & \text{if } k \in \{111, \ldots, 160\} \end{cases}$$

which corresponds to the control input $u(t)$, and the number of positive tests obtained per day

$$\overline{y}_3(k) = \begin{cases} \overline{y}_3'(k) & \text{if } k \in \{1, \ldots, 46\} \\ \overline{y}_3''(k) & \text{if } k \in \{47, \ldots, 110\} \\ \overline{y}_3'''(k) & \text{if } k \in \{111, \ldots, 160\} \end{cases}$$

which corresponds to the model output $y_3(t)$ in (7.5), where $k = 1, \ldots, 160$ are the days of the complete time interval from January 24 to July 01, 2020.

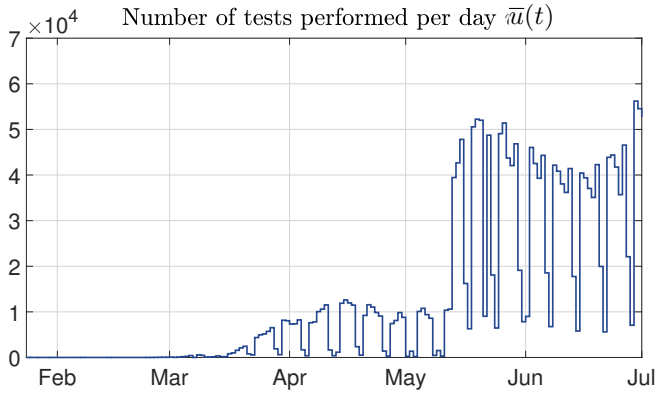


Figure 7.6: Input signal from the data.

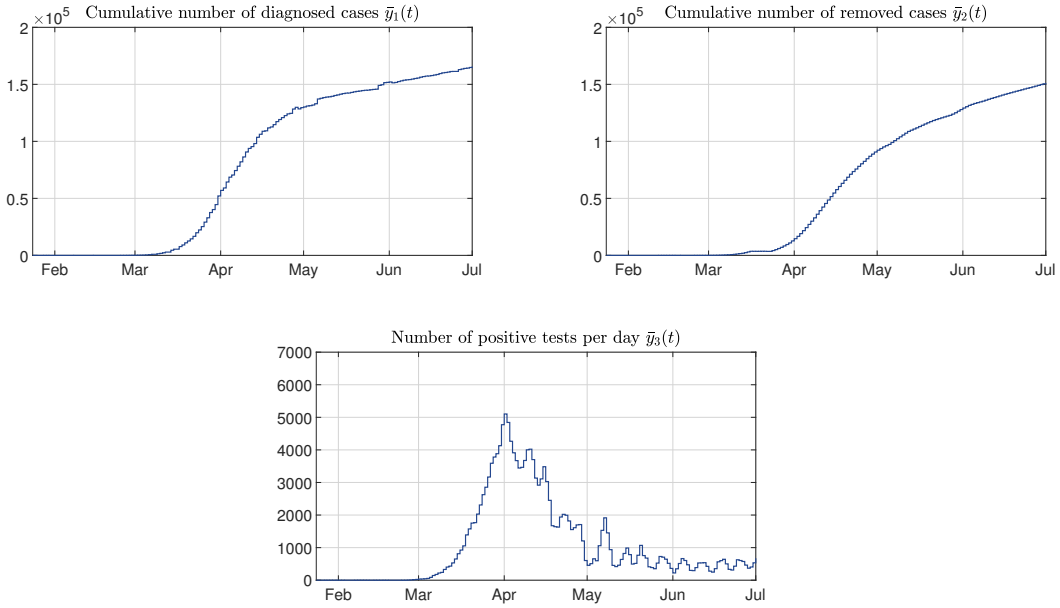


Figure 7.7: Output signals from the data.

7.2.3 Input and output signals from the data

For $k = 0, 1, 2, \ldots, 160$, we define a continuous-time input signal from the data as

$$\overline{u}(t) = \overline{u}(k), \quad \text{for } \lfloor t \rfloor \leq k < \lceil t \rceil$$

which is illustrated in Figure 7.6. We denote by $\overline{y}_i(t)$, for $i = 1, 2, 3$, the outputs obtained from the data. That is, the output signals from the data and the model are related as follows

$$\begin{aligned} \overline{y}_1(t) &= y_1(t) + w_1(t) \\ \overline{y}_2(t) &= y_2(t) + w_2(t) \\ \overline{y}_3(t) &= y_3(t) + w_3(t) \end{aligned}$$

where $w_i(k)$, for $i = 1, 2, 3$, represents the measurement noise. The outputs $\overline{y}_1(t)$, $\overline{y}_2(t)$, and $\overline{y}_3(t)$ correspond to the cumulative number of diagnosed cases, the cumulative number of removed cases, and the number of positive test results per day, respectively. Similar to

the input $u(t)$, we define continuous-time output signals from the data as

$$\begin{cases} \overline{y}_1(t) &= \overline{y}_1(k), \quad \text{for } \lfloor t \rfloor \leq k < \lceil t \rceil \\ \overline{y}_2(t) &= \overline{y}_2(k), \quad \text{for } \lfloor t \rfloor \leq k < \lceil t \rceil \\ \overline{y}_3(t) &= \overline{y}_3(k), \quad \text{for } \lfloor t \rfloor \leq k < \lceil t \rceil \end{cases}$$

which are illustrated in Figure 7.7.

7.3 Estimation of the Model Parameters

In this section, we validate the SIDUR model by estimating the model parameters ρ, β, θ, and γ for the COVID-19 case of France.

7.3.1 Estimation of the removal rate

The removal rate ρ can be directly estimated from the data outputs $\overline{y}_1(k)$ and $\overline{y}_2(k)$. Consider a daily sampling of the model equation (7.1e), which leads to

$$\Delta x_{\mathtt{R}}(k) \approx \rho x_{\mathtt{D}}(k)$$

where Δ stands for the forward difference operator, i.e., $\Delta x_{\mathtt{R}}(k) = x_{\mathtt{R}}(k+1) - x_{\mathtt{R}}(k)$ for $k \in \mathbb{N}$. Therefore, from the relation between $y_1(t)$ and $y_2(t)$ in (7.8), we obtain

$$\Delta \overline{y}_2(k) = \rho\, \overline{y}_{12}(k) + e(k) \tag{7.15}$$

where $\overline{y}_{12}(k) = \overline{y}_1(k) - \overline{y}_2(k)$ and $e(k)$ is the error term due to measurement noise. Then, the problem of estimating ρ can be formulated as follows: Find ρ^* such that

$$\rho^* = \arg \min_{\rho \in [0,1]} \sum_{k=1}^{\tau} \|\Delta \overline{y}_2(k) - \rho\, \overline{y}_{12}(k)\|^2. \tag{7.16}$$

The solution of this problem is obtained through least-square estimation [Ljung1999, Chapter 7].

7.3.2 Estimation of the infection rate, the testing specificity parameter, and the recovery rate

We formulate a problem of fitting the model outputs $y_1(t), y_2(t), y_3(t)$ to the data outputs $\overline{y}_1(k), \overline{y}_2(k), \overline{y}_3(k)$, where $k = 0, 2, \ldots, 160$ represents the days from January 24 to July 01. The model fitting is done by optimizing the parameters β, θ, γ for the time interval $[0, 160]$ under the assumption that γ is constant whereas β and θ are piecewise constants.

To limit the rate of spread of COVID-19, the French government announced to place a lockdown all over France from March 17 to May 10, 2020, which included restricted human mobility, strict social distancing measures, and closure of schools, offices, and marketplaces. However, the essential services and public establishments were authorized to remain open under strict preventive measures. People were allowed to leave their homes with face masks only for necessary groceries, brief exercise within a certain radius of their homes, or for urgent medical reasons. Such an intervention from the public authority is necessary to mitigate the rate of spread of the disease and to reduce the value of infection rate

β. Therefore, in relation to the case of France, we divide the time into three intervals: (i) Before lockdown (January 24 to March 16), i.e., $k = 0, 1, \ldots, 53$, (ii) During lockdown (March 17 to May 10), i.e., $k = 54, \ldots, 108$, and (iii) After lockdown (May 11 to July 01), i.e., $k = 109, \ldots, 160$. We consider a different value of the infection rate β during each of these intervals, i.e.,

$$\beta(k) = \begin{cases} \beta_1, & \text{for } k \in \{0, 1, \ldots, 53\} \\ \beta_2, & \text{for } k \in \{54, \ldots, 108\} \\ \beta_3, & \text{for } k \in \{109, \ldots, 160\} \end{cases}$$

where $\beta_1, \beta_2, \beta_3$ are non-negative real numbers.

For the testing specificity parameter θ, on the other hand, we divide the time into two intervals: (i) January 24 to May 10, i.e., $k = 0, 1, \ldots, 108$, and (ii) May 11 to July 01, i.e., $k = 109, \ldots, 160$, where May 11 corresponds to the change in testing policy in France (see [Hale2020b] and the website of Our-World-in-Data[6]). Thus, we have

$$\theta(k) = \begin{cases} \theta_1, & \text{for } k \in \{0, 1, \ldots, 108\} \\ \theta_2, & \text{for } k \in \{109, \ldots, 160\} \end{cases}$$

where θ_1 and θ_2 are real numbers in the interval $[0, 1]$.

Let $\mathbf{p} = [\ \beta_1 \quad \beta_2 \quad \beta_3 \quad \theta_1 \quad \theta_2 \quad \gamma\ ]^\intercal$ be the parameter vector. Then, the goal is to find $\mathbf{p}^*$ such that

$$\mathbf{p}^* = \arg\min_{\mathbf{p}} \mathcal{J}(\mathbf{p}) \tag{7.17}$$

where the cost function is given by

$$\mathcal{J}(\mathbf{p}) = \sum_{k=0}^{160} \left[(y_1(k, \mathbf{p}) - \overline{y}_1(k))^2 + (y_2(k, \mathbf{p}) - \overline{y}_2(k))^2 + (y_3(k, \mathbf{p}) - \overline{y}_3(k))^2 \right] \tag{7.18}$$

with the model outputs y_i, $i = 1, 2, 3$, depending on the parameter vector $\mathbf{p}$.

To solve this problem, one can also pose it as a least-square estimation, as we did for the removal rate ρ, by defining relations between the data outputs $\overline{y}_1(k), \overline{y}_2(k), \overline{y}_3(k)$. However, such relations include the difference operator Δ applied twice to the data outputs, which is usually not recommended when the data is noisy because it amplifies the measurement noise. Moreover, the gradient-based estimation algorithms [Nelles2001, Chapter 4] are also not suitable due to the difficulty of computing the gradient of the cost function $\mathcal{J}(\mathbf{p})$ online with respect to the parameter vector p. This is because the model outputs $y_1(t), y_2(t), y_3(t)$ do not depend directly on the parameters but through the solution trajectories of the SIDUR model. For simplicity, therefore, we choose the particle swarm optimization (PSO) [Kennedy1995], which is a 'derivative-free' algorithm, to estimate the parameter vector $\mathbf{p}$.

Particle swarm optimization algorithm

We briefly describe the particle swarm optimization (PSO) algorithm, [Kennedy1995], which considers a foraging swarm of n particles who collectively search for an optimal solution of (7.17) in the parameter space. At time step $h = 0, 1, 2, \ldots$, each particle i visits a position $\hat{\mathbf{p}}_h^i$ by moving with velocity $\mathbf{v}_h^i$. Initially, when $h = 0$, the positions $\hat{\mathbf{p}}_0^i$, for all $i \in \{1, \ldots, n\}$, are chosen randomly in the parameter space and the velocities $\mathbf{v}_0^i = \mathbf{0}$. Each particle i stores its personal best pair $(\hat{\mathbf{p}}_h^{i*}, J_h^{i*})$ and the social best pair (s_h^*, J_h^{s*})

[6] https://ourworldindata.org/grapher/covid-19-testing-policy

in memory, where $J_h^{i*} = \mathcal{J}(\hat{\mathbf{p}}_h^{i*})$ and $J_h^{s*} = \mathcal{J}(s_h^*)$ are the costs (7.18) of personal best position $\hat{\mathbf{p}}_h^{i*}$ and social best position $s_h^* = \arg\min_{\hat{\mathbf{p}}_h^{i*}, i\in\{1,\dots,n\}} \mathcal{J}(\hat{\mathbf{p}}_h^{i*})$, respectively. Notice that $J_h^{s*} \leq J_h^{i*}$ for all $i \in \{1,\dots,n\}$. The personal best pair of a particle corresponds to the best position in the parameter space it has visited so far. The social best pair, on the other hand, corresponds to the best position in the parameter space that anyone in the swarm has visited so far.

At every time step, each particle updates its velocity, position, its personal best pair, and the social best pair. The velocity and position are updated as follows:

$$\begin{aligned} \mathbf{v}_{h+1}^i &= w\mathbf{v}_h^i + c_1 r_{h,1}(\hat{\mathbf{p}}_h^{i*} - \hat{\mathbf{p}}_h^i) + c_2 r_{h,2}(s_h^* - \hat{\mathbf{p}}_h^i) \\ \hat{\mathbf{p}}_{h+1}^i &= \hat{\mathbf{p}}_h^i + \mathbf{v}_{h+1}^i \end{aligned}$$

where w is the inertia weight, c_1, c_2 are the acceleration coefficients, and $r_{h,1}, r_{h,2}$ are uniformly distributed random numbers in $[0,1]$ generated at each time step h. There are many ways of choosing these parameters [Clerc2002, Poli2007, Zhan2009].

Each particle i computes the cost $J_{h+1}^i = \mathcal{J}(\hat{\mathbf{p}}_{h+1}^i)$ at its current position and updates its personal best pair as

$$(\hat{\mathbf{p}}_{h+1}^{i*}, J_{h+1}^{i*}) = \begin{cases} (\hat{\mathbf{p}}_{h+1}^i, J_{h+1}^i), & \text{if } J_{h+1}^i \leq J_h^{i*} \\ (\hat{\mathbf{p}}_h^{i*}, J_h^{i*}), & \text{otherwise.} \end{cases}$$

Each particle i then communicates its personal best pair with all the other particles and each of them finds the social best pair for time $h+1$ as

$$(s_{h+1}, J_{h+1}^s) = (\hat{\mathbf{p}}_{h+1}^b, J_{h+1}^b)$$

where $b = \arg\min_{j\in\{1,\dots,n\}} J_{h+1}^j$. Finally, the social best pair is updated as

$$(s_{h+1}^*, J_{h+1}^{s*}) = \begin{cases} (s_{h+1}, J_{h+1}^s), & \text{if } J_{h+1}^s \leq J_h^{s*} \\ (s_h^*, J_h^{s*}), & \text{otherwise.} \end{cases}$$

7.3.3 Estimated parameter values

Infection rate	$\beta_1 = 0.3708 \quad \beta_2 = 0.0707 \quad \beta_3 = 0.3717$
Testing specificity	$\theta_1 = 0.9948 \quad \theta_2 = 0.9967$
Recovery rate	$\gamma = 0.1589$
Removal rate	$\rho = 0.0499$

Table 7.1: Estimated parameters of SIDUR model for the COVID-19 case of France.

The estimated parameter values obtained by solving (7.16) and (7.17) are provided in Table 7.1. The estimated recovery rate γ and removal rate ρ show that an undiagnosed infected person recovers on average in about $1/\gamma \approx 6.3$ days and a diagnosed person recovers or dies on average in about $1/\rho \approx 20$ days. The testing specificity parameter changes slightly from $\theta_1 = 0.9948$ to $\theta_2 = 0.9967$, which can have significant impact on the positive test results because it multiplies with the sum of the susceptible and unidentified recovered population in (7.2) that is in the order of 10^7 in the case of population of France.

The infection rate β changes its value twice. First, it drops from $\beta_1 = 0.3708$ to $\beta_2 = 0.0707$ when the lockdown is implemented in France on March 17, which significantly decreased the rate of the epidemic spread. Then, it rises from $\beta_2 = 0.0707$ to $\beta_3 = 0.3717$ when the lockdown is lifted on May 10. Many restrictions like social distancing and wearing of face masks were still in place after May 10 in order to prevent the spread of COVID-19 in France. However, the increase in the value of β can be explained by the summer vacations when people were allowed to travel everywhere across France and Europe[7]. This made the places with tourist attractions very crowded and resulted in a higher infection rate.

7.3.4 Model validation

Using the estimated values of the model parameters in Table 7.1, we run the model from January 24 to July 01, 2020. The model fits the output signals from data as shown in Figure 7.8.

The basic reproduction number R_0 at the outbreak of the COVID-19 epidemic in France is computed using (7.12) and the average value of effective reproduction number R_t during the three phases (before, during, and after lockdown) are computed using (7.11). For the 'after lockdown' phase, we chose July 01, 2020, to compute R_t because it is the date up to which our data is considered. These computed values are shown in Table 7.2 with a comparison to the ones reported by the government[8] [9].

Epidemic phases	Computed from model	Reported by French government
Outbreak	$R_0 = 2.33$	$R_0 = 2.7$
Before lockdown	$R_t = 2.3$	$R_t = 2.7$
During lockdown	$R_t = 0.33$	$R_t = 0.7$
After lockdown	$R_t = 1$	$R_t = 1$

Table 7.2: The basic reproduction number R_0 at the outbreak and the values of the effective reproduction number R_t at the end of each phase of the COVID-19 epidemic in France.

The change in the value of R_t also influences the evolution of diagnosed population $x_{\mathtt{D}}(t)$. This is because larger value of R_t results in a larger infected population $x_{\mathtt{I}}(t)$ and smaller value of R_t results in a smaller infected population $x_{\mathtt{I}}(t)$, which respectively increases and decreases the probability of detecting an infected person $x_{\mathtt{I}}(t)/x_{\mathtt{T}}(t)$ by a single test. Keeping the number of tests performed per day same, the larger probability of detection $x_{\mathtt{I}}(t)/x_{\mathtt{T}}(t)$ results in a larger diagnosed population $x_{\mathtt{D}}(t)$.

In Table 7.2, we see that the placement and lifting of lockdown on March 17 and May 11, respectively, had a significant impact on the value of R_t. Such an effect on R_t impacted the evolution of the diagnosed population $x_{\mathtt{D}}(t) = y_1(t) - y_2(t)$, which can be interpreted as the number of active confirmed cases and is illustrated in Figure 7.9. The placement of lockdown reduced the value of R_t and resulted in less number of active confirmed cases as compared to the scenario in Figure 7.9 where the lockdown was not placed on March 17. In this scenario, as shown in Figure 7.9, the number of active confirmed cases would have increased to a point that could have challenged the available medical facilities such as hospital beds, ventilators, and ICUs. On the other hand, the lifting of lockdown increased

[7] Sortir à Paris article "Coronavirus: vacances d'été partout en France et en Europe" published on May 28, 2020

[8] Santé Publique France: Epidemiological update of COVID-19, September 2020

[9] Open platform for French public data (data.gouv.fr)

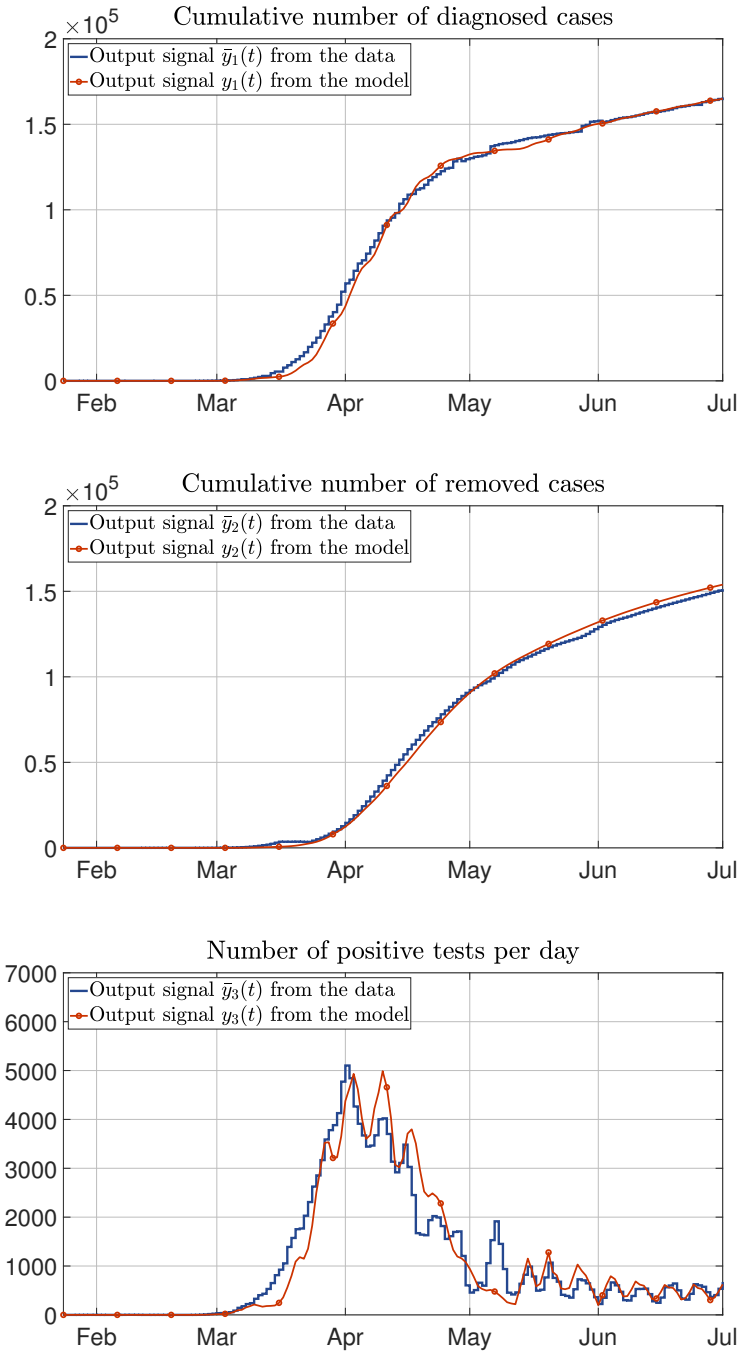


Figure 7.8: Model validation.

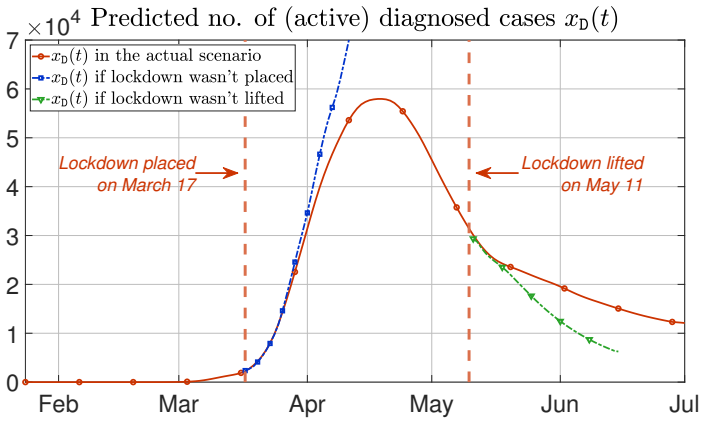


Figure 7.9: The comparison between the number of active diagnosed cases in the actual scenario vs. two scenarios: if the lockdown was not placed on March 17 and if the lockdown was not lifted on May 11.

the value of R_t and resulted in more number of active confirmed cases as compared to the scenario where the lockdown was not lifted on May 11.

7.3.5 Number of active ICU patients and deaths

The number of active ICU patients $B(t)$ is a function of the number of active infected people $A(t)$. Since an infected person starts to show symptoms after the average incubation period of approximately 5 days, [Lauer2020], and a person takes on average 12 days from being diagnosed to being addmitted to ICU, [Zhou2020], we assume $\psi = 5 + 12 = 17$ days to be the average time delay from getting infected to being admitted to ICU for a typical COVID-19 critically ill case. Thus, we model the number of active ICU patients $B(t)$ as a function of $A(t-\psi)$, which is approximated by:

$$B(t) = b_1 A(t-\psi) + b_2\sqrt{A(t-\psi)} \tag{7.19}$$

where b_1 and b_2 are the parameters given in Table 7.3, which are determined via the least-square solution to fit (7.19) to the data on the number of ICU patients. This is illustrated in Figure 7.10.

Parameters of $B(t)$	$b_1 = -0.54 \times 10^4$	$b_2 = 1.25 \times 10^4$
Parameters of $E(t)$	$e_1 = 4.14 \times 10^4$	$e_2 = 7.92 \times 10^5$
	$e_3 = -1.27 \times 10^7$	$e_4 = 9.04 \times 10^7$
	$e_5 = -3.63 \times 10^8$	$e_6 = 8.81 \times 10^8$
	$e_7 = -1.32 \times 10^9$	$e_8 = 1.19 \times 10^9$
	$e_9 = -5.93 \times 10^8$	$e_{10} = 1.25 \times 10^8$

Table 7.3: Estimated parameters b_1 and b_2 in (7.19) and e_i, for $i = 1, \ldots, 10$, in (7.20).

Similar to the case of the number of active ICU patients, a typical non-surviving case has an average incubation period of 5 days and, in addition to that, an average removal period of $\rho^{-1} \approx 20$ days, where ρ is the removal rate, Table 7.1. Thus, assuming $\phi = 5 + \rho^{-1} = 25$ days to be the average time delay from getting infected to death of a typical non-surviving COVID-19 case, we model the number of deaths $E(t)$ as a function of $I(t-\phi)$, which is

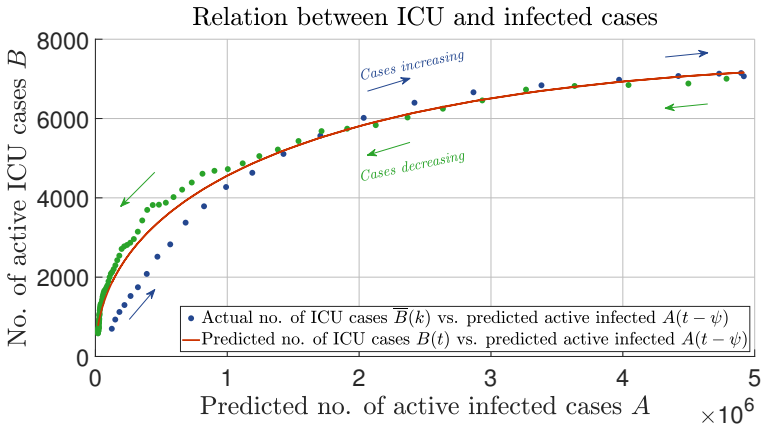


Figure 7.10: Number of active ICU patients $B(t)$ with respect to the number of active infected cases $A(t)$.

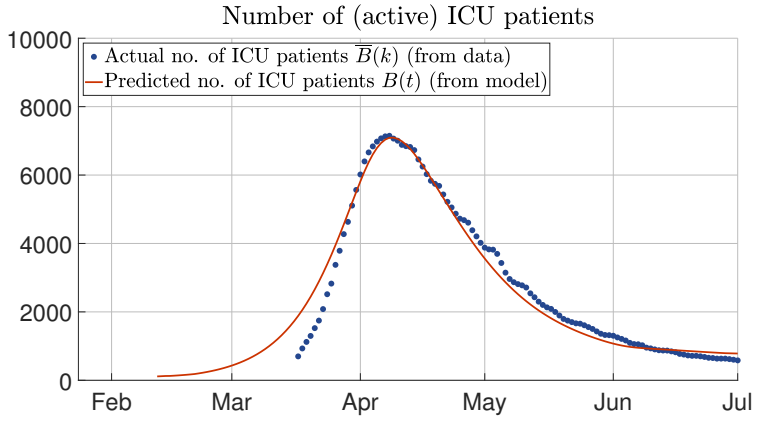


Figure 7.11: Model fit of the data on the number of active COVID-19 ICU cases $B(t)$ in France using the relation (7.19).

approximated by the following polynomial:

$$E(t) = \sum_{i=1}^{10} e_i I^i(t-\phi) \tag{7.20}$$

where e_i, for $i = 1, \ldots, 10$, are the parameters given in Table 7.3, which are determined via the least-square solution to fit (7.20) to the data on the number of deaths. This is illustrated in Figure 7.12.

Using the relations (7.19) and (7.20), we illustrate the model fit of the number of active ICU cases and the cumulative number of deaths with the data in Figure 7.11 and 7.13.

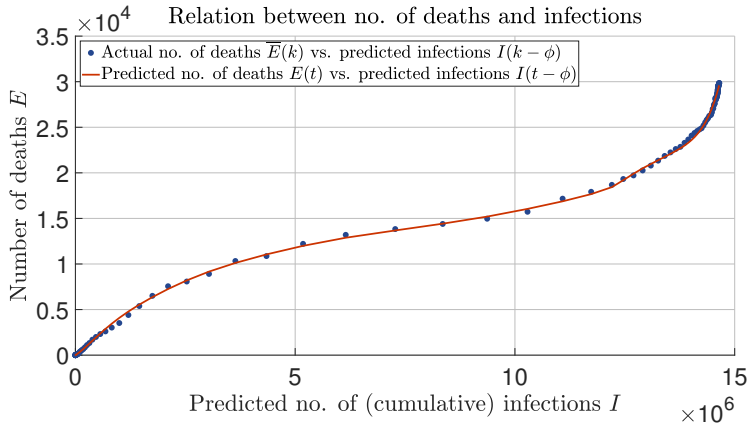


Figure 7.12: Cumulative number of deaths $E(t)$ with respect to the cumulative number of infected cases $I(t)$.

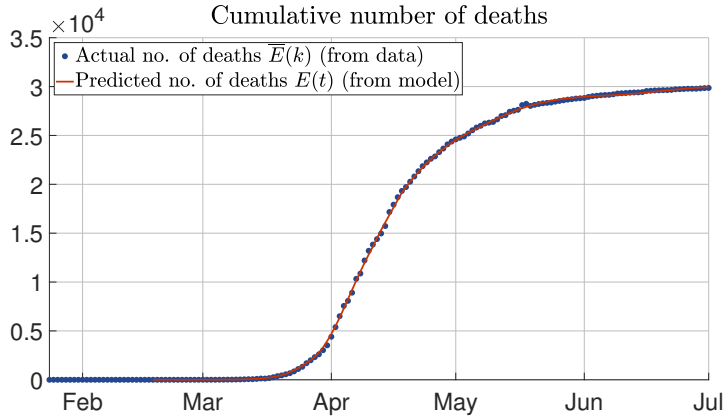


Figure 7.13: Model fit of the data on the number of COVID-19 deaths $E(t)$ in France using the relation (7.20).

7.4 Best-Effort Strategy for Testing

In this section, we design a testing policy for epidemic suppression and use the SIDUR model validated with the COVID-19 data of France to evaluate the testing policy. The testing policy is called *Best-Effort Strategy for Testing (BEST)*, which gives the minimum

number of tests needed to be performed per day in order to stop the epidemic from growing. In other words, if the BEST is applied, then the number of new infections stop to grow with respect to time.

7.4.1 Definition and computation of BEST policy

Assume that a country can manufacture or buy tests continuously during the time of the epidemic, and there is no limitation on the total stockpile of tests during the whole epidemic period. Based on this, we provide a testing policy recommendation on the daily testing capacity starting from a certain time t^* in order to change the course of the epidemic in a sense that is defined below. For simplicity, we further assume that the number of tests performed per day is considered to be the daily testing capacity $c(t)$. In other words, the daily testing capacity is utilized completely each day, i.e., $u(t) = c(t)$.

We say that, at time t, an epidemic is *spreading* if the number of undiagnosed infected population $x_{\mathtt{I}}(t)$ is increasing, i.e., the effective reproduction number $R_t > 1$. On the other hand, an epidemic is *non-spreading* if $x_{\mathtt{I}}(t)$ is not increasing, i.e., the effective reproduction number $R_t \leq 1$.

Definition 7.1 (BEST)**.** The *best effort strategy for testing (BEST) at a given time* t^* is the minimum number of tests to be performed per day from time t^* onward such that the epidemic switches from spreading to non-spreading at t^*.

In other words, the BEST policy provides the smallest lower bound on the testing rate $u(t)$ sufficient to change at given time t^* the course of the epidemic from spreading to non-spreading. Notice that the BEST policy is meaningful only for time t^* when the epidemic is spreading. If the epidemic is already non-spreading, the BEST policy is equal to 0.

Before presenting the BEST policy computation, we first establish the decreasing property of the testable population $x_{\mathtt{T}}(t)$.

Lemma 7.1. The testable population $x_{\mathtt{T}}(t)$ decreases on any interval on which $x_{\mathtt{I}}(t)$ is decreasing and $\theta(t)$ is non-decreasing.

Proof. Consider an interval (t, t'), $t < t'$, on which $x_{\mathtt{I}}$ is decreasing while θ is non-decreasing. From (7.2), we have

$$\begin{aligned} x_{\mathtt{T}}(t) &= \theta(t)(x_{\mathtt{I}}(t) + x_{\mathtt{D}}(t) + x_{\mathtt{R}}(t) - N) + (N - x_{\mathtt{D}}(t) - x_{\mathtt{R}}(t)) \\ &\geq \theta(t')(x_{\mathtt{I}}(t) + x_{\mathtt{D}}(t) + x_{\mathtt{R}}(t) - N) + (N - x_{\mathtt{D}}(t) - x_{\mathtt{R}}(t)) \\ &= \theta(t')x_{\mathtt{I}}(t) + (1 - \theta(t'))(N - x_{\mathtt{D}}(t) - x_{\mathtt{R}}(t)) \end{aligned} \tag{7.21}$$

because $x_{\mathtt{I}}(t) + x_{\mathtt{D}}(t) + x_{\mathtt{R}}(t) - N < 0$ and $\theta(t)$ is non-negative and non-decreasing by assumption. Since $x_{\mathtt{I}}$ is supposed to be decreasing, therefore

$$\theta(t')x_{\mathtt{I}}(t) \geq \theta(t')x_{\mathtt{I}}(t'). \tag{7.22}$$

On the other hand

$$\dot{N} - \dot{x}_{\mathtt{D}} - \dot{x}_{\mathtt{R}} = -u\frac{x_{\mathtt{I}}}{x_{\mathtt{T}}} < 0$$

meaning that $N - x_{\mathtt{D}} - x_{\mathtt{R}}$ is a decreasing function. As $1 - \theta(t') \geq 0$, one gets

$$(1 - \theta(t'))(N - x_{\mathtt{D}}(t) - x_{\mathtt{R}}(t)) \geq (1 - \theta(t'))(N - x_{\mathtt{D}}(t') - x_{\mathtt{R}}(t')). \tag{7.23}$$

Adding the two inequalities (7.22) and (7.23), one deduces from (7.21) that $x_{\mathtt{T}}(t) \geq x_{\mathtt{T}}(t')$ whenever $x_{\mathtt{I}}$ is decreasing on (t, t'). A tighter examination shows that, as both expressions

$\theta(t')$ and $1-\theta(t')$ cannot be zero altogether, at least one of the two inequalities (7.22) and (7.23) is a strict inequality in (t, t'), which implies $x_{\mathtt{T}}(t) > x_{\mathtt{T}}(t')$.

□

Define a function

$$c^*(t) = x_{\mathtt{T}}(t) \left| \frac{\beta(t)}{N} x_{\mathtt{S}}(t) - \gamma \right|_+ \tag{7.24}$$

where, by definition, for any scalar z, $|z|_+ = z$ if $z > 0$ and $|z|_+ = 0$ otherwise.

Proposition 7.2. Assume that the infection rate β is non-increasing while the testing specificity parameter θ is non-decreasing on a time interval $[t^*, t_1)$, for some $t^* < t_1$. Then, the best effort strategy for testing (BEST) at time t^* is given by

$$u(t) = c^*(t^*) = x_{\mathtt{T}}(t^*) \left| \frac{\beta(t^*)}{N} x_{\mathtt{S}}(t^*) - \gamma \right|_+, \quad \forall t \in [t^*, t_1). \tag{7.25}$$

Proof. In order to prove that $c^*(t^*)$ for $t \in [t^*, t_1)$ is the BEST policy at t^*, we show the following:

(i) If $u(t) > c^*(t)$ (resp., $u(t) \geq c^*(t)$) for any $t \in [t^*, t_1)$, then $x_{\mathtt{I}}$ is decreasing (resp., non-increasing) on $[t^*, t_1)$.

(ii) If $u(t) > c^*(t^*)$ (resp., $u(t) \geq c^*(t^*)$) for any $t \in [t^*, t_1)$, then $x_{\mathtt{I}}$ is decreasing (resp., non-increasing) on $[t^*, t_1)$.

Assume that $u(t) > c^*(t)$ on $[t^*, t_1)$. Then, $\Phi(t) := \beta(t)\frac{x_{\mathtt{S}}(t)}{N} - \frac{u(t)}{x_{\mathtt{T}}(t)} - \gamma < 0$ which implies that $x_{\mathtt{I}}$ is decreasing since $\dot{x}_{\mathtt{I}}(t) = \Phi(t) x_{\mathtt{I}}(t)$ almost everywhere. If only the weaker assumption $u(t) \geq c^*(t)$ on $[t^*, t_1)$ is fulfilled, then, by using the continuity of the solutions of ODE with respect to perturbations of the right-hand side, one gets that $x_{\mathtt{I}}$ is non-increasing.

Assume now that $u(t) > c^*(t^*)$ on $[t^*, t_1)$, where $c^*(t^*)$ is constant. Then, by continuity, $u(t) > c^*(t)$ on a certain interval $[t^*, t_2)$, for some $t_2 \in (t^*, t_1]$. As a consequence of the result (i) shown previously, $x_{\mathtt{I}}$ decreases on $[t^*, t_2)$. Moreover, assume that t_2 is the maximal point in $(t^*, t_1]$ having this property. In order to show that $t_2 = t_1$, it is sufficient to show that $u(t_2) > c^*(t_2)$, otherwise one may consider a larger value for t_2 which will lead to a contradiction with the fact that it is maximal. Since $x_{\mathtt{I}}$ decreases on $[t^*, t_2)$ and θ is non decreasing, from Lemma 7.1, we can conclude that $x_{\mathtt{T}}$ also decreases on this interval. On the other hand, since $x_{\mathtt{S}}$ is always decreasing and β is non increasing, one can conclude that $c^*(t)$ also decreases on $[t^*, t_2)$. This is obtained by upper bounding $c^*(t)$. Thus, one has $c^*(t^*) > c^*(t)$, which implies that that $u(t_2) > c^*(t_2)$. Therefore, as $t_2 = t_1$, we have established that $x_{\mathtt{I}}$ decreases on the whole interval $[t^*, t_1)$. For the case where $u(t) \geq c^*(t^*)$, we can use the same argument of continuity of the trajectories.

From the previous results, one deduces that the BEST is given by $c^*(t^*)$ and the testing rate $u(t) \geq c^*(t^*)$ for $t \in [t^*, t_1)$. If $u(t) < c^*(t^*)$, for $t \in [t^*, t_1)$, then one can show easily that the epidemic goes on spreading in the interval $[t^*, t_1)$. Hence, $u(t) = c^*(t^*)$ is the BEST policy at t^*.

□

Proposition 7.2 states that the peak of $x_{\mathtt{I}}(t)$ is uniquely determined by the BEST policy $c^*(t^*)$, where the peak is achieved at time t^*. Therefore, Algorithm 10 can be used to set the peak time t^* once parameters β, γ, and θ are learned from the data.

Algorithm 10 Computation of the BEST policy at time t^*.

1. Inputs: N, β, γ, θ, t^*, $x_{\mathtt{S}}(t^*)$ and $x_{\mathtt{T}}(t^*)$.
2. Compute the BEST policy $c^*(t^*)$ using (7.24).
3. Set $u(t) = c^*(t^*)$, for all $t \geq t^*$.
4. Return to step 2 if β increases or θ decreases.

Remark 7.1. Requiring that β must not increase and θ must not decrease in the interval (t^*, t_1) for some $t_1 > t^*$ is necessary for the BEST policy. It is thus important to keep the external conditions that determine the values of β and θ either constant or such that β decreases (e.g., through the implementation of lockdown) and/or θ increases (e.g., through efficient contact tracing).

Remark 7.2. The case where β decreases and/or θ increases at some time $t_1 > t^*$ has the effect of speeding up the suppression of the epidemic under BEST policy.

Remark 7.3. From (7.1b), we can note that if $x_{\mathtt{S}}(t)/N < \gamma/\beta$, then the epidemic naturally decreases. In this case, doing no testing $u(t) = 0$ is the BEST policy, which, by definition, gives a minimum number of tests to be performed in order to stop the growth of the infected population $x_{\mathtt{I}}$. However, if the testing is resumed in this case, i.e., $u(t) > 0$, it will further speed up the decrease of the infected population.

7.4.2 Evaluation of the BEST policy

Given the COVID-19 data of France, we first compute $c^*(t^*)$ for different values of t^* from January 24 to March 13. Figure 7.14 shows the number of tests per day required by the BEST policy if it is implemented on day k and the corresponding value of peak of infected cases $x_{\mathtt{I}}(k^*)$. One can note that the later the BEST policy is applied the higher is the required number of tests. An exponential increase can be observed from February 28 which corresponds to an acceleration of the infection spread.

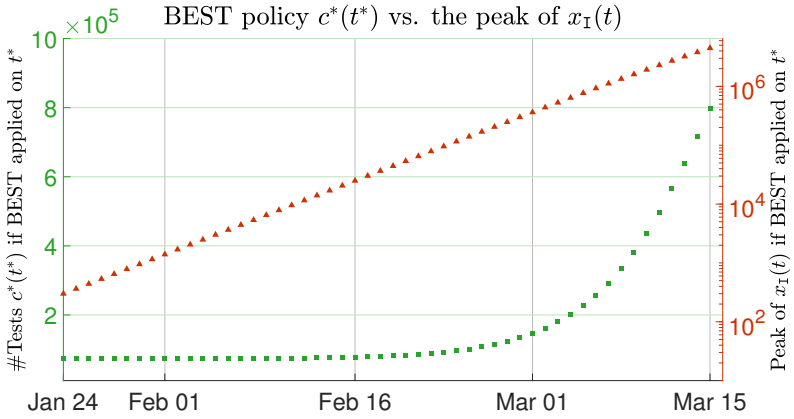


Figure 7.14: Number of tests per day required by the BEST policy (left y-axis, green) vs. peak of infection (right y-axis, red, in logscale) for an implementation day t^*.

Now, we evaluate the BEST policy by considering a scenario where BEST is implemented from March 01, 2020, onward. Figure 7.15 depicts the number of active cases when $u(t)$ is the actual testing scenario (see Figure 7.6) and when $u(t)$ is given by the BEST policy (7.25). For the evaluation, we use $u(t)$ as given by the recorded data on

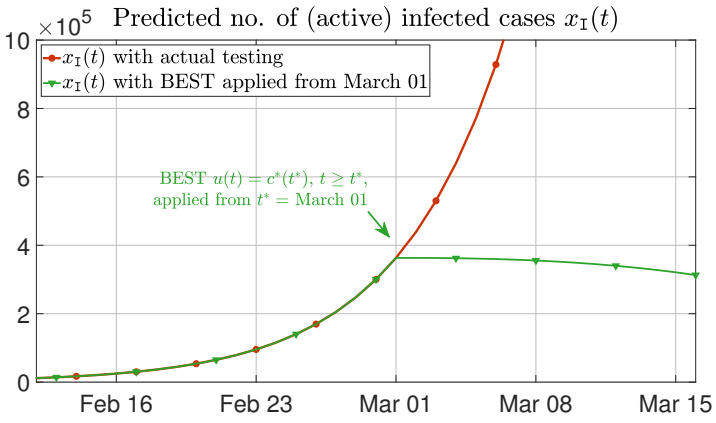


Figure 7.15: Predicted number of infected cases $x_I(t)$: actual testing scenario vs. BEST

the testing rate until March 01 and then use constant $u(t) = c^*(t^*)$ given by (7.25) from March 01 onward. In the first case, the peak of the infected population $x_I(t)$, which are the active undiagnosed cases, is about 6 million. In the second case, the peak of infected population in this case is $363,169$. The required number of tests per day to be performed for the implementation of BEST on March 01 is $c^* \approx 147,000$.

The impact in terms of ICU occupation and number of deaths is now evaluated using the equations (7.19) and (7.20) respectively. The results are illustrated in Figure 7.16 and 7.17. We observe that the peak of the number of active ICU patients could have been reduced by 34.71% and the number of deaths could have been reduced by 74.45% if the BEST policy was applied from March 01, 2020.

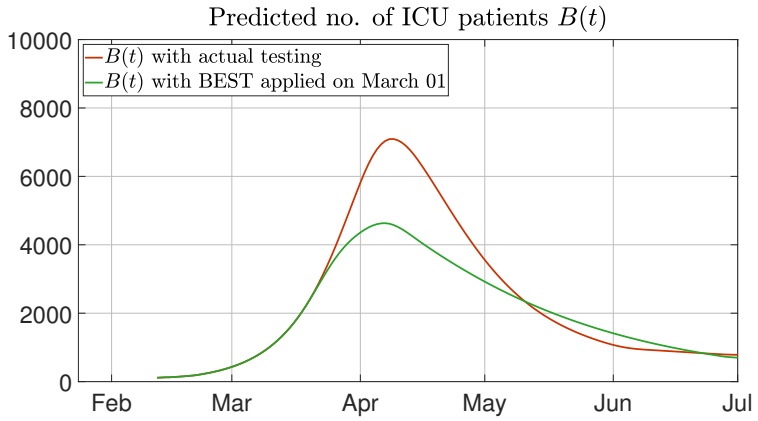


Figure 7.16: The prediction of the number of active ICU cases $B(t)$: actual scenario vs. BEST policy.

7.5 Concluding Remarks

We proposed a SIDUR model for the control of epidemics through testing rate. Testing enables the government to diagnose and isolate the infected people from the susceptible population. We estimated and validated the model for the COVID-19 case of France data. We proposed a best effort strategy for testing (BEST) for epidemic suppression, which

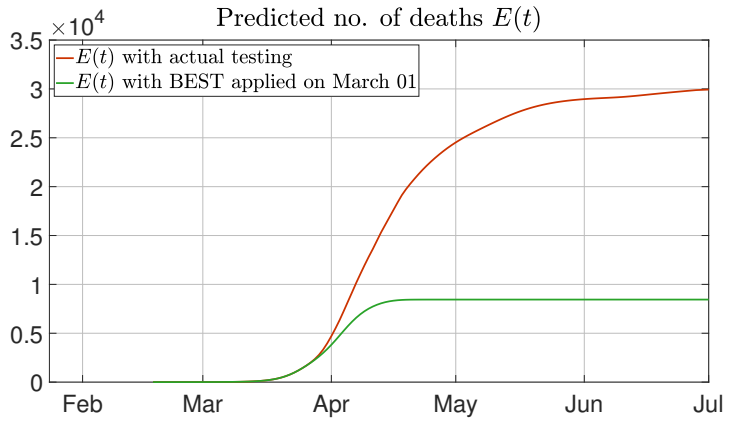


Figure 7.17: The prediction of the cumulative number of deaths $E(t)$: actual scenario vs. BEST policy.

provides the minimum number of tests to be performed from a certain day onward in order to make the increasing infected population non-increasing immediately. That is, it changes the course of epidemic from spreading to non-spreading.

For the COVID-19 case, the control input in SIDUR model corresponds to the number of RT-PCR tests performed per day. However, another type of test, a serology test, which is not considered in the current model because of the unavailability of its data, is also very important. A serology test determines the relevant antibodies in a subject's serum in order to detect whether he/she was infected in the past. By performing serology tests on the testable population, one can detect the unidentified recovered population and transfer them in the identified removed compartment of the model. This reduces the size of the testable population, which in turn increases the testing specificity of RT-PCR tests. In other words, the serology tests complement the RT-PCR tests [de Walque2020, Winter2020]. Therefore, as a future prospect, it will be interesting to consider two control inputs corresponding to both types of test in the SIDUR model.

The model is estimated and validated by fitting the model outputs with the available data of France. This allows us to predict the populations in the unmeasured compartments of the model. However, there is no certainty whether these predicted populations correspond to the reality because of the absence of feedback correction. Therefore, another future prospect is to design an observer for the SIDUR model that estimates the true states of the model.

The BEST policy is easy to compute and implement, however, it is static. Thus, its influence on the control of epidemic is limited. In future, it will be interesting to solve a finite/infinite-horizon optimal control problem to minimize the peak and/or cumulative number of the infected population by a dynamic control input.

8

Control of Urban Human Mobility for Epidemic Mitigation

This chapter develops a model of human mobility between origins (residential areas) and destinations (business parks, industrial areas, schools, markets, etc.) in an urban environment in section 8.1. The SIR epidemic spread process is incorporated into the mobility model in section 8.2. Then, after defining the economic activity of the population and the active infected cases in the city in section 8.3, we formulate and solve the optimal capacity control and the optimal schedule control problems in section 8.4 that maximize the economic activity while keeping the number of active infected cases bounded for epidemic mitigation.

Contents

Controlling human mobility during an epidemic is a fundamental issue faced by policy-makers. Such control can only be done optimally if human mobility is adequately modeled at the scale of a city or metropolis. This chapter, first, develops a model of human mobility that captures the daily patterns of mobility in an urban environment through time-dependent gating functions, which are controlled by the destination schedules and mobility windows. The process of epidemic spread is incorporated at each location that depends on the number of susceptible and infected people present at that location. Then, two optimal control policies are proposed to maximize the economic activity at the destinations while mitigating the epidemic. Precisely, operating capacities and time schedules of destinations are controlled to maximize the economic activity under the constraint that the number of active infected cases remains bounded.

8.1 Formulation of the Urban Human Mobility Model

Consider human mobility in an urban environment between locations of two types: origins and destinations. The origins correspond to locations where people reside—for example, residential areas, neighborhoods, and towns. The destinations, on the other hand, correspond to locations that people visit daily for work, education, shopping, or leisure—for example, industrial zones, business parks, schools, markets, cinemas, etc. We represent this mobility process by a flow network describing the transfer of people between different locations. The main idea of the urban human mobility model is that a certain number of people go from each origin to the destinations every day during specified time intervals and then return later the same day.

The flow of people from one location to another depends on the demand and supply of locations, which depend on the destination schedules, mobility windows, and the number of people in each location. The destination schedules correspond to the daily business hours of destinations during which they are open and people can visit them, whereas the mobility windows between two locations correspond to specified time intervals during which there is mobility of people between those locations. We consider that each destination has an operating capacity that corresponds to the maximum number of people that can visit the destination at any time. The operating capacity of a destination is less than or equal to its nominal capacity and depends on the government's policy during an epidemic in order to reduce the maximum number of people that can gather in the destinations at any time. Therefore, the flow to a certain location stops when the number of people in that location reaches its operating capacity.

8.1.1 Network representation and main assumptions

Let the index set of m origins be

$$\mathcal{V}_o = \{1, \ldots, m\}$$

and the index set of n destinations be

$$\mathcal{V}_d = \{m+1, \ldots, m+n\}.$$

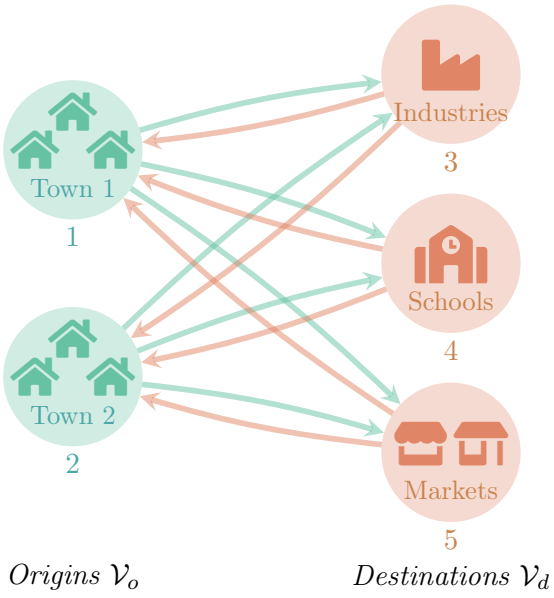


Figure 8.1: An example of an urban human mobility network with two origins and three destinations.

Denote the *total population of origin* $i \in \mathcal{V}_o$ by P_i, which is the number of people who reside in i, and the *nominal capacity of destination* $j \in \mathcal{V}_d$ by C_j, which is the maximum number of people who can visit j nominally at one time. By nominal we mean the times when there is no epidemic and there is no government policy that restricts the maximum number of people who can gather in any destination. Note that the total population of the city is given by

$$P = \sum_{i \in \mathcal{V}_o} P_i.$$

The network of urban human mobility is represented by a bi-directed, bipartite graph $\mathcal{G} = (\mathcal{V}_o, \mathcal{V}_d, \mathcal{E})$, where $\mathcal{E}$ is the set of bi-directed edges—i.e., for every $i \in \mathcal{V}_o$ and $j \in \mathcal{V}_d$, if $(i, j) \in \mathcal{E}$ then $(j, i) \in \mathcal{E}$. An example of a mobility network is illustrated in Figure 8.1.

Assumption 8.1. We adopt the following assumptions:

(i) The total population of the city remains constant.

(ii) The mobility occurs only between pairs of origins and destinations, and not among a pair of different origins or a pair of different destinations.

(iii) The number of people who visit destination j from origin i during a day is equal to the number of people who return to i from j on the same day.

(iv) The mobility pattern between each pair of origins and destinations is periodic and repeats every day, i.e., the period $T_{\text{period}} = 24$ hours. In particular, we ignore mobility patterns of the weekends or holidays that may be different than the normal days.

Note that Assumption 8.1(iv) is a simplifying assumption, which can be relaxed without loss of generality by considering T_{period} to be a week, a month, etc.

8.1.2 Destination categories

Suppose the destinations are divided into $p \leq n$ categories, which correspond to a type of destination such as workplace, school, market, etc. The categories of destinations are represented by a partition

$$\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_p\}$$

where the destinations of category k are grouped in a set $\mathcal{D}_k$, for $k = 1, \ldots, p$, and each destination belongs to only one category, i.e., for every $k, l \in \{1, \ldots, p\}$ and $l \neq k$,

$$\bigcup_{k=1}^{p} \mathcal{D}_k = \mathcal{V}_d \quad \text{and} \quad \mathcal{D}_k \cap \mathcal{D}_l = \emptyset.$$

In the example illustrated in Figure 8.1, we have $p = n$, i.e., three categories and three destinations, because each destination is of a different category. However, it is possible that a mobility network may have multiple destinations of the same category. For instance, a slight modification of the example of Figure 8.1, say with two working places and three schools, could illustrate this immediately.

8.1.3 Operating capacities of destinations

Let $u_k(t) \in [0, 1]$ be the *capacity control input* for destinations of category k, for $k = 1, \ldots, p$, which determines the allowed operating capacity of $\mathcal{D}_k$ in terms of the proportion of nominal capacity at time t in the event of an epidemic. In other words, it can be considered as the government's policy at time t that limits the operating capacities in destinations of category k in order to mitigate the epidemic spread, where

$$\textit{Operating capacity} = C_j u_k(t), \quad \text{for } j \in \mathcal{D}_k.$$

We consider $u_k(t)$ to be piece-wise constant—i.e.,

$$u_k(t) = \begin{cases} \mu_k^1 & \text{if } t \in [0, T_u) \\ \mu_k^2 & \text{if } t \in [T_u, 2T_u) \\ \vdots & \vdots \\ \mu_k^q & \text{if } t \in [(q-1)T_u, qT_u) \end{cases} \tag{8.1}$$

with $\mu_k^h \in [0, 1]$ constant for every $h \in \{1, \ldots, q\}$ and $T = qT_u$ the total time horizon considered by a policymaker. The policy horizon T_u is a multiple of T_{period} and corresponds to the time after which the policy on operating capacities is announced periodically. It can be chosen by the policymaker from at least a week to several months because changing the policy on shorter time intervals may not be practical in terms of implementation.

8.1.4 Destination schedule and mobility window

The *destination schedule* of $j \in \mathcal{V}_d$

$$\mathbb{S}_j = [a_j, b_j), \qquad 0 \leq a_j < b_j \leq 24$$

is the daily time interval during which j is open, where a_j and b_j are the nominal opening and closing hours of j, respectively. The origins, on the other hand, are open throughout

the day, i.e., for every $i \in \mathcal{V}_o$, $\mathbb{S}_i = [0, 24)$. Then, the *supply gating function* (SGF) for $j \in \mathcal{V}_o \cup \mathcal{V}_d$

$$\sigma_j(t) = \begin{cases} 1 & \text{if } t \bmod 24 \in \mathbb{S}_j \\ 0 & \text{otherwise} \end{cases} \tag{8.2}$$

which is periodic with respect to 24 hours.

The *mobility window* of $(i, j) \in \mathcal{E}$

$$\mathbb{D}_{ij} = [t_{ij}, t_{ij} + \tau_{ij}), \qquad 0 \le t_{ij} < t_{ij} + \tau_{ij} \le 24$$

is the daily time interval during which there is mobility from i to j, where $\tau_{ij} > 0$ is the duration of mobility window in hours. Then, the *demand gating function* (DGF) of $(i, j) \in \mathcal{E}$

$$\delta_{ij}(t) = \begin{cases} 1 & \text{if } t \bmod 24 \in \mathbb{D}_{ij} \\ 0 & \text{otherwise} \end{cases} \tag{8.3}$$

which is also periodic with respect to 24 hours.

8.1.5 Model of urban human mobility

Let $N_i(t) \ge 0$ be the number of people in $i \in \mathcal{V}_o \cup \mathcal{V}_d$ at time t (hour). Then, according to the urban human mobility model, *the rate of change of the number of people at any location at time t is equal to the sum of inflows to that location minus the sum of outflows from that location.*

In other words, for any $i \in \mathcal{V}_o \cup \mathcal{D}_k$ and $k \in \{1, \dots, p\}$, the urban human mobility model is given by

$$\dot{N}_i = \sum_{j \in \mathcal{N}_i} (\phi_{ji} - \phi_{ij}) \tag{8.4}$$

where $\mathcal{N}_i$ is the set of neighbors of i in the mobility network $\mathcal{G}$ and $\phi_{ij}(t, N_i(t), N_j(t), u_k(t))$ is the flow from $i \in \mathcal{V}_o$ to $j \in \mathcal{D}_k$ given as

$$\phi_{ij} = \min(\Delta_{ij}, \Sigma_j)$$

with $\Delta_{ij}(t, N_i(t), u_k(t))$ and $\Sigma_j(t, N_j(t), u_k(t))$ the demand and supply functions, respectively. Notice that the flow $\phi_{ji}(t, N_j(t), N_i(t), u_k(t))$ is defined similarly with the subscript ji instead of ij.

The *supply function* $\Sigma_j(t, N_j(t), u_k(t))$ of each location j corresponds to the allowed inflow to j from other locations and is given by

$$\Sigma_j = \begin{cases} \sigma_j \min(F_j, v[C_j u_k - N_j]) & \text{if } j \in \mathcal{D}_k \\ \sigma_j \min(F_j, v[P_j - N_j]) & \text{if } j \in \mathcal{V}_o \end{cases}$$

where $\sigma_j(t)$ is the SGF given by (8.2), $v > 0$ is a regularization parameter taken to be very large (see Remark 8.1), $C_j u_k(t)$ is the operating capacity of $j \in \mathcal{D}_k$ with $u_k(t)$ defined in (8.1), and

$$F_j(t) = \sum_{i \in \mathcal{N}_j} f_{ij}(t)$$

is the maximum inflow to j with

$$f_{ij}(t) = \frac{M_{ij} u_k(t)}{\tau_{ij}} \tag{8.5}$$

the maximum outflow from $i \in \mathcal{V}_o$ to $j \in \mathcal{D}_k$. Here, M_{ij} denotes the nominal number of visitors to j from i and $M_{ij}u_k(t)$ is the number of visitors when the capacity control input $u_k(t)$ is implemented. Notice that, since all visitors return, we have $M_{ji} = M_{ij}$.

The *demand function* $\Delta_{ij}(t, N_i(t), u_k(t))$ of each edge $(i, j) \in \mathcal{E}$ corresponds to the outflow from i towards j and is given by

$$\Delta_{ij} = \delta_{ij} \min\left(vN_i, f_{ij}\right)$$

where $\delta_{ij}(t)$ is the DGF given by (8.3), $v > 0$ is the same regularization parameter introduced in the supply function $\Sigma_j(t, N_j(t), u_k(t))$, and $f_{ij}(t)$ is the maximum outflow from i to j given by (8.5).

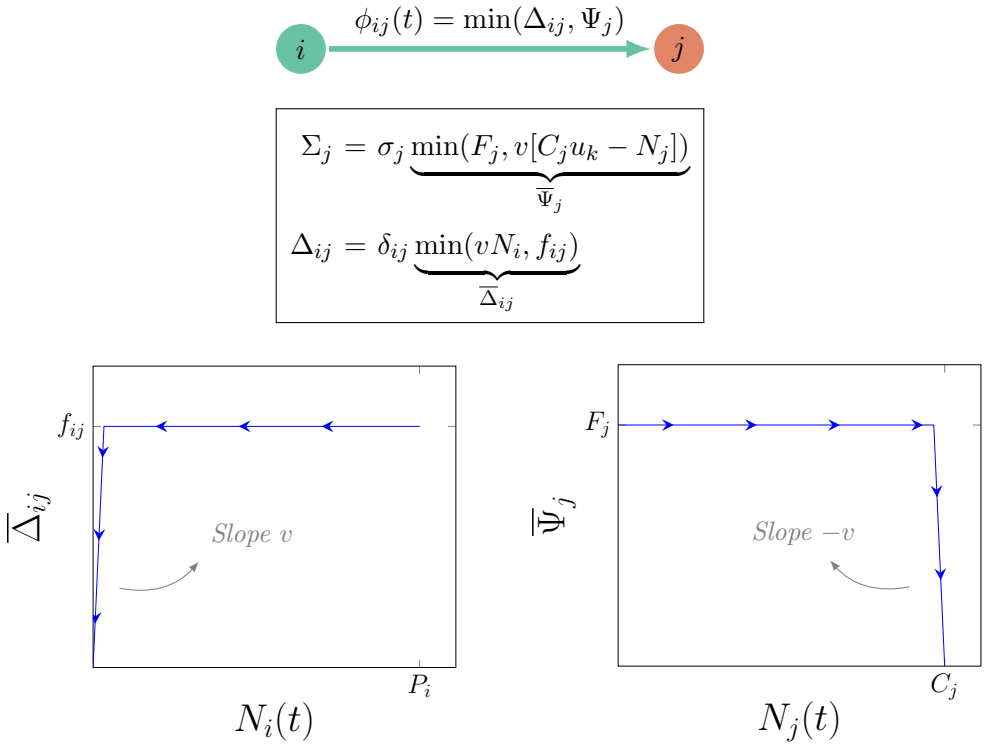


Figure 8.2: An example illustrating the flow $\phi_{ij}(t)$ from origin i to destination j in terms of supply of j and demand of i with respect to j. Here, the arrows on each curve indicate the time evolution.

Suppose $i \in \mathcal{V}_o$ and $j \in \mathcal{D}_k$, then Figure 8.2 illustrates the flow from i to j in terms of demand and supply functions. In the figure, notice that the demand of i moves from right to left with respect to time t, i.e., from being full to being empty, and the supply of j moves from left to right with respect to time t, i.e., from being empty to being full, which is indicated by arrows in the figure.

Remark 8.1. To ensure that the daily number of people going from i to j equals $M_{ij}u_k(t)$, we assume that the demand function $\Delta_{ij}(t, N_i(t), u_k(t)) = \delta_{ij}(t)f_{ij}(t)\mathbb{1}_{N_i(t)>0}$, where $\mathbb{1}_{N_i(t)>0} = 1$ if $N_i(t) > 0$, and 0 otherwise, is the indicator function. Similarly, we assume that the supply function $\Psi_j(t, N_j(t), u_k(t)) = \sigma_j(t)F_j(t)\mathbb{1}_{N_j(t)<C_j u_k(t)}$. However, to avoid the discontinuity posed by the indicator functions, we approximate the demand and supply functions by considering steep slope with a very large regularization parameter v as illustrated in Figure 8.2. ◇

Populations	$P_1 = 3000,$	$P_2 = 2000$	
Capacities	$C_3 = 2000,$	$C_4 = 1500,$	$C_5 = 200$
No. of Visitors	$M_{13} = 1200,$	$M_{14} = 900,$	$M_{15} = 900$
	$M_{23} = 800,$	$M_{24} = 600,$	$M_{25} = 600$
Mobility windows	$\mathbb{D}_{i3} = [8, 9.5),$	$\mathbb{D}_{i4} = [9, 10),$	$\mathbb{D}_{i5} = [8.5, 20)$
	$\mathbb{D}_{3i} = [17, 18.5),$	$\mathbb{D}_{4i} = [16, 17),$	$\mathbb{D}_{5i} = [10, 21)$
		for $i = 1, 2$	
Destination schedules	$\mathbb{S}_3 = [8, 18),$	$\mathbb{S}_4 = [9, 17),$	$\mathbb{S}_5 = [8.5, 20)$
Capacity control policy	$u_1(t) = 0.5,$	$u_2(t) = 0.5,$	$u_3(t) = 1$
		for $t \in [0, 48)$	
Regularization parameter	$v = 30$		

Table 8.1: Parameters related to urban human mobility for the example of Figure 8.1.

Example 8.1. Consider the example of mobility network shown in Figure 8.1 with two origins and three destinations. For the mobility model (8.4), we consider the parameters given in Table 8.1. The mobility profile of two days is plotted in Figure 8.3, where $N_1(t), N_2(t)$ denote the number of people in Town 1 and 2 at time t, respectively, and $N_3(t), N_4(t), N_5(t)$ denote the number of people in the industries, schools, and markets, respectively. As shown in the figure, people go from the origins (1 and 2) to the destinations (3, 4, and 5) and return on the same day according to the destination schedules and mobility windows. Notice that the mobility profiles are the same for both days because, by Assumption 8.1(iv), the destination schedules and mobility windows are the same for every day. Moreover, the nominal capacities of industries and schools are $C_3 = 2000$ and $C_4 = 1500$, however, the capacity control policy $u_1(t) = u_2(t) = 0.5$ reduce the operating capacities to 50% of the nominal capacities. Therefore, the maximum number of people present in these destinations during a day is around $C_3/2 = 1000$ and $C_4/2 = 750$, respectively.

8.2 Incorporating Epidemic Spread Process in Mobility

When people gather at a certain location during the mobility process, the epidemic spreads process occurs at that location which is described here by a SIR model, [Hethcote2000], where the population is divided into Susceptible, Infected, and Recovered classes, and the disease is transmitted according to the local infection rates when the susceptible and infected populations mix in the same location. Notice that any similar epidemiological model could be used instead.

8.2.1 Urban human mobility with epidemic spread

According to the SIR model of epidemic spread, the number of people $N_i(t)$ at each location $i \in \mathcal{V}_o \cup \mathcal{V}_d$ are divided into three classes: number of susceptible $S_i(t)$, infected $I_i(t)$,

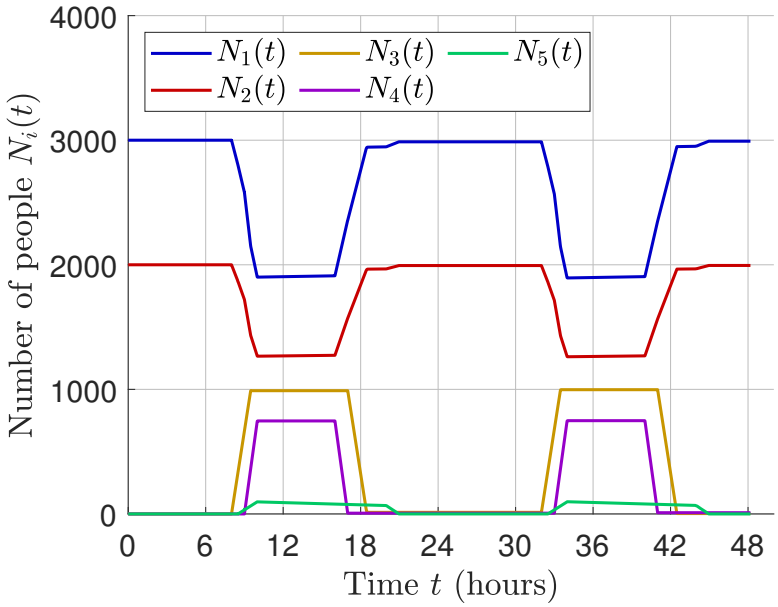


Figure 8.3: The mobility profile of two days for the example in Figure 8.1.

and recovered $R_i(t)$, where, at every time t,

$$N_i(t) = S_i(t) + I_i(t) + R_i(t).$$

The disease transmission at each location i occurs according to the local mass action law

$$\beta_i(t) S_i(t) \frac{I_i(t)}{N_i(t)}$$

where

$$\beta_i(t) = \begin{cases} \overline{\beta}_i \frac{N_i(t)}{P_i} & \text{if } i \in \mathcal{V}_o \\ \overline{\beta}_i \frac{N_i(t)}{C_i} & \text{if } i \in \mathcal{V}_d \end{cases}$$

is the *infection rate* of i at time t with $\overline{\beta}_j$ the *nominal infection rate* of i. The nominal infection rate is defined as the average number of contacts of a person in location i per hour when the number of people in i is maximum. The infection rate $\beta_i(t)$ reduces when the number of people $N_i(t)$ at location i is small and increases when $N_i(t)$ is large. The infected people $I_i(t)$ recover with a *recovery rate* $\gamma \in (0, 1]$, which is a constant that depends on the disease biology and, if available, the treatment methods. The recovery rate γ is defined as the inverse of the average recovery period (in hours) of the infected cases.

Assumption 8.2. The restrictions imposed on the urban human mobility by the government affects all the people, whether susceptible, infected, or recovered, equally.

By Assumption 8.2, the flow from i to j in terms of the number of susceptible, infected, and recovered can be respectively given by

$$\phi_{ij}(t, N_i(t), N_j(t), u_k(t)) \frac{S_i(t)}{N_i(t)}$$

$$\phi_{ij}(t, N_i(t), N_j(t), u_k(t)) \frac{I_i(t)}{N_i(t)}$$

$$\phi_{ij}(t, N_i(t), N_j(t), u_k(t)) \frac{R_i(t)}{N_i(t)}.$$

Let $\mathbf{x}_i(t) = \begin{bmatrix} S_i(t) & I_i(t) & R_i(t) \end{bmatrix}^\intercal \in \mathbb{R}^3_{\geq 0}$ be the state vector of location $i \in \mathcal{V}_o \cup \mathcal{V}_d$ and

$$\boldsymbol{\xi}_i(\mathbf{x}_i(t)) = \begin{bmatrix} -\beta_i(t) S_i(t) \frac{I_i(t)}{N_i(t)} \\ \beta_i(t) S_i(t) \frac{I_i(t)}{N_i(t)} - \gamma I_i(t) \\ \gamma I_i(t) \end{bmatrix} \in \mathbb{R}^3 \tag{8.6}$$

be the vector describing the process of epidemic spread in location i. Then, for $i \in \mathcal{V}_o$ and $j \in \mathcal{D}_k$, the model of urban human mobility with epidemic spread is given by

$$\begin{aligned} \dot{\mathbf{x}}_i &= \boldsymbol{\xi}_i(\mathbf{x}_i) + \sum_{k=1}^{p} \sum_{j \in \mathcal{D}_k} \left[\phi_{ji}(u_k) \frac{\mathbf{x}_j}{N_j} - \phi_{ij}(u_k) \frac{\mathbf{x}_i}{N_i} \right] \\ \dot{\mathbf{x}}_j &= \boldsymbol{\xi}_j(\mathbf{x}_j) + \sum_{i \in \mathcal{V}_o} \left[\phi_{ij}(u_k) \frac{\mathbf{x}_i}{N_i} - \phi_{ji}(u_k) \frac{\mathbf{x}_j}{N_j} \right]. \end{aligned} \tag{8.7}$$

As illustrated in Figure 8.4, there are two aspects of the model. First, inside the locations i and j, there is a process of epidemic spread that transmits the disease from the infected to the susceptible with a local infection rate, and the recovery process that heals the infected with a constant recovery rate. Second, on the edges (i, j) and (j, i), there is a process of human mobility that transfers people from one location to another through the flows $\phi_{ij}(t, N_i(t), N_j(t), u_k(t))$ and $\phi_{ji}(t, N_j(t), N_i(t), u_k(t))$, respectively.

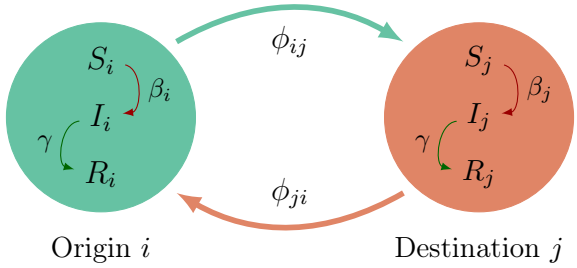


Figure 8.4: The process of urban human mobility happens along the edges whereas the process of epidemic spread happens inside the locations.

8.2.2 Compact representation of the model

The urban human mobility model with epidemic spread given in (8.7) describes the dynamics of the number of susceptible $S_i(t)$, infected $I_i(t)$, and recovered $R_i(t)$ people in location $i \in \mathcal{V}_o \cup \mathcal{V}_d$. These dynamics are controlled by the piece-wise constant capacity control inputs $u_1(t), \ldots, u_p(t)$, defined in (8.1), of the p destination categories in $\mathcal{D}$. Define

$$\mathbf{u}(t) = \begin{bmatrix} u_1(t) & \ldots & u_p(t) \end{bmatrix}^\intercal.$$

Then, to represent the model in a compact form, let

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_o \\ \mathbf{x}_d \end{bmatrix}, \quad \text{where} \quad \begin{cases} \mathbf{x}_o &= \begin{bmatrix} \mathbf{x}_1^\intercal & \cdots & \mathbf{x}_m^\intercal \end{bmatrix}^\intercal \\ \mathbf{x}_d &= \begin{bmatrix} \mathbf{x}_{m+1}^\intercal & \cdots & \mathbf{x}_{m+n}^\intercal \end{bmatrix}^\intercal \end{cases}$$

with $\mathbf{x}_o(t) \in \mathbb{R}^{3m}_{\geq 0}$ the state vector of origins $\mathcal{V}_o$ and $\mathbf{x}_d(t) \in \mathbb{R}^{3n}_{\geq 0}$ the state vector of destinations $\mathcal{V}_d$. Similarly, let

$$\boldsymbol{\xi} = \begin{bmatrix} \boldsymbol{\xi}_o \\ \boldsymbol{\xi}_d \end{bmatrix}, \quad \text{where} \begin{cases} \boldsymbol{\xi}_o &= [\, \boldsymbol{\xi}_1^\intercal \;\; \cdots \;\; \boldsymbol{\xi}_m^\intercal \,]^\intercal \\ \boldsymbol{\xi}_d &= [\, \boldsymbol{\xi}_{m+1}^\intercal \;\; \cdots \;\; \boldsymbol{\xi}_{m+n}^\intercal \,]^\intercal \end{cases}$$

with $\boldsymbol{\xi}_o(\mathbf{x}_o(t)) \in \mathbb{R}^{3m}$ the vector describing the epidemic process in the origins $\mathcal{V}_o$ and $\boldsymbol{\xi}_d(\mathbf{x}_d(t)) \in \mathbb{R}^{3n}$ the vector describing the epidemic process in the destinations $\mathcal{V}_d$. Notice that the vector $\boldsymbol{\xi}_i(\mathbf{x}_i(t))$ describing the epidemic process in each location i is given in (8.6).

The model (8.7) can be represented as

$$\dot{\mathbf{x}} = \boldsymbol{\xi}(\mathbf{x}) + \Phi(\mathbf{x}, \mathbf{u})\mathbf{x} \tag{8.8}$$

where the dependence on t is omitted for brevity. The matrix of flows $\Phi(t, \mathbf{x}(t), \mathbf{u}(t))$ describes the mobility process in the network $\mathcal{G}$ and is given as

$$\Phi(\mathbf{x}, \mathbf{u}) = \begin{bmatrix} \Phi_{oo}(\mathbf{x}, \mathbf{u}) & \Phi_{do}(\mathbf{x}, \mathbf{u}) \\ \Phi_{od}(\mathbf{x}, \mathbf{u}) & \Phi_{dd}(\mathbf{x}, \mathbf{u}) \end{bmatrix} \otimes I_3$$

where $\otimes$ denotes the Kronecker product and

$$\Phi_{oo} = \mathtt{diag}\left[-\sum_{j \in \mathcal{V}_d} \frac{\phi_{1j}}{N_1} \;\; \cdots \;\; -\sum_{j \in \mathcal{V}_d} \frac{\phi_{mj}}{N_m} \right]$$

$$\Phi_{dd} = \mathtt{diag}\left[-\sum_{i \in \mathcal{V}_o} \frac{\phi_{m+1,i}}{N_{m+1}} \;\; \cdots \;\; -\sum_{i \in \mathcal{V}_o} \frac{\phi_{m+n,i}}{N_{m+n}} \right]$$

$$\Phi_{od} = \begin{bmatrix} \frac{\phi_{1,m+1}}{N_1} & \cdots & \frac{\phi_{m,m+1}}{N_m} \\ \vdots & \ddots & \vdots \\ \frac{\phi_{1,m+n}}{N_1} & \cdots & \frac{\phi_{m,m+n}}{N_m} \end{bmatrix}$$

$$\Phi_{do} = \begin{bmatrix} \frac{\phi_{m+1,1}}{N_{m+1}} & \cdots & \frac{\phi_{m+n,1}}{N_{m+n}} \\ \vdots & \ddots & \vdots \\ \frac{\phi_{m+1,m}}{N_{m+1}} & \cdots & \frac{\phi_{m+n,m}}{N_{m+n}} \end{bmatrix}.$$

8.3 Economic Activity and Active Infected Cases

The *economic activity* $E(t) \in \mathbb{R}_{\geq 0}$ in the mobility network $\mathcal{G}$ is defined as

$$E(t) = \sum_{k=1}^{p} \sum_{j \in \mathcal{D}_k} \chi_j \frac{N_j(t)}{C_j} \tag{8.9}$$

where $N_j(t)$ is the number of people in destination $j \in \mathcal{D}_k$ at time t, C_j is the nominal capacity of j, and $\chi_j \in [0, 1]$ is the weight assigned to j according to its economic importance such that $\sum_{j \in \mathcal{V}_d} \chi_j = 1$. Since $N_i(t) = \mathbf{1}_3^\intercal \mathbf{x}_i(t)$, we can write (8.9) as

$$E(t) = \mathbf{e}^\intercal \mathbf{x}(t) \tag{8.10}$$

where $\mathbf{x}(t)$ is the state of (8.8) at time t and

$$\mathbf{e} = \begin{bmatrix} \mathbf{0}_m^\intercal & \frac{\chi_{m+1}}{C_{m+1}} & \cdots & \frac{\chi_{m+n}}{C_{m+n}} \end{bmatrix}^\intercal \otimes \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^\intercal.$$

The number of *active infected cases* $I(t) \in \mathbb{R}_{\geq 0}$ in the mobility network $\mathcal{G}$ at time t is the sum of the number of infected people in all the locations. It is given by

$$I(t) = \sum_{i \in \mathcal{V}_o \cup \mathcal{V}_d} I_i(t) \tag{8.11}$$

where $I_i(t)$ is the number of infected people at location i at time t. Since $I_i(t) = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \mathbf{x}_i(t)$, we can write (8.11) as

$$I(t) = \mathbf{g}^\intercal \mathbf{x}(t)$$

where

$$\mathbf{g} = \begin{bmatrix} \mathbf{1}_m^\intercal & \mathbf{1}_n^\intercal \end{bmatrix}^\intercal \otimes \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}^\intercal.$$

Finally, the *infection peak* is defined as

$$I_{\text{peak}} = \sup_{t \in [0,T]} I(t) = \sup_{t \in [0,T]} \mathbf{g}^\intercal \mathbf{x}(t) \tag{8.12}$$

where $[0, T]$ is a given finite time horizon.

Recovery rate	$\gamma = 1/14$ per 24 hours
Nominal infection rates	$\overline{\beta}_1 = 0.11$ per 24 hours
	$\overline{\beta}_2 = 0.11$ per 24 hours
	$\overline{\beta}_3 = 0.71$ per 24 hours
	$\overline{\beta}_4 = 1.07$ per 24 hours
	$\overline{\beta}_5 = 0.57$ per 24 hours

Table 8.2: Parameters related to local epidemic spread for the example of Figure 8.1.

Example 8.2. Consider the example of mobility network in Figure 8.1 with the mobility parameters given in Table 8.1 and the epidemic parameters given in Table 8.2. Note that the nominal infection rates outside the residences (the "destinations") are assumed higher than the nominal infection rates at the residences (the "origins"). For $T = 1680$ hours (or 10 weeks), we plot the active infected cases $I(t)$ in Figure 8.5 under two circumstances: (i) when there are no restrictions on the operating capacities of all destinations, i.e., $\mathbf{u}(t) = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^\intercal$ for all $t \in [0, T]$, and (ii) when the operating capacities of all destinations are reduced to 50% throughout $[0, T]$, i.e., $\mathbf{u}(t) = \begin{bmatrix} 0.5 & 0.5 & 0.5 \end{bmatrix}^\intercal$ for all $t \in [0, T]$. In Figure 8.5, notice the effect on the number of active infected cases when the operating capacities are reduced. When there are no restrictions on the operating capacities, the infection peak I_{peak} is about 1810 people, whereas, with the restrictions, the peak is about 605 people. In other words, given the mobility and epidemic parameters for the example of Figure 8.1, one can reduce the infection peak by 66.5% through 50% reduction of the operating capacities. ⌟

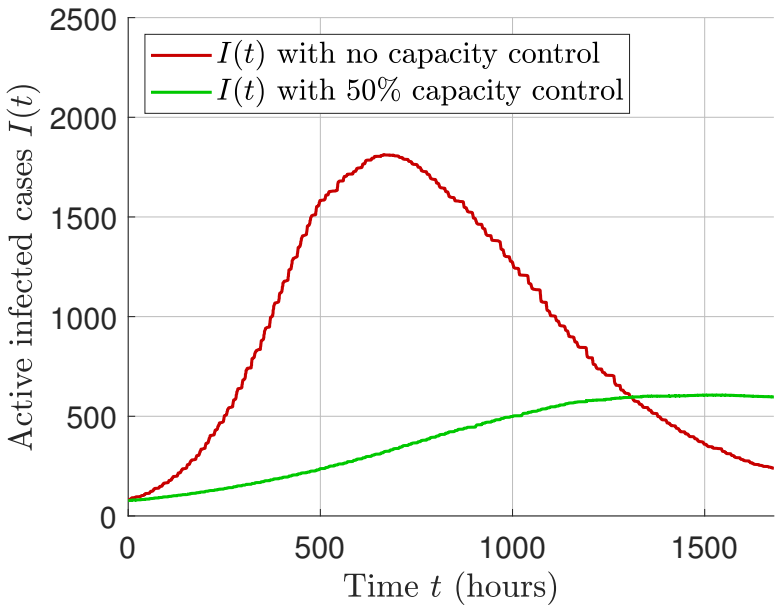


Figure 8.5: Reduction of the number of infected cases $I(t)$ by reducing operating capacities via capacity control policy.

8.4 Optimal Control Policies for Epidemic Mitigation

In this section, we formulate two optimal control problems for epidemic mitigation. The goal of these problems is to maximize the economic activity while keeping the number of active infected cases bounded. First, we formulate the problem of optimal capacity control, which provides an optimal reduction of operating capacities at the destination categories. Second, we formulate the problem of optimal schedule control, which provides an optimal reduction of destination schedules. The problems are solved numerically and simulation results are illustrated.

8.4.1 Optimal capacity control policy

The number of hospitalized cases and deaths due to an epidemic are related to the number of infected cases $I(t)$. A large value of $I(t)$ implies a large number of hospitalizations and loss of lives in the near future. Moreover, if the measures to mitigate the epidemic and limit the number of infected cases are not taken, then the number of hospitalizations may reach a point that could challenge the available medical facilities of the city. As shown in Example 8.2, the number of infected cases can be reduced by reducing the operating capacities of the destinations through the capacity control policies $u_1(t), \ldots, u_p(t)$. However, on the other hand, choosing the values of $u_1(t), \ldots, u_p(t)$ too small can result in a significant reduction of the economic activity $E(t)$, which may result in bankruptcy of businesses and loss of livelihoods. Therefore, our goal in this section is to find optimal capacity control policy that maximize the economic activity under a constraint that the infection peak I_{peak} remains bounded from above.

Suppose a finite time horizon T, a policy horizon $T_u = T/q$ for $q \in \mathbb{N}$, an upper bound $\overline{I} > 0$ on the infection peak, and all the parameters of the model (8.8) be given. Let $\mathbf{u} \in \mathcal{U}$, where $\mathcal{U}$ is the set of admissible capacity control policies $\mathbf{u} : [0, T] \to [0, 1]^p$ such that, for every $t \in [(h-1)T_u, hT_u)$ and $h \in \{1, \ldots, q\}$, $\mathbf{u}(t) = \boldsymbol{\mu}_h$ for some $\boldsymbol{\mu}_h = [\ \mu_1^h \ \cdots \ \mu_p^h\ ]^\intercal \in [0, 1]^p$. Then, the optimal capacity control policy is obtained by solving the following

problem:

$$\begin{aligned} &\underset{\mathbf{u}\in\mathcal{U}}{\text{maximize}} \;\; L(\mathbf{u}) := \frac{1}{T}\int_0^T \mathbf{e}^\intercal \mathbf{x}(t;\mathbf{x}_0,\mathbf{u})dt \\ &\text{subject to} \begin{cases} \dot{\mathbf{x}} = \boldsymbol{\xi}(\mathbf{x}) + \Phi(\mathbf{x},\mathbf{u})\mathbf{x}; \;\; \mathbf{x}(0) = \mathbf{x}_0 \\ I_{\text{peak}}(\mathbf{u}) \leq \overline{I} \end{cases} \end{aligned} \tag{8.13}$$

where $I_{\text{peak}}(\mathbf{u}) = \sup_{t\in[0,T]} \mathbf{g}^\intercal \mathbf{x}(t;\mathbf{x}_0,\mathbf{u})$ is given in (8.12) and the economic activity $E(t) = \mathbf{e}^\intercal \mathbf{x}(t)$ is given in (8.10).

Time horizon	$T = 1680$ hours (10 weeks)
Policy horizon	$T_u = 336$ hours (2 weeks)
Upper bound on infection peak	$\overline{I} = 1000$
Weights of economic importance	$\chi_3 = 0.4$, $\chi_4 = 0.3$, $\chi_5 = 0.3$

Table 8.3: Parameters related to the optimal control problem (8.13) for the example of Figure 8.1.

Example 8.3. Again, consider the example of Figure 8.1 with the mobility and epidemic parameters given in Table 8.1 and 8.2, respectively. Also, consider the parameters in Table 8.3 required by the optimal control problem (8.13). The optimal control problem is solved numerically for a time horizon $T = 1680$ hours (i.e., 10 weeks) using a nonlinear programming solver `fmincon` in MATLAB with interior point algorithm. The solver returns a local minimum $\mathbf{u}_{\text{opt}}(t)$ plotted in Figure 8.6 that satisfies the constraints of (8.13) and is piece-wise constant, where the policy horizon $T_u = 336$ hours (i.e., 2 weeks). In particular, the constraint on the infection peak is satisfied and $I_{\text{peak}} \approx 922$ is less than $\overline{I} = 1000$ as shown in Figure 8.7. Notice that in the beginning the optimal capacity control allows the operating capacities to be around 70-90% of the nominal capacities of destinations. However, as the number of infected cases increase, the value of the optimal capacity control decreases for the next four steps until $8T_u = 1344$ hours (i.e., 8 weeks) to mitigate the infection spread. Then, in the last interval $[8T_u, T]$, the optimal capacity control increases to allow more people visiting the destinations because the infected cases have started to decrease. In Figure 8.8, we plot the economic activity for three cases: (i) when no capacity control policy is implemented, (ii) when the capacity control policy limits the operating capacities to 50% of the nominal capacities of destinations, and (iii) when the optimal capacity control policy $\mathbf{u}_{\text{opt}}(t)$ shown in Figure 8.6 is applied. Notice that the optimal capacity control $\mathbf{u}_{\text{opt}}(t)$ increases the economic activity as compared to the cases when $\mathbf{u}(t) = 0.5\mathbf{1}_3$ while keeping the I_{peak} under the bound $\overline{I}$. ⌟

8.4.2 Optimal schedule control policy

We formulate an epidemic mitigation policy that alters the destination schedules and mobility windows. That is, for every $j \in \mathcal{D}_k$, the destination schedule is altered as

$$\mathbb{S}_j = [a_j, \min(s_k, b_j))$$

where $s_k \in [\underline{s}, 24)$ is the *schedule control* that enforces that all destinations of category k, for $k = 1, \ldots, p$, must be closed after s_k hour, respectively, and $\underline{s} \geq 0$ is the lower bound on s_k. Such a policy limits the spread of infection by reducing the daily amount of time

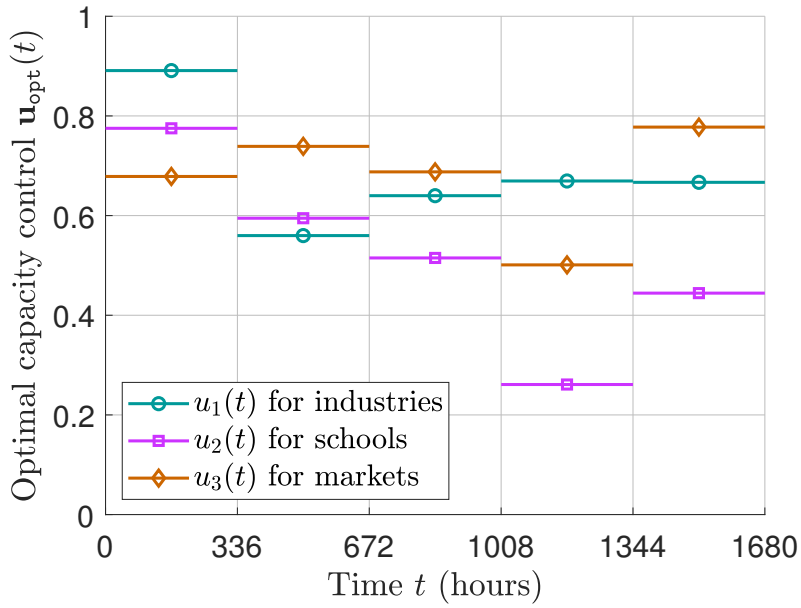


Figure 8.6: Optimal capacity control policy $\mathbf{u}_{opt}$.

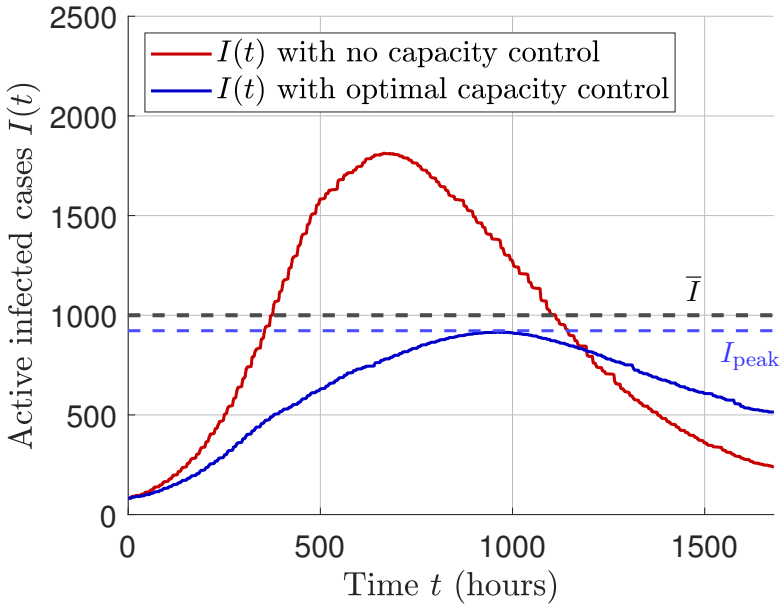


Figure 8.7: Reduction of the number of infected cases $I(t)$ *by controlling the operating capacities via optimal capacity control policy* $\mathbf{u}_{opt}$.

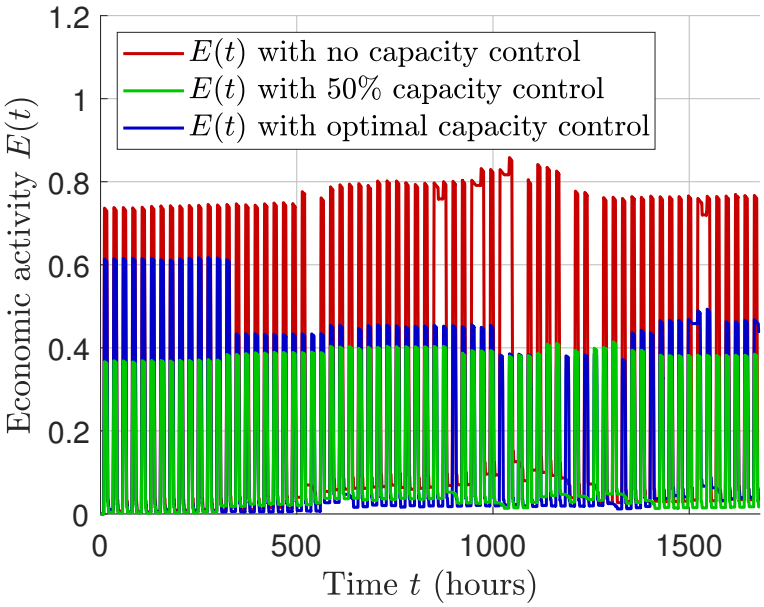


Figure 8.8: Economic activity (i) without capacity control policy, i.e., $\mathbf{u} = \mathbf{1}_3$*, (ii) with 50% capacity control policy* $\mathbf{u} = 0.5\mathbf{1}_3$*, and (iii) with optimal capacity control policy* $\mathbf{u}_{opt}$*.*

people spend at destinations. It also alters the mobility windows

$$\begin{aligned} \mathbb{D}_{ij} &= [t_{ij}, \min(t_{ij} + \tau_{ij}, s_k)) \\ \mathbb{D}_{ji} &= [\min(t_{ji}, s_k), \min(t_{ji}, s_k) + \tau_{ji}). \end{aligned}$$

for $i \in \mathcal{V}_o$ and $j \in \mathcal{D}_k$. That is, people cannot go from i to j after s_k hour and people at j must return to i after s_k hour.

Let $\mathbf{s} = [\ s_1 \quad \cdots \quad s_p\ ]^\intercal$ be the schedule control policy of all destination categories. Then, the problem is to find an optimal $\mathbf{s}$ such that the economic activity is maximized while keeping the infection peak bounded by $\overline{I}$. The schedule control policy complements the capacity control policy obtained by solving (8.13) when there are lower bounds on the capacity control policy. These lower bounds correspond to minimum operating capacities of certain destinations that are required for functioning of the society. This is because some destinations, like hospitals and markets, are essential and their operating capacities cannot be reduced beyond a minimum bound. In other words, for all $k \in \{1, \ldots, p\}$ and $h \in \{1, \ldots, q\}$, we assume constant capacity control policy $\mathbf{u}(t) = \boldsymbol{\mu}$, for all $t \in [0, T]$, where $\boldsymbol{\mu} \in [0, 1]^p$ states the minimum allowed capacity control policy of p destination categories. In the presence of these lower bounds, the problem (8.13) may become infeasible and the infection peak may no longer be bounded. Thus, implementation of optimal schedule control policy $\mathbf{s}$ may help in containing the infections while also allowing economic activity at destinations.

Suppose a finite time horizon T, an upper bound $\overline{I} > 0$ on the infection peak, a constant capacity control policy $\mathbf{u}(t) = \boldsymbol{\mu} \in [0, 1]^p$, $\forall t \in [0, T]$, and the parameters of the model (8.8) are given. Then, the optimal schedule control policy is obtained by solving the following

problem:

$$\begin{aligned} &\underset{\mathbf{s}\in[\underline{s},24)^p}{\text{maximize}} \; L(\mathbf{s}) := \frac{1}{T}\int_0^T \mathbf{e}^\intercal \mathbf{x}(t;\mathbf{x}_0,\mathbf{s})dt \\ &\text{subject to} \begin{cases} \dot{\mathbf{x}} = \boldsymbol{\xi}(\mathbf{x}) + \Phi(\mathbf{x},\mathbf{s})\mathbf{x}; \;\; \mathbf{x}(0) = \mathbf{x}_0 \\ I_{\text{peak}}(\mathbf{s}) \le \overline{I} \end{cases} \end{aligned} \tag{8.14}$$

where $I_{\text{peak}}(\mathbf{s}) = \sup_{t\in[0,T]} \mathbf{g}^\intercal \mathbf{x}(t;\mathbf{x}_0,\mathbf{s})$ is given in (8.12) and the economic activity $E(t) = \mathbf{e}^\intercal \mathbf{x}(t)$ given in (8.10).

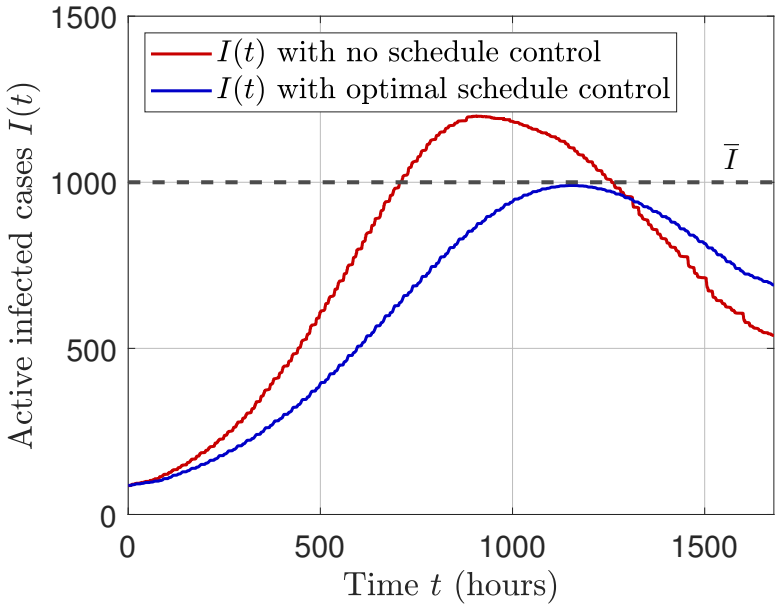


Figure 8.9: Reduction of the number of active infected cases $I(t)$ by controlling destination schedules and mobility windows via optimal schedule control policy $\mathbf{s}_{opt}$.

Example 8.4. Consider again the example of Figure 8.1 and the parameters given in Table 8.1, 8.2, and 8.3, where we change the policy horizon $T_u = T$ since the schedule control policy $\mathbf{s}$ is constant throughout $[0,T]$. Let the capacity control policy $\mathbf{u}(t) = [\ 0.75 \quad 0.75 \quad 0.75\ ]^\intercal$, $\forall t \in [0,T]$, i.e., the operating capacities of destinations should be 75% of their nominal capacities. Then, solving (8.14) numerically by using `fmincon` solver in MATLAB with interior point algorithm, we obtain a local minimum $\mathbf{s}_{\text{opt}} = [\ 14.54 \quad 13.75 \quad 16.10\ ]^\intercal$, which means that the industries must be closed after 14.54 hour (02:32 pm), the schools must be closed after 13.75 hour (01:45 pm), and the markets must be closed after 16.10 hour (04:06 pm). Figure 8.9 shows that the constraint on the infection peak is satisfied. ⌟

8.5 Concluding Remarks

We developed an urban human mobility model on network of origins and destinations that incorporates the process of epidemic spread at each location. The model is described by the flows that transfer people from origins to destinations and back to origins every day. The flows capture the daily patterns of mobility in an urban environment through the gating functions that depend on the destination schedules and mobility windows. At each location, the disease spreads through the interaction of susceptible and infected people, where the infection rate depends on the number of people in that location. We study two optimal control policies, capacity control and schedule control, that mitigate the epidemic spread while maximizing the economic activity at each destination. The optimal capacity control policy maximizes the economic activity by allowing maximum allowable number of people at each destination under the constraint that the infection peak remains bounded. The optimal schedule control policy maximizes the economic activity by allowing maximum allowable time that people spend daily at destinations under the constraint that the infection peak remains bounded. The future investigations include the model predictive control formulation of optimal control problems studied in this chapter. A work in progress is the validation of model with the urban mobility data and regional epidemic statistics, and the development of web interface to illustrate our proof of concept.

Conclusions and Future Outlook

We present general conclusions of our work and discuss the problems that are reserved for the future work.

Part I Aggregated monitoring of large-scale network systems

For the monitoring of large-scale network systems, one has to deal with limited computational and sensing resources. The limitation in computational resources makes the state estimation task infeasible, whereas the limitation in sensing resources renders the network system unobservable. Increasing the computational and sensing resources comes with enormous economic costs and turns out to be impractical under a limited budget. To deal with this issue, we developed techniques in the first part of the thesis for aggregated monitoring of a large-scale network system through a projected network system. The projected network system is a tractable representation obtained by aggregating multiple clusters of nodes in the original network system.

In case the clusters of nodes are specified a priori in a clustered network system, we studied the problem of estimating the average states of clusters from the knowledge of the states of few measured nodes and the model of the projected network system. We provided a minimum-order average observer and stated its design criteria, which is a necessary and sufficient condition for the reconstruction or asymptotic estimation of the average states. Based on the design criteria, we defined the notions of average reconstructability, average observability, and average detectability. Average reconstructability, similar to the usual notion of reconstructability, allows for the reconstruction of current average states of clusters through the average observer that uses the knowledge of past output and input of the network system. On the other hand, average observability allows for the reconstruction of current average states of clusters through the projected network system by taking sufficient derivatives of its output and input. Finally, average detectability allows for the open-loop estimation of the average states through the projected network system and relates to the exponential stability of the average states of clusters. We interpreted these notions graph-theoretically through the clusters to measured nodes, inter-cluster, and intra-cluster topologies of the clustered network system. For average reconstructability and average observability, the measured nodes must span the clustering of the network system. On the other hand, average detectability demands a certain regularity or symmetry of the inter-cluster and intra-cluster topologies. Finally, the design of the average observer under each of these notions is provided.

In case the necessary and sufficient condition for the asymptotic estimation is not satisfied, we devised a methodology for an optimal design of the average observer. The optimal design is achieved by simultaneously stabilizing the average observer and minimizing the effect of average deviation from the average estimation error. We provided sufficient conditions in terms of the induced subgraph topology of unmeasured nodes and clustering of the network system that ensure the stabilizability of the average observer. The efficacy of

the design is illustrated by an application example of a building thermal system.

In case the clusters are not specified in a network system, we presented clustering techniques to facilitate average estimation and illustrated the results by an application example of an SIS epidemic over a large network. Finally, we presented a K-means type clustering technique that facilitates the estimation of state variance of network systems, which is a nonlinear state functional that measures the squared deviation of state trajectories from their average mean. The K-means clustering algorithm allows for the identification of clusters of nodes whose state trajectories eventually converge closer to each other. This enables one to efficiently approximate the state variance through the average states of clusters instead of the states of all nodes. Then, an optimal average observer is employed to estimate the average states and compute the estimated state variance.

The problems that will be studied in the future include optimal sensor placement for asymptotic estimation of average states, cluster selection for combined average observability and average detectability, and variance estimation of multiple clusters. The problems of optimal sensor placement and cluster selection are non-convex, mixed-integer NP-hard optimization problems. Therefore, devising an efficient algorithm is quite challenging. Finally, variance estimation of multiple clusters entails K-means clustering inside each cluster to approximate its state variance. This makes the task computationally challenging.

Part II Modeling and control of epidemics

Mathematical models describing the evolution of epidemics help governments devise policies for the prevention of infection spread in human societies. However, precise identification and integration of the control variables are the crucial parts of an epidemic model. The second part of the thesis, therefore, developed two epidemic models that can be employed to devise control policies related to testing rate and urban human mobility in the event of an epidemic.

Firstly, we developed a five-compartment epidemic model, SIDUR, which considers the testing rate as a control input. The model differentiates between the undetected infected and detected infected people in a population, where the infection is assumed to spread only through the undetected infected people. We estimated and validated the model for the COVID-19 case of France. We devised an epidemic suppression policy called the best-effort strategy for testing (BEST), which provides a minimum testing rate required to stop the growth of the epidemic. The BEST policy is then evaluated by its ability to significantly reduce the cumulative number of deaths and the active number of ICU cases due to COVID-19 in France.

Secondly, we developed a human mobility model in an urban environment that incorporates the process of epidemic spread at each location. We consider mobility between locations of two types, origins and destinations. The origins correspond to residential areas where people reside and the destinations correspond to places that people visit daily. The destinations are divided into multiple categories such as workplaces, schools, markets, etc., whose operating capacities are defined in proportion to their nominal capacities, which correspond to the maximum number of people that can visit each category at a given time during nominal times. In the event of an epidemic, we propose to not only reduce the operating capacities of the destination categories but also limit their schedules of business hours. Thus, we formulate and solve two optimal control problems: optimal capacity control and optimal schedule control. These optimal control problems aim to find a solution that maximizes the economic activity in the destinations while keeping the number of active infected cases bounded.

The future work includes the design of an observer for the SIDUR model, dynamic testing policies on a moving horizon, and the use of the model predictive control framework

and the feedback mechanism for optimal urban human mobility. The observer design for epidemic models is challenging due to the nonlinearity posed by the disease transmission process, which is modeled as the product between the infection rate, susceptible population, and the proportion of infected people. The literature on observer design for epidemic models is scarce, which usually assumes the knowledge of the disease transmission rate to avoid dealing with the nonlinearity. On the other hand, the BEST policy is static that provides a constant testing rate to control the epidemic. However, in the future, we are interested to devise the BEST policy that updates based on the state and parameter estimation of an observer. Similarly, the optimal control policies for urban human mobility are open-loop and solved for the whole time horizon. However, it is more effective to obtain the optimal policies in the framework of model predictive control through a feedback mechanism.

Bibliography

[Abonyi2007] Janos Abonyi and Balazs Feil. Cluster analysis for data mining and system identification. Birkhäuser Basel, 2007.

[Acemoglu2020] Daron Acemoglu, Victor Chernozhukov, Iván Werning and Michael D Whinston. *Optimal targeted lockdowns in a multi-group SIR model*. NBER Working Paper No. 27102, 2020.

[Aguilar2017] Cesar O Aguilar and Bahman Gharesifard. *Almost equitable partitions and new necessary conditions for network controllability*. Automatica, vol. 80, pages 25–31, 2017.

[Albert1999] Réka Albert, Hawoong Jeong and Albert-László Barabási. *Diameter of the world-wide web*. nature, vol. 401, no. 6749, pages 130–131, 1999.

[Aldeen1994] M Aldeen and Hieu Trinh. *Observing a subset of the states of linear systems*. IEE Proceedings-Control Theory and Applications, vol. 141, no. 3, pages 137–144, 1994.

[Aldeen1999] M Aldeen and Hieu Trinh. *Reduced-order linear functional observer for linear systems*. IEE Proceedings-Control Theory and Applications, vol. 146, no. 5, pages 399–405, 1999.

[Allen2008] Linda JS Allen. *An introduction to stochastic epidemic models*. In Mathematical epidemiology, pages 81–130. Springer, 2008.

[Alvarez2020] Fernando E Alvarez, David Argente and Francesco Lippi. *A simple planning problem for COVID-19 lockdown*. NBER Working Paper No. 26981, 2020.

[Anastassopoulou2020] C. Anastassopoulou, L. Russo, A. Tsakris and C. Siettos. *Data-based analysis, modelling and forecasting of the COVID-19 outbreak*. PLoS One, vol. 15, no. e0230405, 2020.

[Anderson1975] Brian Anderson, N Bose and E Jury. *Output feedback stabilization and related problems-solution via decision methods*. IEEE Transactions on Automatic control, vol. 20, no. 1, pages 53–66, 1975.

[Anderson1989] BDO Anderson and Y Liu. *Controller reduction: concepts and approaches*. IEEE Transactions on Automatic Control, vol. 34, no. 8, pages 802–812, 1989.

[Antoulas2005] Athanasios C Antoulas. Approximation of large-scale dynamical systems. Philadelphia, PA, USA: SIAM, 2005.

[Antoulas2010] Athanasios C Antoulas, Christopher A Beattie and Serkan Gugercin. *Interpolatory model reduction of large-scale dynamical systems*. In Efficient modeling and control of large-scale systems, pages 3–58. Springer, 2010.

[Antsaklis2006] Panos J Antsaklis and Anthony N Michel. Linear systems. Springer Birkhauser Boston, 2006.

[Aoki1968] Masanao Aoki. *Control of large-scale dynamic systems by aggregation*. IEEE Trans. Automat. Contr., vol. 13, no. 3, pages 246–253, 1968.

[Arino2003] Julien Arino and P Van den Driessche. *A multi-city epidemic model*. Mathematical Population Studies, vol. 10, no. 3, pages 175–193, 2003.

[Arrow1958] Kenneth J Arrow and Maurice McManus. *A note on dynamic stability*. Econometrica: Journal of the Econometric Society, pages 448–454, 1958.

[Atay2017] Fatihcan M Atay and Lavinia Roncoroni. *Lumpability of linear evolution equations in Banach spaces*. Evolution Equations and Control Theory, vol. 6, no. 1, pages 15–34, 2017.

[Baden2020] Lindsey R. Baden and Eric J. Rubin. *COVID-19 — The Search for Effective Therapy*. New England Journal of Medicine, vol. 382, no. 19, pages 1851–1852, 2020.

[Balcan2010] Duygu Balcan, Bruno Gonçalves, Hao Hu, José J Ramasco, Vittoria Colizza and Alessandro Vespignani. *Modeling the spatial spread of infectious diseases: The GLobal Epidemic and Mobility computational model*. Journal of computational science, vol. 1, no. 3, pages 132–145, 2010.

[Barabási2003] Albert-László Barabási and Eric Bonabeau. *Scale-free networks*. Scientific american, vol. 288, no. 5, pages 60–69, 2003.

[Barabási2009] Albert-László Barabási. *Scale-free networks: a decade and beyond*. science, vol. 325, no. 5939, pages 412–413, 2009.

[Bartle1964] Robert Gardner Bartle. The elements of real analysis, volume 2. Wiley: New York, 1964.

[Bejarano2009] Francisco J Bejarano, Leonid Fridman and Alexander Poznyak. *Unknown input and state estimation for unobservable systems*. SIAM Journal Control Optimization, vol. 48, no. 2, pages 1155–1178, 2009.

[Bejarano2011] Francisco Javier Bejarano, Thierry Floquet, Wilfrid Perruquetti and Gang Zheng. *Observability and detectability analysis of singular linear systems with unknown inputs*. In 2011 50th IEEE Conference on Decision and Control and European Control Conference, pages 4005–4010, 2011.

[Berger2020] D. Berger, K. Herkenhoff and S. Mongey. *An SEIR Infectious Disease Model with Testing and Conditional Quarantine*. NBER Working Paper No. 26901, 2020.

[Bhatia1997] Rajendra Bhatia. Matrix analysis, volume 169. Springer-Verlag New York, 1997.

[Boccaletti2006] Stefano Boccaletti, Vito Latora, Yamir Moreno, Martin Chavez and D-U Hwang. *Complex networks: Structure and dynamics*. Physics reports, vol. 424, no. 4-5, pages 175–308, 2006.

[Bodeau2000] Jérôme Bodeau, Gaël Riboulet and Thierry Roncalli. *Non-uniform grids for pde in finance*. Available at SSRN 1031941, 2000.

[Boduch2009] Michael Boduch and Warren Fincher. *Standards of human comfort: relative and absolute*. 2009. Available online: http://hdl.handle.net/2152/13980.

[Bongiorno Jr1968] JJ Bongiorno Jr and DC Youla. *On observers in multi-variable control systems*. International Journal of Control, vol. 8, no. 3, pages 221–243, 1968.

[Boyd1994] Stephen Boyd, Laurent El Ghaoui, Eric Feron and Venkataramanan Balakrishnan. Linear matrix inequalities in system and control theory. SIAM, 1994.

[Boyd2004] Stephen Boyd and Lieven Vandenberghe. Convex optimization. Cambridge university press, 2004.

[Brauer2012] Fred Brauer, Carlos Castillo-Chavez and Carlos Castillo-Chavez. Mathematical models in population biology and epidemiology, volume 2. Springer, 2012.

[Brotherhood2020] Luiz Brotherhood, Philipp Kircher, Cezar Santos and Michèle Tertilt. *An economic model of the COVID-19 epidemic: The importance of testing and age-specific policies*. CESifo Working Paper No. 8316, 2020.

[Bueno2012] Bruno Bueno, Leslie Norford, Grégoire Pigeon and Rex Britter. *A resistance-capacitance network model for the analysis of the interactions between the energy performance of buildings and the urban climate*. Building and Environment, vol. 54, pages 116–125, 2012.

[Bullo2018] Francesco Bullo. Lectures on network systems. CreateSpace, 1 edition, 2018. With contributions by J. Cortes, F. Dörfler, and S. Martinez. Available online: http://motion.me.ucsb.edu/book-lns.

[Bungartz2010] Hans-Joachim Bungartz, Miriam Mehl, Tobias Neckel and Tobias Weinzierl. *The PDE framework Peano applied to fluid dynamics: an efficient implementation of a parallel multiscale fluid dynamics solver on octree-like adaptive Cartesian grids*. Computational Mechanics, vol. 46, no. 1, pages 103–114, 2010.

[Burer2012] Samuel Burer and Adam N Letchford. *Non-convex mixed-integer nonlinear programming: A survey*. Surveys in Operations Research and Management Science, vol. 17, no. 2, pages 97–106, 2012.

[Campbell2009] Stephen L Campbell and Carl D Meyer. Generalized inverses of linear transformations. SIAM, 2009.

[Cao2020] Xuetao Cao. *COVID-19: Immunopathology and its implications for therapy.* Nature reviews immunology, vol. 20, no. 5, pages 269–270, 2020.

[Carlson1968] David Carlson. *A new criterion for H-stability of complex matrices.* Linear Algebra and its Applications, vol. 1, no. 1, pages 59–64, 1968.

[Casella2020] Francesco Casella. *Can the COVID-19 epidemic be controlled on the basis of daily test reports?* IEEE Control Systems Letters, vol. 5, no. 3, pages 1079–1084, 2020.

[Charpentier2020] A. Charpentier, R. Elie, M. Lauriére and V.C. Tran. *COVID-19 pandemic control: Balancing detection policy and lockdown intervention under ICU sustainability.* arXiv:2005.06526v3, 2020.

[Cheng2017] Xiaodong Cheng, Yu Kawano and Jacquelien MA Scherpen. *Reduction of second-order network systems with structure preservation.* IEEE Transactions on Automatic Control, vol. 62, no. 10, pages 5026–5038, 2017.

[Cheng2018a] Xiaodong Cheng, Yu Kawano and Jacquelien MA Scherpen. *Model Reduction of Multiagent Systems Using Dissimilarity-Based Clustering.* IEEE Transactions on Automatic Control, vol. 64, no. 4, pages 1663–1670, 2018.

[Cheng2018b] Xiaodong Cheng and Jacquelien MA Scherpen. *Clustering approach to model order reduction of power networks with distributed controllers.* Advances in Computational Mathematics, vol. 44, no. 6, pages 1917–1939, 2018.

[Cheng2021] Xiaodong Cheng and JMA Scherpen. *Model Reduction Methods for Complex Network Systems.* Annual Review of Control, Robotics, and Autonomous Systems, vol. 4, no. 1, page null, 2021.

[Chowell2003] G. Chowell, P. Fenimore, M. Castillo-Garsow and C. Castillo-Chavez. *SARS outbreaks in Ontario, Hong Kong and Singapore: The role of diagnosis and isolation as a control mechanism.* Journal of Theoretical Biology, vol. 224, no. 1, pages 1–8, 2003.

[Clerc2002] Maurice Clerc and James Kennedy. *The particle swarm-explosion, stability, and convergence in a multidimensional complex space.* IEEE transactions on Evolutionary Computation, vol. 6, no. 1, pages 58–73, 2002.

[Cobb1984] Daniel Cobb. *Controllability, observability, and duality in singular systems.* IEEE Transactions on Automatic Control, vol. 29, no. 12, pages 1076–1082, 1984.

[Colizza2008] Vittoria Colizza and Alessandro Vespignani. *Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: Theory and simulations.* Journal of theoretical biology, vol. 251, no. 3, pages 450–467, 2008.

[Corfmat1976] Jean-Pierre Corfmat and A Stephen Morse. *Decentralized control of linear multivariable systems.* Automatica, vol. 12, no. 5, pages 479–495, 1976.

[Coxson1984] Pamela G Coxson. *Lumpability and observability of linear systems.* Journal of Mathematical Analysis and Applications, vol. 99, no. 2, pages 435–446, 1984.

[Darouach2000] Mohamed Darouach. *Existence and design of functional observers for linear systems.* IEEE Transactions on Automatic Control, vol. 45, no. 5, pages 940–943, 2000.

[Darouach2019] Mohamed Darouach and Tyrone Fernando. *On the existence and design of functional observers.* IEEE Transactions on Automatic Control, vol. 65, no. 6, pages 2751–2759, 2019.

[Day2020] Michael Day. *COVID-19: Identifying and isolating asymptomatic people helped eliminate virus in Italian village.* BMJ: British Medical Journal (Online), vol. 368, 2020.

[de Walque2020] Damien de Walque, Jed Friedman, Roberta Gatti and Aaditya Mattoo. *How two tests can help contain COVID-19 and revive the economy.* World Bank Research & Policy Briefs, no. 29, 2020.

[Deconinck2016] An-Heleen Deconinck and Staf Roels. *Comparison of characterisation methods determining the thermal resistance of building components from onsite measurements.* Energy and Buildings, vol. 130, pages 309–320, 2016.

[Della Rossa2020] Fabio Della Rossa, Davide Salzano, Anna Di Meglio, Francesco De Lellis, Marco Coraggio, Carmela Calabrese, Agostino Guarino, Ricardo Cardona-Rivera, Pietro De Lellis, Davide Liuzza*et al. A network model of Italy shows that intermittent regional strategies can alleviate the COVID-19 epidemic.* Nature Communications, vol. 11, no. 1, pages 1–9, 2020.

[Deng2010] K. Deng, P. Barooah, P. G. Mehta and S. P. Meyn. *Building thermal model reduction via aggregation of states.* In Proceedings of the 2010 American Control Conference, pages 5118–5123, 2010.

[Deng2014] Kun Deng, Siddharth Goyal, Prabir Barooah and Prashant G Mehta. *Structure-preserving model reduction of nonlinear building thermal models.* Automatica, vol. 50, no. 4, pages 1188–1195, 2014.

[Derbez2014] Mickaël Derbez, Bruno Berthineau, Valérie Cochet, Murielle Lethrosne, Cécile Pignon, Jacques Riberon and Séverine Kirchner. *Indoor air quality and comfort in seven newly built, energy-efficient houses in France.* Building and Environment, vol. 72, pages 173–187, 2014.

[Dieci2001] Luca Dieci and Alessandra Papini. *Conditioning of the exponential of a block triangular matrix.* Numerical Algorithms, vol. 28, no. 1, pages 137–150, 2001.

[Dion2003] Jean-Michel Dion, Christian Commault and Jacob Van der Woude. *Generic properties and control of linear structured systems: a survey.* Automatica, vol. 39, no. 7, pages 1125–1144, 2003.

[Ducrot2020] Arnaud Ducrot, P Magal, Thanh Nguyen and GF Webb. *Identifying the number of unreported cases in SIR epidemic models.* Mathematical medicine and biology: A journal of the IMA, vol. 37, no. 2, pages 243–261, 2020.

[Egerstedt2012] Magnus Egerstedt, Simone Martini, Ming Cao, Kanat Camlibel and Antonio Bicchi. *Interacting with networks: How does structure relate to controllability in single-leader, consensus networks?* IEEE Control Systems Magazine, vol. 32, no. 4, pages 66–73, 2012.

[Eichenbaum2020] Martin S Eichenbaum, Sergio Rebelo and Mathias Trabandt. *The Macroeconomics of Testing and Quarantining.* NBER Working Paper No. 27104, 2020.

[Ely2020] Jeffrey Ely, Andrea Galeotti and Jakub Steiner. *Optimal Test Allocation.* Rapport technique, Mimeo, 2020.

[FDA2020] *Emergency use authorization (EUA) summary: COVID-19 RT-PCR test.* Rapport technique, US Food and Drug Administration, 2020.

[Ferguson2020] Neil Ferguson, Daniel Laydon, Gemma Nedjati Gilani, Natsuko Imai, Kylie Ainslie, Marc Baguelin, Sangeeta Bhatia, Adhiratha Boonyasiri, ZULMA Cucunuba Perez, Gina Cuomo-Dannenburg*et al. Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand.* Imperial College London (16-03-2020), 2020.

[Fernando2010a] Tyrone Fernando, Less Jennings and Hieu Trinh. *Functional observability.* In 2010 Fifth International Conference on Information and Automation for Sustainability, pages 421–423. IEEE, 2010.

[Fernando2010b] Tyrone Lucius Fernando, Hieu Minh Trinh and Les Jennings. *Functional observability and the design of minimum order linear functional observers.* IEEE Transactions on Automatic Control, vol. 55, no. 5, pages 1268–1273, 2010.

[Frias-Martinez2011] Enrique Frias-Martinez, Graham Williamson and Vanessa Frias-Martinez. *An agent-based model of epidemic spread using human mobility and social network information.* In 2011 IEEE 3rd International Conference on Privacy, Security, Risk and Trust and 2011 IEEE 3rd International Conference on Social Computing, pages 57–64, 2011.

[Gartner2002] Nathan H Gartner and Chronis Stamatiadis. *Arterial-based control of traffic flow in urban grid networks.* Mathematical and computer modelling, vol. 35, no. 5-6, pages 657–671, 2002.

[Giordano2020] G. Giordano, F. Blanchini, R. Bruno, P. Colaneri, A. Di Filippo, A. Di Matteo and M. Colaneri. *Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy.* Nature medicine, vol. 26, pages 855–860, June 2020.

[Glover2020] Andrew Glover, Jonathan Heathcote, Dirk Krueger and José-Víctor Ríos-Rull. *Health versus wealth: On the distributional effects of controlling a pandemic.* NBER Working Paper No. 27046, 2020.

[Godsil2001] Chris Godsil and Gordon Royle. Algebraic graph theory. New York: Springer-Verlag, 2001.

[Goh2006] Hock Guan Goh, Moh Lim Sim and Hong Tat Ewe. *Energy efficient routing for wireless sensor networks with grid topology.* In International Conference on Embedded and Ubiquitous Computing, pages 834–843. Springer, 2006.

[Gopinath1971] B Gopinath. *On the control of linear multiple input-output systems.* Bell System Technical Journal, vol. 50, no. 3, pages 1063–1081, 1971.

[Grassberger1983] Peter Grassberger. *On the critical behavior of the general epidemic process and dynamical percolation.* Mathematical Biosciences, vol. 63, no. 2, pages 157–172, 1983.

[Gugercin2008] Serkan Gugercin, Athanasios C Antoulas and Cchristopher Beattie. $\mathcal{H}_2$ *model reduction for large-scale linear dynamical systems.* SIAM journal on matrix analysis and applications, vol. 30, no. 2, pages 609–638, 2008.

[Hache2017] Emmanuel Hache, Déborah Leboullenger and Valérie Mignon. *Beyond average energy consumption in the French residential housing market: A household classification approach.* Energy Policy, vol. 107, pages 82–95, 2017.

[Hale2020a] Thomas Hale, Noam Angrist, Beatriz Kira, Anna Petherick, Toby Phillips and Samuel Webster. *Variation in government responses to COVID-19.* Blavatnik School of Government Working Paper BSG-WP-2020/032, 2020.

[Hale2020b] Thomas Hale, Samuel Webster, Anna Petherick, Toby Phillips and Beatriz Kira. *Oxford COVID-19 Government Response Tracker*, 2020. Blavatnik School of Government.

[Hethcote2000] H.W. Hethcote. *The mathematics of infectious diseases.* SIAM Review, vol. 42, no. 1, pages 599–653, 2000.

[Horn2013] Roger A Horn and Charles R Johnson. Matrix analysis. Cambridge University Press, 2 edition, 2013.

[Ishizaki2014] Takayuki Ishizaki, Kenji Kashima, Jun-Ichi Imura and Kazuyuki Aihara. *Model reduction and clusterization of large-scale bidirectional networks.* IEEE Transactions on Automatic Control, vol. 59, no. 1, pages 48–63, 2014.

[Ishizaki2015] Takayuki Ishizaki, Kenji Kashima, Antoine Girard, Jun-ichi Imura, Luonan Chen and Kazuyuki Aihara. *Clustered model reduction of positive directed networks.* Automatica, vol. 59, pages 238–247, 2015.

[Jennings2011] Les S Jennings, Tyrone Lucius Fernando and Hieu Minh Trinh. *Existence conditions for functional observability from an eigenspace perspective.* IEEE transactions on automatic control, vol. 56, no. 12, pages 2957–2961, 2011.

[Jeong2000] Hawoong Jeong, Bálint Tombor, Réka Albert, Zoltan N Oltvai and A-L Barabási. *The large-scale organization of metabolic networks.* Nature, vol. 407, no. 6804, pages 651–654, 2000.

[Ji2007] Meng Ji and Magnus Egerstedt. *Observability and estimation in distributed sensor networks.* In 46th IEEE Conference on Decision and Control, pages 4221–4226, 2007.

[Kailath1980] Thomas Kailath. Linear systems. Prentice-Hall Englewood Cliffs, NJ, 1980.

[Kalman1960] R. E. Kalman. *A New Approach to Linear Filtering and Prediction Problems.* Journal of Basic Engineering, vol. 82, no. 1, pages 35–45, 1960.

[Kalman1961] Rudolph E Kalman and Richard S Bucy. *New results in linear filtering and prediction theory.* Journal of Basic Engineering, vol. 83, no. 1, pages 95–108, 1961.

[Kalman1963] Rudolf Emil Kalman. *Mathematical description of linear dynamical systems.* Journal of the Society for Industrial and Applied Mathematics, Series A: Control, vol. 1, no. 2, pages 152–192, 1963.

[Kautsky1985] Jaroslav Kautsky, Nancy K Nichols and Paul Van Dooren. *Robust pole assignment in linear state feedback.* International Journal of control, vol. 41, no. 5, pages 1129–1155, 1985.

[Kazantzis2001] Nikolaos Kazantzis and Costas Kravaris. *Discrete-time nonlinear observer design using functional equations.* Systems & Control Letters, vol. 42, no. 2, pages 81–94, 2001.

[Keeling2011] Matt J Keeling and Pejman Rohani. Modeling infectious diseases in humans and animals. Princeton university press, 2011.

[Kennedy1995] James Kennedy and Russell Eberhart. *Particle swarm optimization.* In Proceedings of the International Conference on Neural Networks (ICNN), volume 4, pages 1942–1948. IEEE, 1995.

[Kermack1927] William Ogilvy Kermack and Anderson G McKendrick. *A contribution to the mathematical theory of epidemics.* Proceedings of the Royal Society of London Series A, vol. 115, no. 772, pages 700–721, 1927.

[Khanafer2016] Ali Khanafer, Tamer Başar and Bahman Gharesifard. *Stability of epidemic models over directed graphs: A positive systems approach.* Automatica, vol. 74, pages 126–134, 2016.

[Khargonekar1988] Pramod P Khargonekar, Ian R Petersen and Mario A Rotea. *H/-sub infinity/-optimal control with state-feedback.* IEEE Transactions on Automatic Control, vol. 33, no. 8, pages 786–788, 1988.

[Kimura1977] Hidenori Kimura. *A further result on the problem of pole assignment by output feedback.* IEEE Transactions on Automatic Control, vol. 22, no. 3, pages 458–463, 1977.

[Kleinberg2007] Jon Kleinberg. *The wireless epidemic.* Nature, vol. 449, no. 7160, pages 287–288, 2007.

[Köhler2020] Johannes Köhler, Lukas Schwenkel, Anna Koch, Julian Berberich, Patricia Pauli and Frank Allgöwer. *Robust and optimal predictive control of the COVID-19 outbreak.* arXiv:2005.03580, 2020.

[Kravaris2011] Costas Kravaris. *Functional observers for nonlinear systems.* In 2011 9th IEEE International Conference on Control and Automation (ICCA), pages 501–506, 2011.

[Kravaris2013] Costas Kravaris, Juergen Hahn and Yunfei Chu. *Advances and selected recent developments in state and parameter estimation.* Computers & chemical engineering, vol. 51, pages 111–123, 2013.

[Kravaris2016] Costas Kravaris. *Functional observers for nonlinear systems.* IFAC-PapersOnLine, vol. 49, no. 18, pages 505–510, 2016.

[Lauer2020] Stephen A Lauer, Kyra H Grantz, Qifang Bi, Forrest K Jones, Qulu Zheng, Hannah R Meredith, Andrew S Azman, Nicholas G Reich and Justin Lessler. *The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application.* Annals of internal medicine, vol. 172, no. 9, pages 577–582, 2020.

[Lévy2018] Jean-Pierre Lévy and Fateh Belaïd. *The determinants of domestic energy consumption in France: Energy modes, habitat, households and life cycles.* Renewable and Sustainable Energy Reviews, vol. 81, pages 2104–2114, 2018.

[Li2010] Xiang Li, Christophe Claramunt and Cyril Ray. *A grid graph-based model for the analysis of 2D indoor spaces.* Computers, Environment and Urban Systems, vol. 34, no. 6, pages 532–540, 2010.

[Lin1974] Ching-Tai Lin. *Structural controllability.* IEEE Transactions on Automatic Control, vol. 19, no. 3, pages 201–208, 1974.

[Lin2020] Q. Lin, S. Zhao, D. Gao, Y. Lou, S. Yang, S. Musa, M. Wang, Y. Cai, W. Wang, L. Yang and D. He. *A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan, China with individual reaction and governmental action.* International Journal of Infectious Diseases, vol. 93, pages 211–216, 2020.

[Liu2011] Yang-Yu Liu, Jean-Jacques Slotine and Albert-László Barabási. *Controllability of complex networks.* nature, vol. 473, no. 7346, pages 167–173, 2011.

[Liu2020] Zhihua Liu, Pierre Magal, Ousmane Seydi and Glenn Webb. *Understanding unreported cases in the COVID-19 epidemic outbreak in Wuhan, China, and the importance of major public health interventions.* Biology, vol. 9, no. 3, page 50, 2020.

[Ljung1999] L. Ljung. System identification: Theory for the user. Prentice Hall PTR, Upper Saddle River, NJ-USA, 2 edition, 1999.

[Loayza2020] Norman V Loayza and Steven Pennings. *Macroeconomic policy in the time of COVID-19: A primer for developing countries.* World Bank Research & Policy Briefs, no. 28, 2020.

[Luenberger1964] David G Luenberger. *Observing the state of a linear system.* IEEE transactions on military electronics, vol. 8, no. 2, pages 74–80, 1964.

[Luenberger1966] David Luenberger. *Observers for multivariable systems.* IEEE Transactions on Automatic Control, vol. 11, no. 2, pages 190–197, 1966.

[Luenberger1971] David Luenberger. *An introduction to observers.* IEEE Transactions on Automatic Control, vol. 16, no. 6, pages 596–602, 1971.

[Maasoumy2013] Mehdi Maasoumy, Barzin Moridian, Meysam Razmara, Mahdi Shahbakhti and Alberto Sangiovanni-Vincentelli. *Online Simultaneous State Estimation and Parameter Adaptation for Building Predictive Control.* In ASME 2013 Dynamic Systems and Control Conference, volume 2, 2013.

[Martin2019] Nicolas Martin, Paolo Frasca, Takayuki Ishizaki, Jun-Ichi Imura and Carlos Canudas-de-Wit. *The price of connectedness in graph partitioning problems.* In 2019 18th European Control Conference (ECC), pages 2313–2318. IEEE, 2019.

[Martin2020] Nicolas Martin, Paolo Frasca and Carlos Canudas-de Wit. *Subgraph detection for average detectability of LTI systems.* IEEE Transactions on Network Science and Engineering, vol. 7, no. 4, pages 2787–2798, 2020.

[Martini2010] Simone Martini, Magnus Egerstedt and Antonio Bicchi. *Controllability analysis of multi-agent systems using relaxed equitable partitions.* International Journal of Systems, Control and Communications, vol. 2, no. 1-3, pages 100–121, 2010.

[Massonis2020] Gemma Massonis, Julio R Banga and Alejandro F Villaverde. *Structural Identifiability and Observability of Compartmental Models of the COVID-19 Pandemic.* arXiv:2006.14295, 2020.

[Mathews1994] EH Mathews, PG Richards and C Lombard. *A first-order thermal model for building design.* Energy and Buildings, vol. 21, no. 2, pages 133–145, 1994.

[McFarlane1990] Duncan McFarlane, Keith Glover and M Vidyasagar. *Reduced-order controller design using coprime factor model reduction.* IEEE Transactions on Automatic Control, vol. 35, no. 3, pages 369–373, 1990.

[Mei2017] Wenjun Mei, Shadi Mohagheghi, Sandro Zampieri and Francesco Bullo. *On the dynamics of deterministic epidemic propagation over networks.* Annual Reviews in Control, vol. 44, pages 116–128, 2017.

[Monshizadeh2014] Nima Monshizadeh, Harry L Trentelman and M Kanat Camlibel. *Projection-based model reduction of multi-agent systems using graph partitions.* IEEE Transactions on Control of Network Systems, vol. 1, no. 2, pages 145–154, 2014.

[Morato2020] Marcelo M Morato, Igor ML Pataro, Marcus V da Costa and Julio E Normey-Rico. *A parametrized nonlinear predictive control strategy for relaxing COVID-19 social distancing measures in Brazil.* arXiv:2007.09686, 2020.

[Murdoch1973] P Murdoch. *Observer design for a linear functional of the state vector.* IEEE Transactions on Automatic Control, vol. 18, no. 3, pages 308–310, 1973.

[Murdock2006] Dustin A Murdock, Jose E Ramos Torres, Jeffrey J Connors and Robert D Lorenz. *Active thermal control of power electronic modules.* IEEE Transactions on Industry Applications, vol. 42, no. 2, pages 552–558, 2006.

[Murota1987] Kazuo Murota. Systems analysis by graphs and matroids: Structural solvability and controllability. Springer, Berlin, Heidelberg, 1987.

[Murota2010] Kazuo Murota. Matrices and matroids for systems analysis. Springer, Berlin, Heidelberg, 2010.

[Nadini2020] Matthieu Nadini, Lorenzo Zino, Alessandro Rizzo and Maurizio Porfiri. *A multi-agent model to study epidemic spreading and vaccination strategies in an urban-like environment.* Applied Network Science, vol. 5, no. 1, pages 1–30, 2020.

[Nelles2001] Oliver Nelles. Nonlinear system identification: From classical approaches to neural networks and fuzzy models. Springer-Verlag, Berlin Heidelberg, 2001.

[Newman2002] Mark EJ Newman. *Spread of epidemic disease on networks.* Physical review E, vol. 66, no. 1, page 016128, 2002.

[Niazi2020a] Muhammad Umar B. Niazi, Carlos Canudas-de-Wit and Alain Kibangou. *Thermal Monitoring of Buildings by Aggregated Temperature Estimation.* In IFAC 2020 - IFAC World Congress 2020, Berlin, Germany, 2020.

[Niazi2020b] Muhammad Umar B Niazi, Carlos Canudas-de-Wit and Alain Y Kibangou. *Average state estimation in large-scale clustered network systems.* IEEE Transactions on Control of Network Systems, vol. 7, no. 4, pages 1736–1745, 2020.

[Nikitin2021] Denis Nikitin, Carlos Canudas de Wit and Paolo Frasca. *Control of Average and Deviation in Large-Scale Linear Networks.* IEEE Transactions on Automatic Control, 2021. in press.

[Nowzari2016] Cameron Nowzari, Victor M Preciado and George J Pappas. *Analysis and control of epidemics: A survey of spreading processes on complex networks.* IEEE Control Systems Magazine, vol. 36, no. 1, pages 26–46, 2016.

[Nowzari2017] Cameron Nowzari, Victor M Preciado and George J Pappas. *Optimal resource allocation for control of networked epidemic models.* IEEE Transactions on Control of Network Systems, vol. 4, no. 2, pages 159–169, 2017.

[Oh2020] Juhwan Oh, Jong-Koo Lee, Dan Schwarz, Hannah L Ratcliffe, Jeffrey F Markuns and Lisa R Hirschhorn. *National response to COVID-19 in the Republic of Korea and lessons learned for other countries.* Health Systems & Reform, vol. 6, no. 1, page e1753464, 2020.

[Oldewurtel2010] Frauke Oldewurtel, Alessandra Parisio, Colin N Jones, Manfred Morari, Dimitrios Gyalistras, Markus Gwerder, Vanessa Stauch, Beat Lehmann and Katharina Wirth. *Energy efficient building climate control using stochastic model predictive control and weather predictions.* In Proceedings of the 2010 American Control Conference, pages 5100–5105, 2010.

[Olfati-Saber2007] Reza Olfati-Saber, J Alex Fax and Richard M Murray. *Consensus and cooperation in networked multi-agent systems.* Proceedings of the IEEE, vol. 95, no. 1, pages 215–233, 2007.

[Olivier2020] Laurentz E Olivier, Stefan Botha and Ian K Craig. *Optimized lockdown strategies for curbing the spread of COVID-19: A South African case study.* arXiv:2006.16379, 2020.

[Ostrowski1962] Alexander Ostrowski and Hans Schneider. *Some theorems on the inertia of general matrices.* J. Math. Anal. Appl, vol. 4, no. 1, pages 72–84, 1962.

[Pappalardo2015] Luca Pappalardo, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti and Albert-László Barabási. *Returners and explorers dichotomy in human mobility.* Nature communications, vol. 6, no. 1, pages 1–8, 2015.

[Paré2018a] Philip E Paré, Carolyn L Beck and Angelia Nedić. *Epidemic processes over time-varying networks.* IEEE Transactions on Control of Network Systems, vol. 5, no. 3, pages 1322–1334, 2018.

[Paré2018b] Philip E Paré, Ji Liu, Carolyn L Beck, Barrett E Kirwan and Tamer Başar. *Analysis, estimation, and validation of discrete-time epidemic processes.* IEEE Transactions on Control Systems Technology, vol. 28, no. 1, pages 79–93, 2018.

[Paré2020] Philip E Paré, Carolyn L Beck and Tamer Başar. *Modeling, estimation, and analysis of epidemics over networks: An overview.* Annual Reviews in Control, 2020.

[Pastor-Satorras2001a] Romualdo Pastor-Satorras and Alessandro Vespignani. *Epidemic dynamics and endemic states in complex networks.* Physical Review E, vol. 63, no. 6, page 066117, 2001.

[Pastor-Satorras2001b] Romualdo Pastor-Satorras and Alessandro Vespignani. *Epidemic spreading in scale-free networks.* Physical review letters, vol. 86, no. 14, page 3200, 2001.

[Pastor-Satorras2015] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem and Alessandro Vespignani. *Epidemic processes in complex networks.* Reviews of modern physics, vol. 87, no. 3, page 925, 2015.

[Perkins2020] Alex Perkins and Guido Espana. *Optimal control of the COVID-19 pandemic with non-pharmaceutical interventions.* medRxiv 2020.04.22.20076018, 2020.

[Pezzutto2020] Matthias Pezzutto, Nicolas Bono Rossello, Luca Schenato and Emanuele Garone. *Smart Testing and Selective Quarantine for the Control of Epidemics.* arXiv:2007.15412, 2020.

[Piguillem2020] Facundo Piguillem and Liyan Shi. *Optimal COVID-19 quarantine and testing policies.* CEPR Discussion Paper No. DP14613, 2020.

[Poletto2013] Chiara Poletto, Michele Tizzoni and Vittoria Colizza. *Human mobility and time spent at destination: Impact on spatial epidemic spreading.* Journal of theoretical biology, vol. 338, pages 41–58, 2013.

[Poli2007] Riccardo Poli, James Kennedy and Tim Blackwell. *Particle swarm optimization: An overview.* Swarm intelligence, vol. 1, no. 1, pages 33–57, 2007.

[Ramallo-González2013] Alfonso P Ramallo-González, Matthew E Eames and David A Coley. *Lumped parameter models for building thermal modelling: An analytic approach to simplifying complex multi-layered constructions.* Energy and Buildings, vol. 60, pages 174–184, 2013.

[Ramezani2015] Mohsen Ramezani, Jack Haddad and Nikolas Geroliminis. *Dynamics of heterogeneity in urban networks: aggregated traffic modeling and hierarchical control.* Transportation Research Part B: Methodological, vol. 74, pages 1–19, 2015.

[Rodriguez-Vega2020] Martin Rodriguez-Vega, Carlos Canudas-de-Wit and Hassen Fourati. *Average density detectability in traffic networks using virtual road divisions.* In IFAC 2020-IFAC World Congress 2020, 2020.

[Rodriguez-Vega2021] Martin Rodriguez-Vega, Carlos Canudas-de-Wit and Hassen Fourati. *Average density estimation for urban traffic networks: application to the Grenoble network.* 2021.

[Roger1991] Horn Roger and R Johnson Charles. Topics in matrix analysis. Cambridge University Press, 1991.

[Romagnani2020] Paola Romagnani, Guido Gnone, Francesco Guzzi, Simone Negrini, Andrea Guastalla, Francesco Annunziato, Sergio Romagnani and Raffaele De Palma. *The COVID-19 infection: Lessons from the Italian experience.* Journal of Public Health Policy, pages 1–7, 2020.

[Rotella2011] Frédéric Rotella and Irène Zambettakis. *Minimal single linear functional observers for linear systems.* Automatica, vol. 47, no. 1, pages 164–169, 2011.

[Rotella2015] Frédéric Rotella and Irène Zambettakis. *A note on functional observability.* IEEE Transactions on Automatic Control, vol. 61, no. 10, pages 3197–3202, 2015.

[Rotella2016] Frédéric Rotella and Irène Zambettakis. *A direct design procedure for linear state functional observers.* Automatica, vol. 70, pages 211–216, 2016.

[Rothman2008] Kenneth J Rothman, Sander Greenland and Timothy L Lash. Modern epidemiology. Lippincott Williams & Wilkins, 3 edition, 2008.

[Sadamoto2017] Tomonori Sadamoto, Takayuki Ishizaki and Jun-Ichi Imura. *Average State Observers for Large-Scale Network Systems.* IEEE Transactions on Control of Network Systems, vol. 4, no. 4, pages 761–769, 2017.

[Sakhraoui2018] Imane Sakhraoui, Baptiste Trajin and Frédéric Rotella. *Discrete linear functional observer for the thermal estimation in power modules.* In 2018 IEEE 18th International Power Electronics and Motion Control Conference (PEMC), pages 812–817. IEEE, 2018.

[Salathé2020] Marcel Salathé, Christian L Althaus, Richard Neher, Silvia Stringhini, Emma Hodcroft, Jacques Fellay, Marcel Zwahlen, Gabriela Senti, Manuel Battegay, Annelies Wilder-Smith*et al. COVID-19 epidemic in Switzerland: On the importance of testing, contact tracing and isolation.* Swiss medical weekly, vol. 150:w20225, 2020.

[Sandberg2004] Henrik Sandberg and Anders Rantzer. *Balanced truncation of linear time-varying systems.* IEEE Transactions on Automatic Control, vol. 49, no. 2, pages 217–229, 2004.

[Sandberg2009] Henrik Sandberg and Richard M Murray. *Model reduction of interconnected linear systems.* Optimal Control Applications and Methods, vol. 30, no. 3, pages 225–245, 2009.

[Sattenspiel1995] Lisa Sattenspiel and Klaus Dietz. *A structured epidemic model incorporating geographic mobility among regions.* Mathematical biosciences, vol. 128, no. 1-2, pages 71–91, 1995.

[Scardovi2009] Luca Scardovi and Rodolphe Sepulchre. *Synchronization in networks of identical linear systems.* Automatica, vol. 45, no. 11, pages 2557–2562, 2009.

[Sinani2019] Klajdi Sinani and Serkan Gugercin. $\mathcal{H}_2(t_f)$ *optimality conditions for a finite-time horizon.* Automatica, vol. 110, page 108604, 2019.

[Skadron2002] K. Skadron, T. Abdelzaher and M. R. Stan. *Control-theoretic techniques and thermal-RC modeling for accurate and localized*

dynamic thermal management. In Proceedings Eighth International Symposium on High Performance Computer Architecture, pages 17–28, Feb 2002.

[Song2020] Sirui Song, Zefang Zong, Yong Li, Xue Liu and Yang Yu. *Reinforced Epidemic Control: Saving Both Lives and Economy.* arXiv preprint arXiv:2008.01257, 2020.

[Trinh2011] Hieu Minh Trinh and Tyrone Lucius Fernando. Functional observers for dynamical systems. Springer-Verlag Berlin Heidelberg, 2011.

[Van den Driessche2008] P Van den Driessche and James Watmough. *Further notes on the basic reproduction number.* In Mathematical epidemiology, pages 159–178. Springer, 2008.

[van der Schaft2013] Arjan van der Schaft, Shodhan Rao and Bayu Jayawardhana. *On the mathematical structure of balanced chemical reaction networks governed by mass action kinetics.* SIAM Journal on Applied Mathematics, vol. 73, no. 2, pages 953–973, 2013.

[Walker2020] Patrick Walker, Charles Whittaker, Oliver Watson, Marc Baguelin, K Ainslie, Sangeeta Bhatia, Samir Bhatt, A Boonyasiri, O Boyd, L Cattarino*et al. Report 12: The global impact of COVID-19 and strategies for mitigation and suppression.* Imperial College London (26-03-2020), 2020.

[Walter1999] Gilbert G Walter and Martha Contreras. Compartmental modeling with networks. Springer-Birkhäuser Boston, 1999.

[Wang006a] S. Wang and X. Xu. *Simplified building model for transient thermal performance estimation using GA-based parameter identification.* International Journal of Thermal Sciences, vol. 45, no. 4, pages 419–432, 2006a.

[Wang006b] Shengwei Wang and Xinhua Xu. *Parameter estimation of internal thermal mass of building dynamic models using genetic algorithm.* Energy conversion and management, vol. 47, no. 13-14, pages 1927–1941, 2006b.

[Wang1973] Shih-Ho Wang and EJ Davison. *On the stabilization of decentralized control systems.* IEEE Transactions on Automatic Control, vol. 18, no. 5, pages 473–478, 1973.

[Wei1969] James Wei and James C W Kou. *A lumping analysis in monomolecular reaction systems.* Industrial & Engineering Chemistry Fundamentals, vol. 8, pages 114–123, 1969.

[WHO2020] *Advice on the use of point-of-care immunodiagnostic tests for COVID-19: Scientific brief, 8 April 2020.* Technical documents, World Health Organization, 2020.

[Winter2020] Amy K Winter and Sonia T Hegde. *The important role of serology for COVID-19 control.* The Lancet Infectious Diseases, vol. 20, no. 7, pages 758–759, 2020.

[Wonham1967] W Wonham. *On pole assignment in multi-input controllable linear systems.* IEEE transactions on automatic control, vol. 12, no. 6, pages 660–665, 1967.

[Wu2004] Jianjun Wu, Ziyou Gao, Huijun Sun and Haijun Huang. *Urban transit system as a scale-free network.* Modern Physics Letters B, vol. 18, no. 19n20, pages 1043–1049, 2004.

[Yip1981] Elizabeth Yip and Richard Sincovec. *Solvability, controllability, and observability of continuous descriptor systems.* IEEE Transactions on Automatic Control, vol. 26, no. 3, pages 702–707, 1981.

[Zhan2009] Zhi-Hui Zhan, Jun Zhang, Yun Li and Henry Shu-Hung Chung. *Adaptive particle swarm optimization.* IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 39, no. 6, pages 1362–1381, 2009.

[Zhou2020] Fei Zhou, Ting Yu, Ronghui Du, Guohui Fan, Ying Liu, Zhibo Liu, Jie Xiang, Yeming Wang, Bin Song, Xiaoying Gu*et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study.* The Lancet, 2020.

List of Figures

List of Tables