2019

Analysis And Control Of Networked Systems Using Structural And Measure-Theoretic Approaches

Ximing Chen
University of Pennsylvania, simon.cxm.1120@gmail.com

Recommended Citation

Chen, Ximing, "Analysis And Control Of Networked Systems Using Structural And Measure-Theoretic Approaches" (2019). *Publicly Accessible Penn Dissertations*. 3428.
https://repository.upenn.edu/edissertations/3428

Analysis And Control Of Networked Systems Using Structural And Measure-Theoretic Approaches

Abstract

Network control theory provides a plethora of tools to analyze the behavior of dynamical processes taking place in complex networked systems. The pattern of interconnections among components affects the global behavior of the overall system. However, the analysis of the global behavior of large scale complex networked systems offers several major challenges. First of all, analyzing or characterizing the features of large-scale networked systems generally requires full knowledge of the parameters describing the system's dynamics. However, in many applications, an exact quantitative description of the parameters of the system may not be available due to measurement errors and/or modeling uncertainties. Secondly, retrieving the whole structure of many real networks is very challenging due to both computation and security constraints. Therefore, an exact analysis of the global behavior of many real-world networks is practically unfeasible. Finally, the dynamics describing the interactions between components are often stochastic, which leads to difficulty in analyzing individual behaviors in the network.

In this thesis, we provide solutions to tackle all the aforementioned challenges. In the first part of the thesis, we adopt graph-theoretic approaches to address the problem caused by inexact modeling and imprecise measurements. More specifically, we leverage the connection between algebra and graph theory to analyze various properties in linear structural systems. Using these results, we then design efficient graph-theoretic algorithms to tackle topology design problems in structural systems. In the second part of the thesis, we utilize measure-theoretic techniques to characterize global properties of a network using local structural information in the form of closed walks or subgraph counts. These methods are based on recent results in real algebraic geometry that relates semidefinite programming to the multidimensional moment problem. We leverage this connection to analyze stochastic networked spreading processes and characterize safety in nonlinear dynamical systems.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Electrical & Systems Engineering

First Advisor
Victor M. Preciado

Keywords
Algorithms, Epidemic models, Graph theory, Multidimensional moment problem, Networked systems, Structural systems

Subject Categories
Applied Mathematics | Engineering

ANALYSIS AND CONTROL OF NETWORKED SYSTEMS USING STRUCTURAL AND MEASURE-THEORETIC APPROACHES

Ximing Chen

A DISSERTATION

in

Electrical and Systems Engineering

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2019

Supervisor of Dissertation

Victor M. Preciado, Associate Professor

Graduate Group Chairperson

Victor M. Preciado, Associate Professor

Dissertation Committee

George J. Pappas, Joseph Moore Professor and Chair

Manfred Morari, Distinguished Faculty Fellow

Masaki Ogura, Assistant Professor of Division of Information Science, Nara Institute of Science and Technology

ANALYSIS AND CONTROL OF NETWORKED SYSTEMS USING STRUCTURAL
AND MEASURE-THEORETIC APPROACHES

Acknowledgement

The last six years at Penn has been a joyful journey for me. I would like to express my great gratitude towards my advisor Dr. Victor M. Preciado, who has patiently trained me from an immature researcher to a qualified Ph.D. candidate. His patient advisory, and superb vision towards research has influenced me in a great deal. I am especially thank him for all his endeavors on helping me finesse every single detail in my paper in the most rigorous way possible.

It is of great honor to invite Dr. George J. Pappas, Dr. Manfred Morari, and Dr. Masaki Ogura to serve in my dissetation committee. They have raised keen comments on my research as well as the scope of this thesis. I would like to thank Dr. Pappas for his sharp insights towards my research area, Dr. Morari for his mentorship in my late Ph.D. career, and Dr. Ogura for his active advice and continuing guidance. Their experience in the field of control theory has certainly enlightened the path towards research. It is very fortunate to have them in my committee and no words can express my gratitude except an ancient Chinese proverb: "A single conversation across the table with a wise man is better than ten years mere study of books".

This thesis would probably not be completed without the help of talented researchers in this community. I would like to thank Jingqi Li and Dr. Sergio Pequito for collaborating on the first part of the thesis, and Dr. Khem. R. Ghusinga, Dr. Abhyudai Singh, and Dr. Masaki Ogura for contributing valuable ideas in the second part of this thesis. I shall say that it has been an wonderful opportunity and extraordinary experience. Dr. Ufuk Topcu, Dr. Shuo Han and Dr. Fei Miao mentored me in my early Ph.D. career and helped me get through numerous difficulties, for which I am extraordinarily thankful.

Moreover, I would like to express my sincere gratitude to my previous supervisor and mentors in HKUST, Dr. Li Qiu, Dr. Ling Shi, and Dr. Shenghui Song who, not only helped me build up solid analytical and critical thinking abilities, but also encouraged me to pursue an academia career. I would also like to thank Dr. Tara Javidi and Dr. Alexandre Megretski for providing extraordinary warm-hearted advice on conducting research during my exchange and summer internship program. To Dr. Zhihao Jiang, Dr. Chungwei Lin, Dr. Eunsuk Kang, and Dr. Shiraishi Shinichi in Toyota InfoTechnology Center for offering me an opportunity to experience cutting-edge technologies in the autonomous-driving industry.

The content presented in this manuscript is also inspired and benefited from discussions with friends at the University of Pennsylvania including, but not limited to, Dr. Weiyu Huang, Dr. Min Wen, Jihua Huang, Dr. Meng Xu, Dr. Xiaowei Zhou, Dr. Sikang Liu, Dr. Jinwook Heo, Dr. Bhoram Lee, Dr. Daewon Lee, Sangdon Park, Yitong Wang, Tao Zhang, Clark Zhang, as well as dearest group members and companions Dr. Cassiano Becker, Dr. Mahyar Fazlyab, Alp Aydinoglu, Jorge F. Barreras, Shaoru Chen, Mikhail Hayhoe, and Jacob Seidman. May the friendship be with us.

I would like to thank my parents and my grandparents for putting up with my untiring mischievous behavior, raising me up, and always wishing me *a bright future*. Last but not the least, I would like to express my deepest gratitude to my significant other half Siyi Luo for her care, support and love. Time and distance did not break us apart and nothing will. The adventure we are about to embark upon together will be joyful and wonderful.

Ximing Chen
Philadelphia, May 2019.

ABSTRACT

ANALYSIS AND CONTROL OF NETWORKED SYSTEMS USING STRUCTURAL AND MEASURE-THEORETIC APPROACHES

Ximing Chen

Victor M. Preciado

Network control theory provides a plethora of tools to analyze the behavior of dynamical processes taking place in complex networked systems. The pattern of interconnections among components affects the global behavior of the overall system. However, the analysis of the global behavior of large scale complex networked systems offers several major challenges. First of all, analyzing or characterizing the features of large-scale networked systems generally requires full knowledge of the parameters describing the system's dynamics. However, in many applications, an exact quantitative description of the parameters of the system may not be available due to measurement errors and/or modeling uncertainties. Secondly, retrieving the whole structure of many real networks is very challenging due to both computation and security constraints. Therefore, an exact analysis of the global behavior of many real-world networks is practically unfeasible. Finally, the dynamics describing the interactions between components are often stochastic, which leads to difficulty in analyzing individual behaviors in the network.

In this thesis, we provide solutions to tackle all the aforementioned challenges. In the first part of the thesis, we adopt graph-theoretic approaches to address the problem caused by inexact modeling and imprecise measurements. More specifically, we leverage the connection between algebra and graph theory to analyze various properties in linear structural systems. Using these results, we then design efficient graph-theoretic algorithms to tackle topology design problems in structural systems. In the second part of the thesis, we utilize measure-theoretic techniques to characterize global properties of a network using local structural information in the form of closed walks or subgraph counts. These methods are based on recent results in real algebraic geometry that relates semidefinite programming to the multidimensional moment problem. We leverage this connection to analyze stochastic networked spreading processes and characterize safety in nonlinear dynamical systems.

Contents

Notations

$\mathbf{x}$	A vector
x_i	The i-th element in a vector $\mathbf{x}$
$\mathbf{x}^{\boldsymbol{\alpha}}$	The monomial $\Pi_{i=1}^{n} x_i^{\alpha_i}$
$\mathbf{x}^\top$	The transpose of a vector $\mathbf{x}$
$\lvert\mathbf{x}\rvert$	The 1-norm of a vector $\mathbf{x} = \sum_{i=1}^{n} \lvert x_i \rvert$
$\mathbf{1}_n$	The n-dimensional vector of all ones
$\mathcal{S}, \lvert\mathcal{S}\rvert$	A set and the cardinality of the set $\mathcal{S}$
$\mathbf{1}_{\mathcal{S}}$	The indicator function of the set $\mathcal{S}$
$\mathbb{R}$	The set of real numbers
$\mathbb{R}_{+}^{n}, \mathbb{R}_{++}^{n}$	The set of n-dimensional vectors with nonnegative, positive entries
$\mathbb{C}$	The set of complex numbers
$\mathbb{N}$	The set of non-negative integers
$\mathbb{N}_{r}^{n}$	The set of integer vectors $\{\mathbf{x} \in \mathbb{N}^n : \lvert\mathbf{x}\rvert \leq r\}$.
$[n]$	The set of integers $\{1, \ldots, n\}$
$[M]_{ij}$	The element in the i-th row and j-th column of a matrix M
$\bar{M}$	The structural pattern of a matrix M
$M \succeq 0$	The symmetric matrix M is positive semidefinite.
$M^\top$	The transpose of a matrix M
$G(\mathcal{V}, \mathcal{E})$	A directed graph with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$
$\mathcal{G}$	A mixed graph
$\mathcal{N}^{-}(\mathcal{S}), \mathcal{N}^{+}(\mathcal{S})$	The set of in-neighbors and out-neighbors of a set of nodes $\mathcal{S}$ in G

Chapter 1

Introduction

Natural and artificial systems often consist of a large number of components interconnected via a complex pattern of connections [1–3]. Examples of such complex systems include biological [4–6], brain [7–10], social [11–14] and communication [15–18] networks, to mention a few. In particular, the pattern of interconnections among these components affects the global behavior of the overall system. In this direction, network control theory powerful tools to characterize and analyze the structure and function of complex networked systems, examples include epidemic outbreaks in human contact networks [19], information spreading in social networks [20], or synchronization in power systems [21], among with others [22–24].

The underlying goal of using networks to model various natural and engineered systems is to reveal how the interconnection patterns between components affect the global behavior of the overall system. For example, how fast a meme is spreading in a social network is related to how individuals are connected to each other. However, analyzing the global behavior of large-scale complex networked systems often faces a few challenges. Firstly, analyzing or characterizing the dynamics of large-scale networked systems often requires full knowledge of the parameters describing the system's structure. In many applications involving large-scale networks, an exact quantitative description of the parameters of the system may not be available due to measurement errors and/or modeling uncertainties [25, 26]. Secondly, characterizing the global behavior of a system cannot be done without having access to the entire structure of the network. However, the sheer size of real-world networks makes the problem computationally challenging. In this direction, it is typically impossible to retrieve the whole structure of many real networks due to both computation and/or security constraints. For example, the spectral radius of a directed graph is closely related to many networked system properties such as the graph independence number and the speed of networked spreading processes. Nonetheless, the spectral radius can only be computed when the structure of the network is completely known. Finally, in certain networked systems, the dynamics describing the evolution of the network states is often stochastic, which makes the analysis challenging.

In this thesis, I provide solutions to tackle *all* the aforementioned challenges. More specifically, in the first part of my thesis, I use tools from structural system theory and graph theory to address the problem caused by inexact modeling and imprecise measurements. In the second part of my thesis, I use measure-theoretic techniques to analyze the global behavior of a network using local structural information. This method is based on recent development in functional analysis providing a connection between semidefinite programming and the multivariate moment problem. Furthermore, I leverage this connection to analyze stochastic spreading processes on networks, among other applications such as safety synthesis of nonlinear dynamical systems.

1.1 Structural and Graph-theoretic Approach for System Analysis and Design

Complex networks have been shown to be a powerful tool for modeling dynamical systems [26–28]. In particular, when analyzing and designing networked dynamical systems, it is crucial to verify their controllability, i.e., the existence of an input sequence allowing us to drive the states of the system towards an arbitrary state within finite time. Nonetheless, verifying such a property requires full knowledge of the parameters describing the system's dynamics [29]. However, in many applications involving large-scale networks, those parameters are difficult, or even impossible, to obtain [26]. Alternatively, it is practically more viable to identify the presence or absence of dynamical interconnections between each pair of nodes in the network, without characterizing the strength of these interactions. Subsequently, it is of interest to analyze system properties, such as controllability, using exclusively information about the system's structure and tools from graph theory [30]. In other words, even though an exact quantification on the edge-weights in the network may not be available, it is still possible to analyze network control properties resorting to tools developed in the context of structural systems theory [31–34].

Seminal work on graph-theoretic analysis of controllability can be found in [31], in which the notion of *structural controllability* was introduced. Following this seminal work, the authors in [32, 33, 35, 36] provided necessary and sufficient conditions for structural controllability of multi-input linear time-invariant (LTI) systems using various graph-theoretic notions. Nonetheless, existing results on structural controllability assumed implicitly that the parameters are either fixed zeros or independent free variables (see Figure 1-1 for an illustration). Nonetheless, such an assumption is often violated in practical scenarios, for instance, when the system is characterized by undirected networks [37], or when different interconnections in the system are strongly correlated [38]. Consequently, it is of interest to provide necessary and sufficient conditions for structural systems with special weight constraints, which is the main focus of Chapter 2. Similar problems are considered in [39] and [40]. However, the result in [39] is not applicable to systems modeled by undirected graph, whereas the matrix net approach in [40] may suffer from computational complexity in large-scale systems. Moreover, the authors in [41] and [42] also proposed, separately and independently, graph-theoretic necessary and sufficient conditions for structural controllability of dynamical systems modeled by a symmetric graph.

Another objective of Chapter 2 is to provide necessary and sufficient conditions for structural output controllability of LTI systems with symmetric state matrices. This is motivated by the following concern: While controllability is concerned about our ability

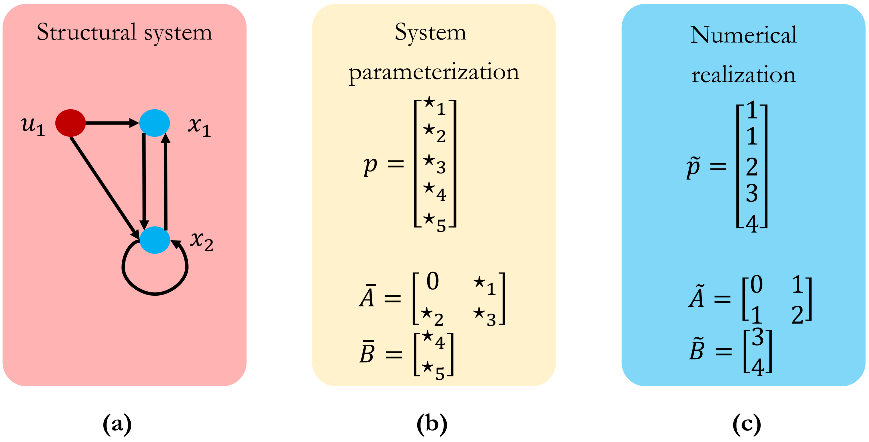

Structural system
u_1
x_1
x_2
System parameterization
$p = \begin{bmatrix} \star_1 \\ \star_2 \\ \star_3 \\ \star_4 \\ \star_5 \end{bmatrix}$
$\bar{A} = \begin{bmatrix} 0 & \star_1 \\ \star_2 & \star_3 \end{bmatrix}$
$\bar{B} = \begin{bmatrix} \star_4 \\ \star_5 \end{bmatrix}$
Numerical realization
$\tilde{p} = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$
$\tilde{A} = \begin{bmatrix} 0 & 1 \\ 1 & 2 \end{bmatrix}$
$\tilde{B} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$
(a)
(b)
(c)

Figure 1-1: (a) An examplary graph representation of a linear structural system with two states and one input. The input is marked by red circle whereas the states are marked by blue circles. (b) The algebraic representation of the structural system, where $p \in \mathbb{R}^5$ is a vector consisting of 5 independent parameters, denoted using $\star_i, i = 1, \ldots, 5$. Each of the parameter in p represents an unknown weight value of edges in the graph in (a). (c) A numerical realization of the structural system in which the parameters are set to be $\tilde{p}$.

to steer all the states in the system arbitrarily, in certain scenarios involving large-scale systems, we are only concerned about our ability to steer a specific subset of states. In this context, we consider the notion of *(structural) target controllability* [43, 44]. More generally, it is shown in [45] that target controllability is a particular case of output controllability, i.e., our ability steer the system's outputs arbitrarily. Although there are recent work providing necessary and sufficient conditions for *strong* structural target controllability [46, 47], to the best of our knowledge, providing necessary and sufficient conditions for structural target controllability and structural output controllability remains an unsolved problem [48]. As a result, we provide the first result for characterizing structural output controllability in structural linear systems.

Loosely speaking, a network is structurally controllable if it is controllable for almost all realizations of edge weights (see Section 3.1 in Chapter 2 for a formal description of this concept). When a networked system is not structurally controllable, then it is impossible to design a controllable system by tuning the weights of the network. In this case, one has to resort to structural changes in the system. This can be done by either (*i*) adding actuation capabilities to the networked system, or (*ii*) modifying the topology of the dynamical network by, for example, adding new edges to the network topology. The former case is explored in [28, 48–54]. Briefly, in [28, 48, 54], proposed graph-theoretical algorithms to find the minimum number of driving nodes to ensure structural controllability in complex networks. In [49, 50], the authors complement this work to obtain the minimum number of driven nodes in polynomial-time. Subsequently, the minimum number of driven nodes required while accounting for actuation costs was addressed in [51, 52]. Alternatively, if one seeks the minimum collection of inputs from an a priori defined collection of actuation capabilities, then the problem is NP-hard [53]. Notwithstanding, there are several cases when adding actuation capabilities to the network is either too expensive or not feasible. Therefore, whenever possible or cost-efficient, one can opt to modify the topology of the dynamical network. This is the main focus of the Chapter 3 of this thesis, where we propose a polynomial-time algorithm to determine the minimum number of extra connections that must be added to a given structural system in order to ensure structural controllability.

We find in the literature several similar and earlier work in this direction. In [55], Wang et al. proposed an approach to perturb the structure of an *undirected* network to ensure structurally controllability when only one driving node was considered. In [56], Ding et al. studied a similar problem for directed networks. However, they assumed that all the nodes are already reachable from the driving nodes. Altough they solved the problem using a constrained integer program which is, in general, NP-hard [57], they did not discuss the complexity of their algorithm. In contrast with previous works, we address the case of arbitrary directed network topologies with any number of driving nodes and show that the problem can be solved in polynomial time without any assumption on reachability. Parallel work on the same topic was also explored by [58], in which the authors provided complete characterizations of the computation complexity of adding a number of directed edges that incurs into minimum total cost to render a structurally controllable system.

In Chapter 4, we further extend our design strategies introduced in Chapter 3 to consider topology design problems in symmetrically structured systems. More specifically, we consider the minimum-cost edge selection problem, in which we aim to add *undirected* edges to a given graph to render a symmetrically structurally target controllable system. Moreover, we assume that each edge can be added at a given cost and we aim to find a configuration of edges incurring into minimum total cost. We provide full characterizations on the computation complexity of this problem and provide polynomial-time algorithms that are able to obtain optimal solutions to a few polynomially-solvable instances.

While controllability of networked dynamical systems has attracted a great amount of research interest [59–61], it is of equal interest to consider a less restrictive system property called stabilizability, especially in typical engineering system design problems [59–61]. Instead of requiring full-steering ability of all states of the system, stabilizability only requires that the system states can be steered to the origin asymptotically by injecting proper controls. However, similar to controllability, assessing whether a system is stabilizable requires the exact parameters of the system.

In this direction, assessing the stabilizability from the structural information on the system dynamics model has been an active topic of research [49, 62, 63]. In Chapter 5, we provide graph-theoretic characterization on structural stabilizability of symmetrically structured systems. We further utilize the result to consider design problems from a network security perspective. Unlike existing work on control of networks under malicious attackes [58, 64–72], we investigate a fully novel problem of optimal attack/recovery against stabilizability by manipulating network topological structure through disabling/adding actuators.

1.2 Measure-theoretic Approach for System Analysis

In the first part of this thesis, we use structural systems theory and graph theory to analyze networked dynamical systems with uncertain weights. In the second part of the thesis, we will mainly be concerned about the following two challenges: (i) analyzing global system properties when the entire structure of the network is impossible to obtain, and (*ii*) when the dynamics describing the transitions between the network states are stochastic. Before addressing these two challenges, we first provide background and motivation about the importance of these two problems.

1.2.1 Analyzing Global Properties of Networks Using Local Structural Information

A common approach to model complex networks is via synthetic random models, such as the Erdős-Rényi random graph [73], the Watts-Strogatz small-world model [74], or the Barabási-Albert model [75], among many others [15, 76–78]. Synthetic models have been used to analyze, for example, the behavior of many networked dynamical processes, such as synchronization of coupled oscillators [74, 79, 80], network diffusion [81, 82] or epidemic spreading [19, 83, 84] (see, for example [85, 86], for a thorough exposition). A fruitful path to analyze the dynamics of networked processes exploits the connection between network eigenvalues and dynamics. For example, the eigenvalues of the adjacency matrix can be used to characterize the speed of spreading of epidemic processes in networks [19, 83, 84, 87, 88].

Even though network eigenvalues are of utmost importance, its computation in large-scale networks is a very challenging problem [89]. On the one hand, the sheer size of real-world networks makes this problem computationally challenging. On the other hand, it is typically impossible to retrieve the whole structure of many real networks due to privacy and/or security constraints. In contrast, it is usually feasible to extract local samples of the network structure in the form of ego-networks [90] or subgraph counts [6, 91–93] using graph crawlers. It is, therefore, of interest to analyze the role of local structural samples on the global eigenvalue spectrum of a complex network.

We find in the literature many works aiming to upper and lower bound the spectral radius of a graph from local structural information [94–105]. In [94] and [95], the authors derived upper bound the spectral radius of a matrix from its symmetric and skew-symmetric components. Merikoski et al. [98] provided bounds on the sum of selected eigenvalues using the trace and the determinant. Instead of bounding the eigenvalues of arbitrary square matrices, the works in [96, 97, 101] provide lower bounds on the spectral radius of general non-negative matrices. Most of these bounds are based on the traces of the matrix and/or its second power. In [99], the authors use the traces of even-order powers of a matrix to provide upper bounds on the spectral radius, assuming that the eigenvalues are all real. In [104, 105], the authors bound the spectral radius of an *undirected* graph using subgraph counts. Similar results were obtained for the spectral gap of the Laplacian matrix in [106, 107].

In Chapter 6, we develop a measure-theoretic framework to obtain upper and lower bound on the spectral radius (a global system property) of large *directed* graphs using counts of small subgraphs (local structural information). By exploiting recent results in the multi-dimensional moment problem [108], we propose a hierarchy of small semidefinite programs [109] providing converging sequences of upper and lower bounds on the spectral radius. We numerically show that our framework provides accurate upper and lower bounds in real-world directed networks, as well as random synthetic digraphs.

While eigenvalues of the adjacency matrix is of importance, in certain applications, it is of interest to study the eigenvalue spectrum of the Laplacian matrix. Particularly relevant is the second smallest eigenvalue of the Laplacian matrix, also called the algebraic connectivity [110], which is crucial in the analysis of consensus algorithms [111, 112] and network synchronization [79]—see [113] and the references therein for a detailed expositions of the applications of algebraic connectivity.

Due to the practical importance of the algebraic connectivity in systems and control, several papers have been recently published on estimation of Laplacian eigenvalues [113–121]. In [115], the authors presented a scheme in which agents in a network run a continuous-time dynamics able to induce oscillations whose frequencies match the Laplacian eigenvalues. In [116], a decentralized power iteration approach is introduced to efficiently estimate the algebraic connectivity and the corresponding eigenvector using a continuous-time implementation. Analogous methods are developed to estimate algebraic connectivity in directed graphs by authors in [117]. In [121], the authors provide a continuous-time nonlinear dynamics on manifolds to distributively compute the top largest (or smallest) eigenvalues of both the Laplacian and adjacency matrices. A discrete-time dynamics inspired by the power iteration is proposed in [118]. Additionally, the authors provided upper and lower bounds on the quality of estimation. In [119], the authors used Khatri-Rao product of matrices to estimate Laplacian eigenvalues and eigenvectors from a finite sequence of observations of a consensus-like dynamics.

Although there are a variety of work on algebraic connectivity estimation, few work are able to achieve this goal using only local topological information of the network. In this direction, the authors in [122] provided an upper bound on algebraic connectivity from local structural samples of the network. In Section 8.1 of Chapter 8, we extend the idea in [122] to provide a sequence of lower bounds on the algebraic connectivity of an undirected graph. We show that, by leveraging the measure-theoretic framework presented in Chapter 6, we can obtain nontrivial lower bounds using, solely, local structural information of the graph.

1.2.2 Analysis and Control of Spreading Processes

Modeling and analysis of spreading processes taking place in complex networks have found applications in a wide range of scenarios, such as modeling the propagation of malware in computer networks [123], failures in technological networks [124], memes in social networks [125], and diseases in human populations [19, 84, 126]. We find in the literature a wide variety of models to characterize the dynamics of spreading processes over networks. In the epidemiological literature, these models consider the spread of a disease in human contact networks in which individuals and their relationships are modeled via complex networks. Some of the most popular models in the literature are the *Susceptible-Infected-Susceptible* (SIS) [127], the *Susceptible-Infected-Recovered* (SIR) [128], and their variants [129, 130].

During the last decade, several mathematical techniques have been developed to determine whether a disease spreading over a network will be eradicated quickly or, in contrast, will spread widely over time, causing a large epidemic outbreak. These techniques can then be used to design efficient strategies to contain, or even eradicate, the spread of the disease by distributing medical resources throughout the network [131–134]. One of the most important characteristics in the global behavior of these models is the presence of phase transitions, or epidemic thresholds. These phase transitions can be described as a dynamical bifurcations in the dynamics, where the system transition from a single stable equilibrium at the origin (i.e., the disease-free state) towards the existence of (potentially many) nontrivial equilibria. The authors in [87] presented an approximate analysis to show that the networked SIS model presents a phase transition that can be characterized in terms of the largest eigenvalue of the adjacency matrix representing the network structure. A rigorous analysis of this phase transition was presented in [127], where the authors use Markov processes to model the exact stochastic dynamics of the networked SIS spreading process. Following this approach, the authors in [130] characterized the global dynamics of a more general spreading model which includes the SIS as a particular case.

A common idea behind the aforementioned results is to construct a Markov transition model and analyze the transition probabilities among network states. However, the number of possible network states grows exponentially with the number of nodes. Consequently, the analysis of the resulting Markov process is both computationally and analytically challenging to study. An alternative approach to overcome this challenge is to analyze the probability of each node being in a particular state at a given time. For example, in the SIS spreading model, it is mathematically convenient to analyze the time evolution of the probabilities of infection of each node in the network. However, as illustrated in [127], the ODEs describing the evolution of these infection probabilities depend on pairwise correlations (second-order moments) between the states of connected nodes in the network. As shown in [135], the governing dynamics of these second-order moments can also be described as ODEs involving third-order moments. In general, the ODEs describing the dynamics of k-th order moments depend on $(k+1)$-th order moment. Therefore, a complete characterization of the dynamics requires, in general, an exponential number of ODEs. To address this issue, it is common to resort to moment-closure techniques—a method to obtain a closed system of ODEs by approximating higher-order moments using lower-order ones [136]. Hence, it would be possible to use this technique to obtain a polynomial number of ODEs approximating the dynamics of moments of order k as a function of moments of order up to k. For instance, the popular mean-field approximation (MFA) is a moment-closure techniques in which pairwise correlations are approximated by the products of two first-order expectations [127], resulting in a linear number of ODEs. In [137], the authors proposed to close second-order moments using Fréchet inequalities; whereas the authors in [138] proposed to close third-order moments by products of first- and second-order moments.

Existing moment-closure techniques suffer from the following pitfalls. Firstly, there is no theoretical guarantee on the quality of the approximation obtained. Secondly, in the particular case of the SIS dynamics, these techniques often fail to lower bound the evolution of the probabilities of infection of each node. These pitfalls are partly addressed by [137, 139]; in particular, in [139] the authors showed that the moment-closure problem can be directly related to the *multidimensional moment problem* in functional analysis [108]. In Chapter 7, we propose a mathematical and computational framework to obtain a polynomial number of ODEs describing the dynamics of all k-th order moments of the SIS stochastic model for an arbitrary integer k and an arbitrary contact network. Our framework utilizes recent results in real algebraic geometry relating the multidimensional moment problem with semidefinite programming [108]. As part of this framework, we provide upper and lower bounds on the evolution of an arbitrary k-th moment of the SIS stochastic model. Moreover, we provide a simplified expression for $k = 1$ to approximate the dynamics of the means of each node state using a linear number of piecewise-affine differential equations. Finally, we extend our framework to other compartmental spreading processes over networks, such as the SI and SIR models.

In the first part of Chapter 7, we study the dynamic behavior of single-disease processes in single-layer networks. However, these models do not fully characterize how information spread throughout networks in the real world. For example, it is possible for a human to obtain news via online and/or offline social networks. Aiming to provide a more realistic model, Sahneh et al. proposed in [140] a modeling framework to analyze spreading processes in multilayer networks. A further extension was proposed by Funk and Jansen [141] by analyzing the case of two competitive viruses in a two-layer network. In a similar direction, Wei et al. in [142] derived sufficient conditions for exponential die-out of two competitive SIS viruses on an arbitrary two-layer network. A rigorous nonlinear analysis of this model can be found in [143]. In Section 7.5 of Chapter 7, we build on previous approaches to analyze and control the dynamics of (i) an arbitrary number of diseases spreading through (ii) a multilayer contact network of non-identical agents.

A central problem in public health is the development of vaccination strategies to tame disease spreading. Several works have been recently propose in this direction. In [144], several heuristics were proposed to immunize distribute vaccines throughout the nodes of a network. In the control systems literature, Wan et al. proposed in [134] a method to control the spread of a virus using eigenvalue sensitivity analysis. Our work is closely related to [145], where the epidemic control problem was studied from an optimization point of view, and [132, 146], where the authors developed different strategies for optimal resource allocation in the single-disease, single-layer case using convex optimization. In Section 7.5 of Chapter 7, we extend this framework to find the optimal budget allocation to fabricate and distribute different types of vaccines to simultaneously control several diseases in a (possibly directed) multi-layered network of non-identical agents.

1.3 Contributions

The work presented in this thesis have been published or submitted for publication in conferences and journals such as IEEE Conference on Decision and Control, IEEE Transactions on Control of Networked Systems, IEEE Transactions on Automatic Control, and SIAM Journal on Matrix Analysis and Applications (see [147–151]). In addition to these papers, the author have also published works on engineering applications related to smart cities and connected vehicles (see [152, 153]).

Part I

Structural System Analysis and Design via Graph Theory

Chapter 2

Graph-theoretic Characterization of Symmetric Linear Structural Systems

We begin this chapter by introducing notion related to structural system theory and graph theory that are necessary for the development of the results presented in the first part of the thesis (i.e., Chapter 2–5).

2.1 Preliminaries on Graph theory and Algebra

2.1.1 Notations in Algebra

We denote the cardinality of a set $\mathcal{S}$ by $|\mathcal{S}|$. We adopt the notation $[n]$ to represent the set of integers $\{1, \ldots, n\}$. Let $\mathbf{0}_{n\times m} \in \mathbb{R}^{n\times m}$ be the matrix with all entries equals to zero. Whenever clear from the context, $\mathbf{0}_{n\times m}$ is abbreviated as $\mathbf{0}$.

Given $M_1 \in \mathbb{R}^{n\times m_1}$ and $M_2 \in \mathbb{R}^{n\times m_2}$, we let $[M_1, M_2] \in \mathbb{R}^{n\times(m_1+m_2)}$ be the concatenation of M_1 and M_2. The ij-th entry of $M \in \mathbb{R}^{n\times n}$ is denoted by $[M]_{ij}$. Moreover, we let $[M]_{i_1,\ldots,i_k}^{j_1,\ldots,j_k}$ be the $k \times k$ submatrix of M formed by collecting $i_1, \ldots, i_k$-th rows and

$j_1, \dots, j_k$-th columns of M. The determinant of a matrix $M \in \mathbb{R}^{n \times n}$ is defined by the expansion:

$$\det M = \sum_{\sigma \in \mathcal{S}_n} \left(\operatorname{sgn}(\sigma) \prod_{i=1}^{n} [M]_{i\sigma(i)} \right), \tag{2.1}$$

where $\mathcal{S}_n$ is the set of all permutations of $\{1, \dots, n\}$, and $\operatorname{sgn}(\sigma)$ is the *signature*[1] of a permutation $\sigma \in \mathcal{S}_n$.

A matrix $\bar{M} \in \{0, \star\}^{n \times m}$ is called a *structured matrix*, if $[\bar{M}]_{ij}$ is either a fixed zero or an independent free parameter denoted by $\star$. In particular, we define a matrix $\bar{M} \in \{0, \star\}^{n \times n}$ to be *symmetrically structured*, if the value of the free parameter associated with $[\bar{M}]_{ij}$ is constrained to be the same as the value of the free parameter associated with $[\bar{M}]_{ji}$, for all j and i. For example, consider $\bar{M}$ and $\bar{A}$ be specified by

$$\bar{M} = \begin{bmatrix} 0 & m_{12} \\ m_{21} & m_{22} \end{bmatrix} \text{ and } \bar{A} = \begin{bmatrix} 0 & a_{12} \\ a_{12} & a_{22} \end{bmatrix},$$

where m_{12}, m_{21}, m_{22} and a_{12}, a_{22} are independent parameters. In this case, $\bar{M}$ is a structured matrix whereas $\bar{A}$ is symmetrically structured.

In the rest of the thesis, we refer to $\tilde{M}$ as a *numerical realization* of a (symmetrically) structured matrix $\bar{M}$, i.e., $\tilde{M}$ is a matrix obtained by independently assigning real numbers to each independent free parameter in $\bar{M}$. In addition, we say that the structured matrix $\bar{M} \in \{0, \star\}^{n \times m}$ is the *structural pattern* of the matrix $M \in \mathbb{R}^{n \times m}$, where $[\bar{M}]_{ij} = \star$ if and only if $[M]_{ij} \neq 0$, for $\forall i \in [n], \forall j \in [m]$.

Given a (symmetrically) structured matrix $\bar{M}$, we let $n_{\bar{M}}$ be the number of its independent free parameters and we associate with $\bar{M}$ a parameter space $\mathbb{R}^{n_{\bar{M}}}$. Furthermore, we use vector $\mathbf{p}_{\tilde{M}} = (p_1, \dots, p_{n_{\bar{M}}})^\top \in \mathbb{R}^{n_{\bar{M}}}$ to encode the value of independent free entries of $\bar{M}$ in a numerical realization $\tilde{M}$.

In what follows, a set $V \subseteq \mathbb{R}^n$ is called a *variety* if there exist polynomials $\varphi_1, \dots, \varphi_k$, such that $V = \{x \in \mathbb{R}^n \colon \varphi_i(x) = 0, \forall i \in \{1, \dots, k\}\}$, and V is a *proper variety* when $V \neq$

[1] The signature of a permutation equals to 1 if $|\{(x, y) : x < y, \sigma(x) > \sigma(y)\}|$ is even, and -1 otherwise.

$\mathbb{R}^n$. We denote by $V^c := \mathbb{R}^n \setminus V$ its complement.

The *term rank* [30] of a (symmetrically) structured matrix $\bar{M}$, denoted as t–rank$(\bar{M})$, is the largest integer k such that, for some suitably chosen distinct rows $i_1, \ldots, i_k$ and distinct columns $j_1, \ldots, j_k$, all of the entries $\{[\bar{M}]_{i_\ell j_\ell}\}_{\ell=1}^k$ are $\star$-entries. Additionally, a (symmetrically) structured matrix $\bar{M} \in \{0, \star\}^{n \times m}$ is said to have *generic rank* k, denoted as g–rank$(\bar{M}) = k$, if there exists a numerical realization $\tilde{M}$ of $\bar{M}$, such that rank$(\tilde{M}) = k$. If g–rank$(\bar{M}) > 0$, it is worth noting that the set of parameters describing all possible realizations forms a proper variety when rank$(\tilde{M}) <$ g–rank$(\bar{M})$ [36].

2.1.2 Preliminaries on Graph Theory

In the rest of the paper, we let $G = (\mathcal{V}, \mathcal{E})$ denote a directed graph whose vertex-set and edge-set are denoted by $\mathcal{V} = \{v_1, \ldots, v_n\}$ and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, respectively. A graph G is called *undirected* if $(i, j) \in \mathcal{E}$ implies $(j, i) \in \mathcal{E}$ for all pairs of vertices $i, j \in \mathcal{V}$. Given an edge $(i, j) \in \mathcal{E}$, we say that the 'tail' vertex i is pointing towards the 'head' vertex j, which we denote by $i \to j$. The *order* of G is defined by the number of its vertices. The *out-neighborhood* of vertex $i \in \mathcal{V}$ is defined as $\mathcal{N}_i^+ = \{j \in \mathcal{V} \colon (i, j) \in \mathcal{E}\}$. Similarly, we define the *in-neighborhood* of vertex i as $\mathcal{N}_i^- = \{j \in \mathcal{V} \colon (j, i) \in \mathcal{E}\}$.

A *walk* of length k in G is defined as an ordered sequence of vertices $(i_0, i_1, \ldots, i_k)$ with $(i_\ell, i_{\ell+1}) \in \mathcal{E}$ for all $\ell = 0, \ldots, k-1$. If $i_0 = i_k$, the walk is said to be *closed*; otherwise, the walk is said to be *open*. We say that a vertex $i \in \mathcal{V}$ has a *self-loop* if $(i, i) \in \mathcal{E}$. A *path* $\mathcal{P}$ in G is defined as an ordered sequence of distinct vertices $\mathcal{P} = (v_1, \ldots, v_k)$ with $\{v_1, \ldots, v_k\} \subseteq \mathcal{V}$ and $(v_i, v_{i+1}) \in \mathcal{E}$ for all $i = 1, \ldots, k-1$. A graph contains a *multiedge* if any directed edge appears more than once in $\mathcal{E}$. A digraph is said to be *simple* if the digraph does not have self-loops or multiedges. A *cycle* is either a path $(v_1, \ldots, v_k)$ with an additional edge (v_k, v_1) (denoted as $\mathcal{C} = (v_1, \ldots, v_k, v_1)$), or a vertex with an edge to itself (i.e., self-loop, denoted as cycle $\mathcal{C} = (v_1, v_1)$). We denote by $\mathcal{V}_{\mathcal{C}} \subseteq \mathcal{V}$ the set of vertices in $\mathcal{C}$, and $\mathcal{E}_{\mathcal{C}} \subseteq \mathcal{E}$ the set of edges in $\mathcal{C}$. The length of a cycle $\mathcal{C}$, is defined as the number of distinct vertices in $\mathcal{C}$, and is denoted by $|\mathcal{C}|$. A vertex $v_2 \in \mathcal{V}$ is *reachable* from $v_1 \in \mathcal{V}$ if there exists a path in G from v_1 to v_2.

A directed graph $G_s = (\mathcal{V}_s, \mathcal{E}_s)$ is a *sub-graph* of G if $\mathcal{V}_s \subseteq \mathcal{V}$ and $\mathcal{E}_s \subseteq \mathcal{E}$. In particular, if $\mathcal{V}_s = \mathcal{V}$, then G_s is said to *span* G. Conversely, given a set $\mathcal{S}$ of vertices in G, we let $G_{\mathcal{S}} = (\mathcal{S}, \mathcal{S} \times \mathcal{S} \subset \mathcal{E})$ be the *subgraph of* G *induced by* $\mathcal{S}$. We also call $G_{\mathcal{S}}$ the $\mathcal{S}$*-induced subgraph* of G. We say that $G_{\mathcal{S}}$ can be covered by *disjoint cycles* if there exists $\mathcal{C}_1, \ldots, \mathcal{C}_l$, such that $\mathcal{S} = \bigcup_{i=1}^{l} \mathcal{V}_{\mathcal{C}_i}$ and $\mathcal{V}_{\mathcal{C}_i} \cap \mathcal{V}_{\mathcal{C}_j} = \emptyset$, for all $i \neq j$, $i, j \in \{1, \ldots, l\}$. Given a set $\mathcal{S} \subseteq \mathcal{V}$, we define the *in-neighbour set* of $\mathcal{S}$ as $\mathcal{N}(\mathcal{S}) = \{v_i \in \mathcal{V} | (v_i, v_j) \in \mathcal{E}, v_j \in \mathcal{S}\}$. We say a vertex v_i is *reachable* from vertex v_j in $G(\mathcal{V}, \mathcal{E})$, if there exists a path from vertex v_j to vertex v_i.

A graph is said to be *strongly connected* if there exists a path between any two vertices in the graph. A *strongly connected component* (SCC) is a maximal subgraph G_s that is strongly connected. A *condensation* of G is a *directed acyclic graph* (DAG) generated by representing each SCC in G as a virtual vertex in the condensation and a directed edge between two virtual vertices in the condensation exists, if and only if, there exists a directed edge connecting the corresponding SCCs in G [154]. An SCC is said to be *linked* if it has at least one incoming/outgoing edge from another SCC. In particular, a *source SCC* has no incoming edges from another SCC and a *sink SCC* has no outgoing edges to another SCC.

Given a directed graph $G = (\mathcal{V}, \mathcal{E})$ and two sets $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathcal{V}$, we define the *bipartite graph* $\mathcal{B}(\mathcal{S}_1, \mathcal{S}_2, \mathcal{E}_{\mathcal{S}_1,\mathcal{S}_2})$ as an undirected graph, whose vertex set is $\mathcal{S}_1 \cup \mathcal{S}_2$ and edge set[2] $\mathcal{E}_{\mathcal{S}_1,\mathcal{S}_2} = \{\{s_1, s_2\} \colon (s_1, s_2) \in \mathcal{E}, s_1 \in \mathcal{S}_1, s_2 \in \mathcal{S}_2\}$. Given $\mathcal{B}(\mathcal{S}_1, \mathcal{S}_2, \mathcal{E}_{\mathcal{S}_1,\mathcal{S}_2})$, and a set $\mathcal{S} \subseteq \mathcal{S}_1$ or $\mathcal{S} \subseteq \mathcal{S}_2$, we define *bipartite neighbor set* of $\mathcal{S}$ as $\mathcal{N}_{\mathcal{B}}(\mathcal{S}) = \{j \colon \{j, i\} \in \mathcal{E}_{\mathcal{S}_1,\mathcal{S}_2}, i \in \mathcal{S}\}$. A *matching* $\mathcal{M}$ is a set of edges in $\mathcal{E}_{\mathcal{S}_1,\mathcal{S}_2}$ that do not share vertices, i.e., given edges $e = \{s_1, s_2\}$ and $e' = \{s_1', s_2'\}$, $e, e' \in \mathcal{M}$ only if $s_1 \neq s_1'$ and $s_2 \neq s_2'$. The vertex v is said to be *right-unmatched* with respect to a matching $\mathcal{M}$ associated with $\mathcal{B}(\mathcal{S}_1, \mathcal{S}_2, \mathcal{E}_{\mathcal{S}_1,\mathcal{S}_2})$ if $v \in \mathcal{S}_2$, and v does not belong to an edge in the matching $\mathcal{M}$. A matching is said to be maximum if it is a matching with the maximum number of edges

[2] We denote undirected edges using curly brackets $\{v_i, v_j\}$, in contrast with directed edges, for which we use parenthesis.

among all possible matchings. Additionally, a matching is called a *perfect matching* if it does not contain right-unmatched vertices. Given a bipartite graph $\mathcal{B}(S_1, S_2, \mathcal{E}_{S_1,S_2})$, the maximum matching problem can be solved efficiently in $\mathcal{O}(\sqrt{|S_1 \cup S_2|}|\mathcal{E}_{S_1,S_2}|)$ time [154].

2.2 Problem Statements

We consider a linear time-invariant system whose dynamics is captured by

$$\dot{x} = Ax + Bu, \quad y = Cx, \tag{2.2}$$

where $x \in \mathbb{R}^n$, $y \in \mathbb{R}^k$ and $u \in \mathbb{R}^m$ are the state, the output and the input vectors, respectively. In addition, the matrix $A \in \mathbb{R}^{n\times n}$ is the state matrix, $B \in \mathbb{R}^{n\times m}$ is the input matrix and $C \in \mathbb{R}^{k\times n}$ is the output matrix. In this chapter, we consider the following assumption:

Assumption 1. *The state matrix $A \in \mathbb{R}^{n\times n}$ is symmetric, i.e., $A = A^\top$.*

This symmetry assumption is motivated by control problems arising in undirected networked dynamical systems. Furthermore, this assumption will be crucial when establishing graph-theoretic results characterizing structural controllability problems in undirected networks. Hereafter, we use the 3-tuple (A, B, C) to represent the system (2.2). In particular, we use the pair (A, B) to denote a system without a measured output. A pair (A, B) is called *reducible* if there exists a permutation matrix P, such that

$$PAP^{-1} = \begin{bmatrix} A_{11} & \mathbf{0} \\ A_{21} & A_{22} \end{bmatrix}, \quad PB = \begin{bmatrix} \mathbf{0} \\ B_2 \end{bmatrix}, \tag{2.3}$$

where $A_{11} \in \mathbb{R}^{q\times q}$ and $B_2 \in \mathbb{R}^{(n-q)\times m}$, $1 \leq q < n$. The pair (A, B) is called *irreducible* otherwise. Furthermore, we use $\bar{A}$ and $\bar{B}$ to represent the structural pattern of A and B, respectively. In particular, by Assumption 1, we consider $\bar{A}$ to be symmetrically structured. Thus, $(\bar{A}, \bar{B})$ is referred to as the *structural pair* of the system (A, B).

Given a structured matrix $\bar{A}$, we associate it with a directed graph $G(\bar{A}) = (\mathcal{X}, \mathcal{E}_{\mathcal{X},\mathcal{X}})$, which we refer to as the *state digraph*, where $\mathcal{X} = \{x_1, \ldots, x_n\}$ is the state vertex set, and $\mathcal{E}_{\mathcal{X},\mathcal{X}} = \{(x_j, x_i) : [\bar{A}]_{ij} = \star\}$ is the set of edges. Similarly, we associate a directed graph $G(\bar{A}, \bar{B}) = (\mathcal{X} \cup \mathcal{U}, \mathcal{E}_{\mathcal{X},\mathcal{X}} \cup \mathcal{E}_{\mathcal{U},\mathcal{X}})$ with the structural pair $(\bar{A}, \bar{B})$, where $\mathcal{U} = \{u_1, \ldots, u_m\}$ is the set of input vertices and $\mathcal{E}_{\mathcal{U},\mathcal{X}} = \{(u_j, x_i) \colon [\bar{B}]_{ij} = \star\}$ is the set of edges from input vertices to state vertices. We refer to $G(\bar{A}, \bar{B})$ as the *system digraph*.

Definition 1 (Structural Controllability [31])**.** *A structural pair $(\bar{A}, \bar{B})$ is structurally controllable if there exists a numerical realization $(\tilde{A}, \tilde{B})$, such that the controllability matrix $Q(\tilde{A}, \tilde{B}) := [\tilde{B}, \tilde{A}\tilde{B}, \ldots, \tilde{A}^{n-1}\tilde{B}]$ has full row rank.*

While controllability is concerned about the ability to steer all the states of a system to a desired final state, under certain circumstances, it is more preferred to control the behavior of only a subset of states. More specifically, given a set $\mathcal{T} \subseteq [n]$, which we refer to as the *target set*, it is of interest to consider whether the set of selected states can be steered arbitrarily. If so, we say that the pair (A, B) is *target controllable* with respect to $\mathcal{T}$ [43]. Notice that this does not exclude the possibility of some other states indexed by $[n] \setminus \mathcal{T}$ being controllable as well. Similarly, we introduce the notion of *structural target controllability* in the context of structural pairs.

Definition 2 (Structural Target Controllability [44])**.** *Given a structural pair $(\bar{A}, \bar{B})$, and a target set $\mathcal{T} = \{i_1, \ldots, i_k\} \subseteq [n]$, let $\mathcal{X}_\mathcal{T}$ be the set of state vertices corresponding to $\mathcal{T}$ in $G(\bar{A}, \bar{B})$. We define a matrix $C_\mathcal{T} \in \mathbb{R}^{k \times n}$ by*

$$[C_\mathcal{T}]_{\ell j} = \begin{cases} 1, & \textit{if } j = i_\ell,\ i_\ell \in \mathcal{T}, \\ 0, & \textit{otherwise}. \end{cases} \tag{2.4}$$

The structural pair $(\bar{A}, \bar{B})$ is structurally target controllable with respect to $\mathcal{T}$ if there exists a numerical realization $(\tilde{A}, \tilde{B})$, such that the target controllability matrix $Q_\mathcal{T}(\tilde{A}, \tilde{B}) := C_\mathcal{T}[\tilde{B}, \tilde{A}\tilde{B}, \ldots, \tilde{A}^{n-1}\tilde{B}]$ has full row rank.

Note that structural controllability is equivalent to structural target controllability when $\mathcal{T} = [n]$. Therefore, the necessary and sufficient conditions for structurally target controllable undirected networks can be applied to characterize structural controllability. Subsequently, in this paper, we consider the following problem:

Problem 1. *Given a structural pair $(\bar{A}, \bar{B})$, where $\bar{A}$ is symmetrically structured and $\bar{B}$ is a structured matrix, and a target set $\mathcal{T} \subseteq [n]$, find a necessary and sufficient condition for $(\bar{A}, \bar{B})$ to be structurally target controllable with respect to $\mathcal{T}$.*

2.3 Characterizing Structural Properties using Graph Theory

In this section, we first introduce a proposition that is crucial for developing our solution to Problem 1. Then, we characterize the generic rank of symmetrically structured matrices in Lemma 1. Subsequently, we characterize the relationship between the term-rank of a symmetrically structured matrix and the presence of non-zero simple eigenvalues in a numerical realization in Lemma 2. This allows us to obtain a result characterizing the relationship between irreducibility and structural controllability of a structural pair involving symmetrically structured matrix (see Lemma 3). Based on these results, we propose graph-theoretic necessary and sufficient conditions for structural controllability and structural target controllability in Theorems 1 and 2, respectively.

Proposition 1 (Popov-Belevitch-Hautus (PBH) test [155])**.** *The pair (A, B) is uncontrollable if and only if there exists a $\lambda \in \mathbb{C}$ and a nontrivial vector $e \in \mathbb{C}^n$, such that $e^\top A = \lambda e^\top$ and $e^\top B = 0$.*

Given a pair (A, B), where $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times m}$, we say that the mode (λ, e) of A, where $\lambda \in \mathbb{C}$ and $e \in \mathbb{C}^n$, is an *uncontrollable mode* if $e^\top A = \lambda e^\top$ and $e^\top B = 0$.

2.3.1 Generic Properties of Symmetrically-structured Matrices

If a symmetrically structured matrix is generically full rank, then any numerical realization has almost surely no zero eigenvalue. In this subsection, we characterize the generic rank of a symmetrically structured matrix in Lemma 1, which lays the foundation for a further characterization of spectral properties of numerical realizations.

Lemma 1. *Consider an $n \times n$ symmetrically structured matrix $\bar{A}$, and a set $\mathcal{T} = \{i_1, \ldots, i_k\} \subseteq [n]$. Let $G(\bar{A}) = (\mathcal{X}, \mathcal{E}_{\mathcal{X},\mathcal{X}})$ be the digraph representation of $\bar{A}$, $\mathcal{X}_{\mathcal{T}} \subseteq \mathcal{X}$ be the set of vertices indexed by $\mathcal{T}$, and $C_{\mathcal{T}}$ be defined as in (2.4). The generic-rank of $C_{\mathcal{T}}\bar{A}$ equals to k if and only if $|\mathcal{N}(\mathcal{S})| \geq |\mathcal{S}|$, $\forall \mathcal{S} \subseteq \mathcal{X}_{\mathcal{T}}$.*

Proof. See Appendix A.1. □

Lemma 1 establishes a relationship between the generic rank of a submatrix of a symmetrically structured matrix and the topology of its corresponding digraph. Subsequently, Corollary 1 follows, which characterizes the generic rank of the concatenation of a symmetrically structured matrix $\bar{A} \in \{0, \star\}^{n \times n}$ and a structured matrix $\bar{B} \in \{0, \star\}^{n \times m}$.

Corollary 1. *Consider a structural pair $(\bar{A}, \bar{B})$, where $\bar{A}$ is symmetrically structured, and a set $\mathcal{T} = \{i_1, \ldots, i_k\} \subseteq [n]$. Let $G(\bar{A}, \bar{B}) = (\mathcal{X} \cup \mathcal{U}, \mathcal{E}_{\mathcal{X},\mathcal{X}} \cup \mathcal{E}_{\mathcal{U},\mathcal{X}})$ be the digraph representation of $(\bar{A}, \bar{B})$, and $\mathcal{X}_{\mathcal{T}} \subseteq \mathcal{X}$ be the set of vertices indexed by $\mathcal{T}$. If $|\mathcal{N}(\mathcal{S})| \geq |\mathcal{S}|$, $\forall \mathcal{S} \subseteq \mathcal{X}_{\mathcal{T}}$, then* $\text{g–rank}(C_{\mathcal{T}}[\bar{A}, \bar{B}]) = k$.

Proof. See Appendix A.1. □

In the remaining subsections, we aim to provide necessary and sufficient conditions for structural controllability. To achieve this goal, we notice that the eigenvalues of the state matrix are closely related to controllability, as indicated by Proposition 1. Besides, the approach in [35] shows that for an irreducible structural pair with no symmetric parameter dependencies, all the nonzero modes of its numerical realization are almost surely simple and controllable. Similarly, to characterize structural controllability of undirected networks, we will provide characterizations of the modes in the numerical realization of a structural pair involving symmetrically structured matrix. Instead of using the maximum order of principle minor as in [35], we derive below a condition based on the term rank to ensure that generically the numerical realization of a symmetrically structured matrix has k nonzero simple eigenvalues.

Lemma 2. *Given an $n \times n$ symmetrically structured matrix $\bar{A}$, if* $\text{t–rank}(\bar{A}) = k$, *then there exists a proper variety $V_1 \subset \mathbb{R}^{n_{\bar{A}}}$, such that for any numerical realization $\tilde{A}$,* *where the numerical values assigned to free parameters of $\bar{A}$ are encoded in the vector $\mathbf{p}_{\tilde{A}} \in \mathbb{R}^{n_{\bar{A}}} \setminus V_1$, $\tilde{A}$ has k nonzero simple eigenvalues.*

Proof. See Appendix A.1. □

Remark 1. *The challenge in the proof of Lemma 2 is to construct a finite number of nonzero polynomials, i.e., the polynomials of where not every coefficient is zero, such that the numerical values assigned to free parameters of $\bar{A}$ in a numerical realization $\tilde{A}$, where $\tilde{A}$ does not have k nonzero simple eigenvalues, are the zeros of those polynomials. Since the set of zeros of a nonzero polynomial has Lebesgue measure zero [156], it follows that for any numerical realization $\tilde{A}$, $\tilde{A}$ has almost surely k nonzero simple eigenvalues.*

Remark 2. *Lemma 2 generally is not true for a structured matrix. For example, consider $\bar{M} = \begin{bmatrix} 0,\star \\ 0,0 \end{bmatrix}$, $\text{t–rank}(\bar{M}) = 1$, but for any numerical realization, $\tilde{M}$ has no nonzero mode.*

As shown in [31, 35], irreducibility is a necessary condition for structural controllability. We can expect that irreducibility also plays a similar role in symmetrically structured systems. Moreover, we show below that irreducibility ensures that all nonzero simple modes of $\tilde{A}$ are controllable, generically.

Lemma 3. *Given a structural pair $(\bar{A}, \bar{B})$, where $\bar{A}$ is symmetrically structured and t–rank$(\bar{A}) = k$, if $(\bar{A}, \bar{B})$ is irreducible, then there exists a proper variety $V \subset \mathbb{R}^{n_{\bar{A}}+n_{\bar{B}}}$, such that for any numerical realization $(\tilde{A}, \tilde{B})$ with $[\mathbf{p}_{\tilde{A}}, \mathbf{p}_{\tilde{B}}] \in \mathbb{R}^{n_{\bar{A}}+n_{\bar{B}}} \setminus V$, $\tilde{A}$ has k nonzero, simple and controllable modes.*

Proof. See Appendix A.1. □

2.3.2 Symmetric Structural Controllability

We have shown that irreducibility guarantees that generically all non-zero simple modes of $(\tilde{A}, \tilde{B})$ are controllable. Subsequently, to characterize the controllability, it remains to find conditions to ensure that generically all the zero modes are also controllable. In this subsection, Theorem 1 proposes conditions guaranteeing that generically both the nonzero and zero modes of $(\tilde{A}, \tilde{B})$ are controllable, therefore establishes a graph-theoretic necessary and sufficient condition for structural controllability in symmetrically structured system.

Theorem 1. *Let $(\bar{A}, \bar{B})$ be a structural pair, with $\bar{A}$ being a symmetrically structured matrix, and let $\mathcal{X}$ be the set of state vertices in $G(\bar{A}, \bar{B})$. The structural pair $(\bar{A}, \bar{B})$ is structurally controllable, if and only if, the following conditions hold simultaneously in $G(\bar{A}, \bar{B})$:*

1. *all the state vertices are input-reachable;*

2. $|\mathcal{N}(\mathcal{S})| \geq |\mathcal{S}|$, $\forall \mathcal{S} \subseteq \mathcal{X}$.

Proof. See Appendix A.1. □

Notice that Conditions *1)* and *2)* in Theorem 1 admits a similar form as the conditions for structural controllability (see, for example [31]). Subsequently, if a structural pair with symmetric parameter dependencies is structurally controllable, then the structural pair with the same structural pattern without symmetric parameter dependencies, will also be structurally controllable. However, the converse cannot be trivially derived due to symmetric parameter dependencies in Assumption 1.

2.3.3 Symmetric Structural Target Controllability

We now extend the solution approach in Theorem 1 to establish graph-theoretic necessary and sufficient conditions for structural target controllability of the given structural pair $(\bar{A}, \bar{B})$ and target set $\mathcal{T}$.

Theorem 2. *Consider a structural pair $(\bar{A}, \bar{B})$, with $\bar{A}$ being symmetrically structured, and a target set $\mathcal{T} \subseteq [n]$. Let $\mathcal{X}_{\mathcal{T}}$ be the set of state vertices corresponding to $\mathcal{T}$ in $G(\bar{A}, \bar{B})$. The structural pair $(\bar{A}, \bar{B})$ is structurally target controllable with respect to $\mathcal{T}$, if and only if, the following conditions hold simultaneously in $G(\bar{A}, \bar{B})$:*

1. *all the states vertices in $\mathcal{X}_{\mathcal{T}}$ are input-reachable;*

2. $|\mathcal{N}(\mathcal{S})| \geq |\mathcal{S}|$, $\forall \mathcal{S} \subseteq \mathcal{X}_{\mathcal{T}}$.

Proof. See Appendix A.1. □

Remark 3. *Condition 2) in Theorem 2 can be verified using local topological information in the network. In particular, this condition is satisfied if there exists a matching in the bipartite graph $\mathcal{B}(\mathcal{S}_1, \mathcal{S}_2, \mathcal{E}_{\mathcal{S}_1,\mathcal{S}_2})$ associated with $G(\bar{A}, \bar{B})$, where $\mathcal{S}_1 = \mathcal{X} \cup \mathcal{U}$ and $\mathcal{S}_2 = \mathcal{X}_{\mathcal{T}}$, such that all vertices in $\mathcal{S}_2$ are right-matched. The existence of such a matching can be verified in $\mathcal{O}(\sqrt{|\mathcal{S}_1 \cup \mathcal{S}_2|}|\mathcal{E}_{\mathcal{S}_1,\mathcal{S}_2}|)$ time [154, §23.6].*

Through the proof of Theorem 2, we notice that the characterization of structural target controllability relies on the assumption that the state matrix is symmetric. More specifically, since the state matrix is symmetric, the eigenvectors of the state matrix form a complete basis of the state space, which allows us to generalize the PBH test in the context of target controllability problems. On the contrary, when the system is characterized by a directed network, the state matrix A is, in general, non-diagonalizable, which prevents us from generalizing PBH test to characterize the target controllability problems - see [48, Example 3] for a reference.

In addition, the proof of Theorem 2 suggests that, even for the case where $\bar{A}$ is not symmetrically structured, the violation of either Conditions *1)* or *2)* results in that $(\bar{A}, \bar{B})$ is not structurally target controllable. Therefore, in general, when the structured matrix $\bar{A} \in \{0, \star\}^{n \times n}$ is not symmetrically structured, the Conditions *1)* and *2)* in Theorem 2 are necessary but not sufficient conditions for the structural target controllability of the pair $(\bar{A}, \bar{B})$.

2.4 Extensions to Structural Output Controllability

Next, we utilize the developed framework to analyze a more general notion of controllability, defined as follows.

Definition 3 (Structural Output Controllability)**.** *A structural triple $(\bar{A}, \bar{B}, \bar{C})$ is structurally output controllable if there exists a numerical realization $(\tilde{A}, \tilde{B}, \tilde{C})$ such that the output controllability matrix $Q(\tilde{A}, \tilde{B}, \tilde{C}) = \tilde{C}[\tilde{B}, \tilde{A}\tilde{B}, \cdots, \tilde{A}^{n-1}\tilde{B}]$ has full row rank.*

Notice that structural target controllability is a special case of structural output controllability [44], provided that C takes the particular form in (2.4). In the context of output controllability, each output is regarded as a weighted linear combination of a collection of states. This underlying connection between structural target controllability and structural output controllability motivates us to use the results we have developed to provide necessary and sufficient conditions for structural output controllability.

By utilizing Theorem 2 and arguing the relationship between a target set and the state-to-output connections characterized by the output matrix C, we can characterize graph-theoretic conditions for (symmetric) structural output controllability, as shown in the following theorem.

Theorem 3. *Consider a structural system $(\bar{A}, \bar{B}, \bar{C})$, where $\bar{A}$ is a symmetrically structured matrix, while $\bar{B}, \bar{C}$ are structured matrices. The structural system $(\bar{A}, \bar{B}, \bar{C})$ is structurally output controllable, if and only if, the following conditions hold simultaneously:*

1. *there exists a target set $\mathcal{T} \subseteq [n]$ such that $(\bar{A}, \bar{B})$ is structurally target controllable with respect to $\mathcal{T}$;*
2. *there is no right-unmatched vertex in $\mathcal{B}(\mathcal{X}_{\mathcal{T}}, \mathcal{Y}, \mathcal{E}_{\mathcal{X}_{\mathcal{T}},\mathcal{Y}})$, where $\mathcal{Y} = \{y_i\}_{i=1}^{k}$, $\mathcal{X}_{\mathcal{T}} = \{x_i \in \mathcal{X} \colon i \in \mathcal{T}\}$, and $\mathcal{E}_{\mathcal{X}_{\mathcal{T}},\mathcal{Y}} = \{\{x_j, y_i\} \colon [\bar{C}]_{ij} = \star\}$.*

Proof. See Appendix A.1. □

The conditions in Theorem 3 require us to find a target set $\mathcal{T}$ for which a matching condition in a bipartite graph $\mathcal{B}(\mathcal{X}_{\mathcal{T}}, \mathcal{Y}, \mathcal{E}_{\mathcal{X}_{\mathcal{T}},\mathcal{Y}})$ is satisfied. Naively, there are exponentially many possible target sets $\mathcal{T}$, implying that it may be computationally challenging to verify structural output controllability through the conditions in Theorem 3 Indeed, we show in Theorem 4 that verifying those conditions is NP-hard:

Theorem 4. *Consider a structural system $(\bar{A}, \bar{B}, \bar{C})$, where $\bar{A} \in \{0, \star\}^{n \times n}$ is a symmetrically structured matrix. The problem of verifying the necessary and sufficient conditions in Theorem 3 is NP-hard.*

Proof. See Appendix A.1. □

2.5 Illustrative Examples

2.5.1 Examples on Symmetric Structural Controllability

In this section, we provide an example to illustrate our necessary and sufficient conditions in Theorem 1 and Theorem 2. We consider a symmetrically structured system with 10 states and 2 inputs modeled by an undirected network with unknown link weights. The structural representations of its state and input matrix are denoted by $\bar{A} \in \{0, \star\}^{10\times 10}$ and $\bar{B} \in \{0, \star\}^{10\times 2}$, as follows.

$$\bar{A} = \begin{bmatrix} 0 & a_{12} & 0 & a_{14} & a_{15} & 0 & 0 & 0 & 0 & 0 \\ a_{12} & 0 & a_{23} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & a_{23} & 0 & a_{34} & 0 & 0 & 0 & 0 & 0 & 0 \\ a_{14} & 0 & a_{34} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ a_{15} & 0 & 0 & 0 & 0 & 0 & a_{57} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & a_{66} & a_{67} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & a_{57} & a_{67} & 0 & 0 & a_{79} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & a_{89} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & a_{79} & a_{89} & 0 & a_{910} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & a_{910} & 0 \end{bmatrix}, \bar{B} = \begin{bmatrix} 0 & b_{12} \\ b_{21} & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & b_{52} \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

In addition, we let the target set be $\mathcal{T} = \{2, 6, 8\}$. Subsequently, $C_{\mathcal{T}}$, defined according to (2.4), is equal to

$$C_{\mathcal{T}} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

We also associate the structural pair $(\bar{A}, \bar{B})$ with the digraph $G(\bar{A}, \bar{B}) = (\mathcal{X} \cup \mathcal{U}, \mathcal{E}_{\mathcal{X},\mathcal{X}} \cup \mathcal{E}_{\mathcal{U},\mathcal{X}})$ as depicted in Figure 5-1, where $\mathcal{X} = \{x_1, \ldots, x_{10}\}$, $\mathcal{U} = \{u_1, u_2\}$ and $\mathcal{X}_{\mathcal{T}} = \{x_2, x_6, x_8\}$. Notice that by letting $\mathcal{S} = \{x_8, x_{10}\}$, we have $\mathcal{N}(\mathcal{S}) = \{x_9\}$. As a re-

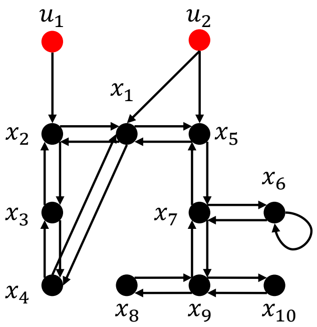


Figure 2-1: The digraph representation of the structural pair $(\bar{A}, \bar{B})$, where the red and black vertices represent input and state vertices, respectively. The black arrows represent edges in $G(\bar{A}, \bar{B})$.

sult, according to Theorem 1, $\exists \mathcal{S} \subseteq \mathcal{X}$, $|\mathcal{N}(\mathcal{S})| < |\mathcal{S}|$ implies that the system is not structurally controllable.

However, since all the vertices in $\mathcal{X}_\mathcal{T}$ are input-reachable, and $|\mathcal{N}(\mathcal{S})| \geq |\mathcal{S}|$, $\forall \mathcal{S} \subseteq \mathcal{X}_\mathcal{T}$, by Theorem 2, $(\bar{A}, \bar{B})$ is structurally target controllable with respect to $\mathcal{T}$. This example also shows that, if the input-reachability of the vertices in $\mathcal{X}_\mathcal{T}$ is guaranteed, then the structural target controllability in undirected networks can be verified by only local topological information.

2.5.2 Examples on Symmetric Structural Output Controllability

In this subsection, we provide an example to illustrate Theorems 2 and 3 We consider a symmetrically structured system with 7 states, 2 inputs, and 3 outputs. Let the target set be $\mathcal{T} = \{2, 4, 6\}$. The structural representations of the state, input, output, and target matrices are,

$$\bar{A} = \begin{bmatrix} 0 & a_{12} & a_{13} & a_{14} & 0 & 0 & 0 \\ a_{12} & 0 & 0 & 0 & 0 & 0 & 0 \\ a_{13} & 0 & 0 & 0 & 0 & 0 & 0 \\ a_{14} & 0 & 0 & a_{44} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & a_{56} & a_{57} \\ 0 & 0 & 0 & 0 & a_{56} & 0 & a_{67} \\ 0 & 0 & 0 & 0 & a_{57} & a_{67} & 0 \end{bmatrix}, \bar{B} = \begin{bmatrix} b_{11} & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & b_{42} \\ 0 & b_{52} \\ 0 & 0 \\ 0 & 0 \end{bmatrix},$$

$$\bar{C} = \begin{bmatrix} 0 & c_{12} & 0 & c_{14} & 0 & 0 & 0 \\ 0 & 0 & 0 & c_{24} & 0 & 0 & 0 \\ 0 & 0 & 0 & c_{34} & 0 & c_{36} & c_{27} \end{bmatrix}, C_\mathcal{T} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

We also associate the structural pair $(\bar{A}, \bar{B})$ with the mixed graph $\mathcal{G}(\bar{A}, \bar{B}) = (\mathcal{X} \cup \mathcal{U}, \mathcal{E}_u(\bar{A}), \mathcal{E}_{\mathcal{U},\mathcal{X}})$, depicted in Figure 5-1, where $\mathcal{X} = \{x_i\}_{i=1}^{7}$, $\mathcal{U} = \{u_1, u_2\}$ and $\mathcal{X}_\mathcal{T} = \{x_2, x_4, x_6\}$. Notice that by letting $\mathcal{S} = \{x_2, x_3\}$, we have $\mathcal{N}(\mathcal{S}) = \{x_1\}$. As a re-

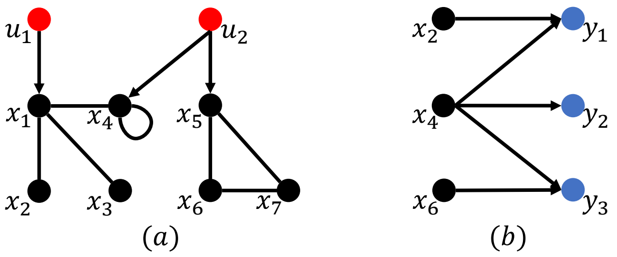


Figure 2-2: The subfigure (a) is the mixed graph representation of the structural pair $(\bar{A}, \bar{B})$, where the red and black vertices represent input and state vertices, respectively. The black lines and arrows represent edges in $G(\bar{A}, \bar{B})$; The subfigure (b) is the bipartite graph $\mathcal{B}(\mathcal{X}_{\mathcal{T}}, \mathcal{Y}, \mathcal{E}_{\mathcal{X}_{\mathcal{T}},\mathcal{Y}})$, where $\mathcal{X}_{\mathcal{T}} = \{x_2, x_4, x_6\}$ and $\mathcal{Y} = \{y_1, y_2, y_3\}$. The black and blue vertices are target vertices $\mathcal{X}_{\mathcal{T}}$ and output vertices $\mathcal{Y}$, respectively.

sult, according to Theorem 1, $\exists \mathcal{S} \subseteq \mathcal{X}$, $|\mathcal{N}(\mathcal{S})| < |\mathcal{S}|$ implies that the system is not structurally controllable. However, since all the vertices in $\mathcal{X}_{\mathcal{T}}$ are input-reachable, and $|\mathcal{N}(\mathcal{S})| \geq |\mathcal{S}|$, $\forall \mathcal{S} \subseteq \mathcal{X}_{\mathcal{T}}$, by Theorem 2, $(\bar{A}, \bar{B})$ is structurally target controllable with respect to $\mathcal{T}$. This example also shows that, if the input-reachability of the vertices in $\mathcal{X}_{\mathcal{T}}$ is guaranteed, then the structural target controllability in undirected networks can be verified by only local topological information. Finally, to verify the structural output controllability, we notice that there exists a target set $\mathcal{T} = \{2, 4, 6\}$ such that $(\bar{A}, \bar{B})$ is structurally target controllable with respect to $\mathcal{T}$ and there is no right-unmatched vertices with respect to any maximum matching in $\mathcal{B}(\mathcal{X}_{\mathcal{T}}, \mathcal{Y}, \mathcal{E}_{\mathcal{X}_{\mathcal{T}},\mathcal{Y}})$, where $\mathcal{Y} = \{y_1, y_2, y_3\}$ is the set of output vertices. By Theorem 3, $(\bar{A}, \bar{B}, \bar{C})$ is structurally output controllable.

Chapter 3

Topology Design in Asymmetric Linear Structural Systems

Structural controllability extends the classical controllability concept to the case of networks with uncertain edges and it is a generic property, i.e., a structural system is structurally controllable if it is controllable for almost all realizations of edge weights. Subsequently, whenever a system is not structurally controllable, one cannot retain controllablity by designing the weights of parameters. In order to retain controllability, it is necessary to perturb the structure of the system. In this chapter, we consider the perturbation in the form of edge additions. More precisely, given a structurally uncontrollable system, we aim to add new *directed* edges to the system to retain structural controllablity. It is worth noting that we do not consider symmetric parameter dependencies in this chapter, and the case when symmetric constraint is concerned will be explored later in Chapter 4.

The rest of this chapter is organized as follows. A formal description of the problem under consideration are introduced in Section 3.1. Preliminaries on graph theory and structural system theory are introduced in Section 3.2. The main results are provided in Section 3.3. In Section 3.4, we illustrate our results in several complex network topologies.

3.1 Problem Statements

The dynamics of a linear networked dynamical system can be described as follows:

$$\dot{x}(t) = Ax(t) + Bu(t), \tag{3.1}$$

where $x(t) \in \mathbb{R}^n$ is the state vector, $u(t) \in \mathbb{R}^m$ is the input vector, $A \in \mathbb{R}^{n\times n}$ is the state transition matrix and $B \in \mathbb{R}^{n\times m}$ is the input matrix. In the sequel, we refer to the system (3.1) by the matrix pair (A, B), and if the system is controllable, we say that the pair (A, B) is controllable. Furthermore, we define $\bar{A} \in \{0, 1\}^{n\times n}$ to be the structural pattern of A, i.e., $\bar{A}_{ij} = 0$ if $[A]_{ij} = 0$, and $\bar{A}_{ij} = 1$ otherwise. Similarly, $\bar{B} \in \{0, 1\}^{n\times m}$ encodes the sparsity pattern of B where $\bar{B}_{ij} = 0$ of $[B]_{ij} = 0$, and $\bar{B}_{ij} = 1$ otherwise.

We say that the structural pattern $(\bar{A}, \bar{B})$ is *structually controllable* if there exists a pair $(\hat{A}, \hat{B})$ with the same structural pattern as $(\bar{A}, \bar{B})$ that is controllable – see Definition 1 or [157]. Furthermore, if such pair $(\hat{A}, \hat{B})$ exists, then almost all possible matrix pairs with the same structural pattern as $(\bar{A}, \bar{B})$ are controllable [157].

Given a structurally uncontrollable pair $(\bar{A}, \bar{B})$, we are interested in the problem of adding a minimum number of entries in $\bar{A}$ to obtain a structurally controllable system. Intuitively, if we add sufficient edges in the network such that the resulting network is a complete graph, then the resulting system is structurally controllable, provided that at least one node is actuated, i.e., $\bar{B} \neq 0$. Nonetheless, adding new edges corresponds, in practice, to building new infrastructure. Therefore, from a design and implementation perspective, one seeks to add the minimum number of edges to attain the design objective, which, in our case, consists in ensuring structural controllability. Formally, the problem is described as follows:

Problem 2. *Given the pair* $(\bar{A}, \bar{B})$ *with* $\bar{B} \neq 0$, *find*

$$\begin{aligned} \tilde{A}^* = \arg \min_{\tilde{A}\in\{0,1\}^{n\times n}} \quad & \|\tilde{A}\|_0 \\ \textit{s.t.} \ (\bar{A} + \tilde{A}, \bar{B}) & \textit{ is structurally controllable,} \end{aligned} \tag{3.2}$$

where $\|\tilde{A}\|_0$ *denotes the number of non-zero entries in a matrix* $\tilde{A}$*, and the operator* $+ :$ $\{0,1\}^{n\times n} \times \{0,1\}^{n\times n} \to \{0,1\}^{n\times n}$ *is the element-wise exclusive-or for binary matrices.* ∘

If $(\bar{A} + \tilde{A}, \bar{B})$ is structurally controllable, we refer to the matrix $\tilde{A}$ as a *feasible edge-addition matrix*, and to $\tilde{A}^*$ in (3.2) as the *optimal edge-addition matrix*. As part of the solution proposed in this paper, we provide a characterization of all possible optimal edge-addition matrices by resorting to graph-theoretical tools. Further, we provide a polynomial-time algorithm to obtain one such solution.

3.2 General Structural Controllability

In the rest of the chapter, we adopt standard notations in graph theory introduced in Subsection 2.1.2 of Chapter 2. Furthermore, to introduce our problem, we recall the following definitions from Subsection 2.2 in Chapter 2.

Given a structural pair $(\bar{A}, \bar{B})$, we associate a directed graph $G(\bar{A}, \bar{B}) = (\mathcal{X} \cup \mathcal{U}, \mathcal{E}_{\mathcal{X},\mathcal{X}} \cup \mathcal{E}_{\mathcal{U},\mathcal{X}})$, which we refer to as the *system digraph*, where $\mathcal{X} = \{x_1, \ldots, x_n\}$ and $\mathcal{U} = \{u_1, \ldots, u_m\}$ denote the set of state vertices and input vertices. Moreover, the sets $\mathcal{E}_{\mathcal{X},\mathcal{X}} = \{(x_i, x_j) \colon [\bar{A}]_{ji} \neq 0\}$ and $\mathcal{E}_{\mathcal{U},\mathcal{X}} = \{(u_j, x_i) \colon [\bar{B}]_{ij} \neq 0\}$ denote the edge sets of G.

In the remaining of this chapter, unless otherwise specified, a state vertex being reachable means that it is reachable from some input vertex. Similarly, a vertex set is reachable if every vertex in the set is reachable. Also, due to the graph representation of the pair $(\bar{A}, \bar{B})$, when $(\bar{A}, \bar{B})$ is structural controllable, we interchangeably say that $G(\bar{A}, \bar{B})$ is structurally controllable. In addition, we can associate an undirected bipartite graph with $G(\bar{A}, \bar{B})$, called the *system bipartite graph* and denoted by $\mathcal{B}(\bar{A}, \bar{B}) = \mathcal{B}(\mathcal{X}^+ \cup \mathcal{U}^+, \mathcal{X}^-, \mathcal{E}_{\mathcal{X}^+,\mathcal{X}^-} \cup \mathcal{E}_{\mathcal{U}^+,\mathcal{X}^-})$, in which $\{x_i^+, x_j^-\} \in \mathcal{E}_{\mathcal{X}^+,\mathcal{X}^-}$ if $(x_i, x_j) \in \mathcal{E}_{\mathcal{X},\mathcal{X}}$, and $\{u_i^+, x_j^-\} \in \mathcal{E}_{\mathcal{U}^+,\mathcal{X}^-}$ if $(u_i, x_j) \in \mathcal{E}_{\mathcal{U},\mathcal{X}}$.

For ease of notation, we use a *signal-notation mapping* $s \colon \mathcal{E}_{\mathcal{X},\mathcal{X}} \cup \mathcal{E}_{\mathcal{U},\mathcal{X}} \to \mathcal{E}_{\mathcal{X}^+,\mathcal{X}^-} \cup \mathcal{E}_{\mathcal{U}^+,\mathcal{X}^-}$ to map edges from the system digraph into edges of the system bipartite graph, as follows: $s((u_i, x_j)) = \{u_i^+, x_j^-\}$ and $s((x_i, x_j)) = \{x_i^+, x_j^-\}$. In addition, due to the bijectivity of the signal-notation mapping, we have that $s^{-1}(\{u_i^+, x_j^-\}) = (u_i, x_j)$ and $s^{-1}(\{x_i^+, x_j^-\}) = (x_i, x_j)$.

The concepts introduced in this section can be used to determine if a structural system is structurally controllable, as follows:

Theorem 5 ([34, 49])**.** *The pair $(\bar{A}, \bar{B})$ is structurally controllable if and only if the following two conditions hold:*

(a) every state vertex $x \in \mathcal{X}$ in the system digraph $G(\bar{A}, \bar{B}) = (\mathcal{X} \cup \mathcal{U}, \mathcal{E}_{\mathcal{X},\mathcal{X}} \cup \mathcal{E}_{\mathcal{U},\mathcal{X}})$ is reachable (from some input vertex $u \in \mathcal{U}$);

(b) any maximum matching M of the system bipartite graph

$$\mathcal{B}(\bar{A}, \bar{B}) = \mathcal{B}(\mathcal{X}^+ \cup \mathcal{U}^+, \mathcal{X}^-, \mathcal{E}_{\mathcal{X}^+,\mathcal{X}^-} \cup \mathcal{E}_{\mathcal{U}^+,\mathcal{X}^-})$$

has no right-unmatched vertices. ◇

Notice that both conditions in Theorem 5 can be verified in polynomial time [34]. Hence, one could naively try to ensure both conditions by adding edges iteratively, but such an approach is, in general, non-optimal and does not provide optimality guarantees.

3.3 Minimum Edge Addition for Structural Controllability

In this section, we provide the main results of the paper. First, in Section 3.3.1, we reformulate Problem 2 as a graph-theoretical problem. Next, in Section 3.3.2, we sharpen our intuition by exploring two particular network topologies. In Section 3.3.3, we show that iterative solutions are sub-optimal. Next, using graph-theoretical tools, we characterize the set of feasible solutions to Problem 1 (Theorem 6). Subsequently, we obtain a feasible solution containing the minimum number of additional edges to ensure structural controllability (Theorem 7). Finally, we provide a polynomial-time algorithm (Algorithm 8) to obtain an optimal solution to Problem 2, whose correctness and computational complexity are proved in Theorem 8.

3.3.1 Graph-theoretical Optimization Problem

At a first glance, Problem 2 may seem a purely combinatorial problem. Naively, one may find a solution by exhaustively exploring the set of $n \times n$ binary matrices. However, Theorem 5 can be leveraged to shrink the search domain of (3.2). This motivates us to recast (3.2) as the following graph-theoretical problem.

Recall that the system digraph is given by $G(\bar{A}, \bar{B}) = (\mathcal{X} \cup \mathcal{U}, \mathcal{E}_{\mathcal{X},\mathcal{X}} \cup \mathcal{E}_{\mathcal{U},\mathcal{X}})$. Therefore, given a feasible edge-addition matrix $\tilde{A}$, we can associate a digraph with the perturbed structural system $(\bar{A} + \tilde{A}, \bar{B})$, which we denote by $G(\bar{A} + \tilde{A}, \bar{B}) = (\mathcal{X} \cup \mathcal{U}, \mathcal{E}_{\mathcal{X},\mathcal{X}} \cup \mathcal{E}_{\mathcal{U},\mathcal{X}} \cup \tilde{\mathcal{E}})$, where the edge set $\tilde{\mathcal{E}} \subseteq \mathcal{X} \times \mathcal{X}$ is such that $(x_i, x_j) \in \tilde{\mathcal{E}}$ if and only if $\tilde{A}_{ji} = 1$. Subsequently, since there is an one-to-one correspondence between $\tilde{\mathcal{E}}$ and the structural matrix $\tilde{A}$, we can provide the following equivalent formulation of Problem 1:

Problem 3. *Given the system digraph $G(\bar{A}, \bar{B}) = (\mathcal{X} \cup \mathcal{U}, \mathcal{E}_{\mathcal{X},\mathcal{X}} \cup \mathcal{E}_{\mathcal{U},\mathcal{X}})$, find*

$$\begin{aligned} \tilde{\mathcal{E}}^* = \arg \min_{\tilde{\mathcal{E}} \subseteq \mathcal{X} \times \mathcal{X}} \quad & |\tilde{\mathcal{E}}| \\ \textit{s.t.} \quad & G(\bar{A} + \tilde{A}, \bar{B}) = (\mathcal{X} \cup \mathcal{U}, \mathcal{E}_{\mathcal{X},\mathcal{X}} \cup \mathcal{E}_{\mathcal{U},\mathcal{X}} \cup \tilde{\mathcal{E}}) \\ & \textit{is structurally controllable.} \end{aligned}$$

Additionally, we define a *feasible edge-addition configuration* as a set of edges that is a feasible solution of Problem 3. Also, an *optimal edge-addition configuration* is defined as an optimal solution of Problem 3.

3.3.2 Special Cases

Next, before showing that iterative strategies can be suboptimal, we discuss two special cases to sharpen our intuition. First, recall that according to Theorem 5, the pair $(\bar{A}, \bar{B})$ is structurally controllable, if and only if, two conditions are satisfied. Therefore, we explore two special cases, where in each case only one of the conditions in Theorem 5 is satisfied; hence, only the remaining condition needs to be ensured to attain feasibility.

<u>*Case I*</u>: Consider a structured system $(\bar{A}, \bar{B})$ such that only Condition (a) in Theorem 1 holds, while Condition (b) is not satisfied. In other words, all state vertices are reachable while there exists a maximum matching of the system bipartite graph with right-unmatched vertices. As a result, the cardinality of a maximum matching M with respect to $\mathcal{B}(\bar{A}, \bar{B})$ is strictly less than n. Subsequently, let us denote by $U_L = \{v_i^l \colon i \in \{1, \ldots, n_l\}\}$ and $U_R = \{v_i^r \colon i \in \{1, \ldots, n_r\}\}$ the left- and right-unmatched vertices associated with a maximum matching M, respectively. In particular, notice that $n_l \geq n_r$ since $|\mathcal{X}^+ \cup \mathcal{U}^+| \geq |\mathcal{X}^-|$, and $|M| = n - n_r$. Therefore, to ensure that $G(\bar{A} + \tilde{A}, \bar{B})$ is structurally controllable, it is sufficient to add edges between U_L and U_R without common end-points and such that all right-unmatched vertices belong to one of such edges. However, such approach is not necessarily a solution to Problem 3 since some of the newly considered edges may correspond to edges between input and state vertices, while we are only allowed to connect pairs of state vertices. Consequently, let (without loss of generality) $U_L^{\mathcal{X}} = \{v_i^l \colon i \in \{1, \ldots, n_r\}\} \subseteq U_L$ be the set of n_r left-unmatched state vertices. Therefore, an optimal edge-addition configuration can be obtained as $\mathcal{E}^* = \{(v_i^l, v_i^r) \colon v_i^l \in U_L^{\mathcal{X}},\, v_i^r \in U_R,\, i \in \{1, \ldots, n_r\}\}$. In other words, $M \cup \mathcal{E}^*$ is a maximum matching with respect to the bipartite graph $\mathcal{B}(\bar{A} + \tilde{A}, \bar{B})$ without right-unmatched vertices, which implies that Theorem 5-(b) holds. Thus, $\tilde{\mathcal{E}}^* = \{s^{-1}(\{v_i^l, v_i^r\})) \colon \{v_i^l, v_i^r\} \in \mathcal{E}^*\}$ is an optimal solution to Problem 3. Since there may exist multiple maximum matchings of the system bipartite graph, the optimal edge-addition configuration constructed using the above procedure may not be unique. However, the number of right-unmatched vertices are the same for all maximum matchings due to maximality. As a result, in this case, all optimal edge-addition configurations contain n_r edges. ○

Remark 4. *Under the assumption that all state vertices in the system digraph are reachable, Problem 1 can be also solved via an integer program, as proposed in [56].* ◇

<u>*Case II*</u>: Suppose that a network $(\bar{A}, \bar{B})$ is such that Condition (b) in Theorem 1 holds,

while Condition (a) does not, i.e., some state vertex might be unreachable in $G(\bar{A}, \bar{B})$. Since at least one state vertex is assumed to be actuated (i.e., $\bar{B} \neq 0$), the set of reachable state vertices is non-empty. Therefore, we propose to partition the state vertices of the system digraph into two disjoint sets according to their reachability. Let $\mathcal{R}_1$ and $\mathcal{N}$ be the sets containing all the reachable and unreachable state vertices, respectively. Then, we define G_r (respectively, G_u) as the $\mathcal{R}_1$-induced (respectively, $\mathcal{N}$-induced) subgraph.

Now, notice that if an edge is added to ensure the reachability of any vertex v in some source SCC in G_u (i.e., the tail of the edge is a reachable state vertex), then all state vertices reachable from this particular source SCC become reachable as well. Consequently, to ensure reachability of all state vertices, it is sufficient to add edges to ensure reachability of one vertex per each unreachable source SCCs. Additionally, it is also necessary to have an edge pointing towards each source SCC in G_u, since otherwise the vertices belonging to it remain unreachable. Therefore, we first need to identify the source SCCs in the DAG associated with the unreachable subgraph G_u (these source SCCs can be efficiently found using, for example, [154]). Also, without loss of generality, assume there are r of these source SCCs, whose vertex sets are denoted by $\mathcal{S}_j \subseteq \mathcal{N}$, $j = 1, \ldots, r$. Subsequently, to ensure the reachability of all state vertices in $\mathcal{N}$, we need to add r edges which tails are in a reachable vertex and each head points towards one of the vertices in one of the r source SCCs. Thus, the set $\tilde{\mathcal{E}}^* = \{(v_r, v_j) \colon v_r \in \mathcal{R}_1, v_j \in \mathcal{S}_j, j \in \{1, \ldots, r\}\}$ is an optimal edge-addition configuration. Notwithstanding, notice that $\tilde{\mathcal{E}}^*$ does not characterize all possible optimal edge-addition configurations, since when an edge is added from a reachable vertex towards an unreachable source SCC, all state vertices reachable from this particular source SCC become reachable; thus, the tail of an edge in an optimal edge-addition configuration should be in $\mathcal{R}_1$ and its head should be in an unreachable source SCC. ◦

From Case I, we notice that selecting new edges for the edge-addition configuration do not increase the number of right-unmatched vertex associated with the system bipartite graph. Similarly, adding more edges never decreases the number of reachable state vertices in the system digraph. As a consequence, one may select edges to ensure both

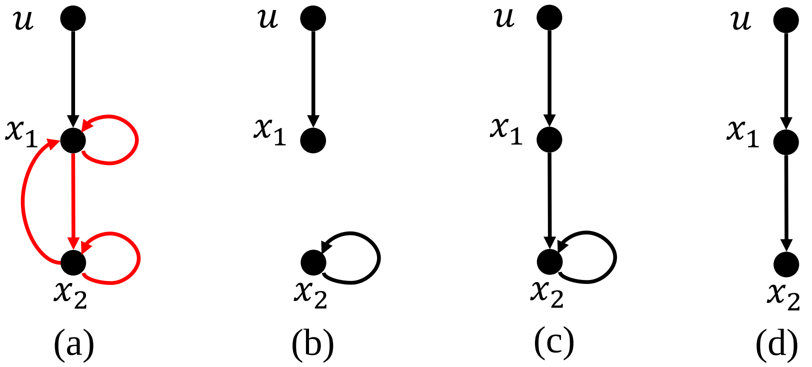


Figure 3-1: In (a), we illustrate a system digraph $G(\{u, x_1, x_2\}, \{(u, x_1)\})$ with three vertices and one edge depicted in black. The goal is to find the smallest subset of state edges (depicted by red edges) to ensure structural controllability. Let us consider the iterative strategy described in Subsection 3.3.3. In (b), we depict a possible solution to the first step described in Case I, i.e., the edge (x_2, x_2) suffices to satisfy Theorem 5-(b). In (c), we depict a possible solution to the second step described in Case II when the system digraph considered is the one depicted in (b). In contrast, the edge (x_1, x_2) suffices to satisfy Theorem 5-(a), resulting in the system digraph in (d).

conditions in Theorem 1 are satisfied iteratively. Nonetheless, such a selection scheme often leads to sub-optimal solutions, as we show next.

3.3.3 Iterative Solutions are Sub-optimal

In order to motivate the need for an algorithm that solves a general instance of the problem proposed in Problem 3, we describe below a naive iterative approach leading to suboptimal solutions. The steps in this iterative algorithm are based on the cases described in Section 3.3.2. Specifically, each iteration consists of a two-stage process. In the first stage, we find the minimum number of edges required to satisfy Theorem 5-(b) using the methodology described in Case I. The second stage in each iteration is described in Case II, whose aim is to satisfy Condition (a) in Theorem 5.

To show how this iterative approach can lead to suboptimal solutions, we show in Figure 3-1 an instance where we initially use the method proposed in Case I to ensure that Theorem 5-(b) holds, followed by the method proposed in Case II is applied to ensure Theorem 5-(a). As we explain in the caption of Figure 3-1, the naive strategy requires two edges, whereas the digraph depicted in Figure 3-1-(d) is also feasible and

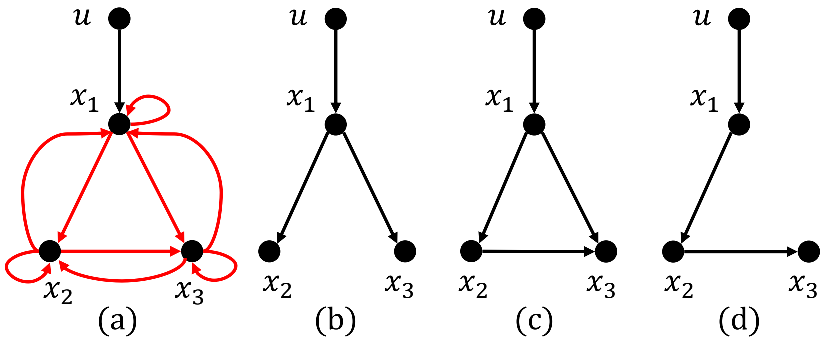


Figure 3-2: In (a), we illustrate a system digraph $G(\{u, x_1, x_2, x_3\}, \{(u, x_1)\})$ in black. The goal is to find the smallest subset of state edges (depicted by red) to ensure structural controllability. Let us consider the iterative strategy described in Subsection 3.3.3. In (b), we depict a possible solution to the first step described in Case II. In (c), we depict a possible solution to the second step, which was computed by performing the solution described in Case I when the system digraph considered is the one depicted in (b). In contrast, the edge (x_2, x_3) suffices to satisfy Theorem 5-(b), resulting in the system digraph in (d).

requires only one edge. Alternatively, in Figure 3-2, we provide an instance where the strategy adopted aims first to ensure Theorem 5-(a), followed by Theorem 5-(b), using the solutions in Case II and Case I, respectively. Again, in this case, the naive strategy requires three edges, whereas the digraph depicted in Figure 3-2-(d) is also feasible and requires only two edges. In summary, naive strategies are (in general) sub-optimal.

3.3.4 General Case

Hereafter, we characterize the solutions to Problem 3 when no assumptions are made on the topology of the network. First, we introduce a definition required to characterize the smallest collection of edges needed to attain reachability, i.e., satisfy Condition (a) in Theorem 5. In order to introduce this definition, we need to define the following notation. Let $G(\bar{A}, \bar{B}) = (\mathcal{X} \cup \mathcal{U}, \mathcal{E}_{\mathcal{X},\mathcal{X}} \cup \mathcal{E}_{\mathcal{U},\mathcal{X}})$ be the system digraph, and partition the set of state vertices $\mathcal{X}$ into two sets based on their reachability (from an input), namely, $\mathcal{X} = \mathcal{R}_1 \cup \mathcal{N}$, where $\mathcal{R}_1$ is the set of reachable vertices and $\mathcal{N}$ is the set of unreachable vertices. Additionally, without loss of generality, let us assume there are r source SCCs that are unreachable, which vertex sets are denoted by $\mathcal{N}_1, \ldots, \mathcal{N}_r \subseteq \mathcal{N}$. Also, let $\Delta(\mathcal{N}_h)$ denote the set of vertices that are reachable in $G(\bar{A}, \bar{B})$ from the vertices in $\mathcal{N}_h$, for $h = 1, \ldots, r$.

Definition 4. *A set S_B is called a set of bridging edges if it can be generated by the following recursive algorithm:*

```
Algorithm 1: Set of bridging edges
  Input: Sets $\mathcal{R}_1$ and $\mathcal{N}_1, \ldots, \mathcal{N}_r$;
  Initialize $\mathcal{K} = \{1, \ldots, r\}$, $t_1$ as any value in $\mathcal{K}$, and the set $S_B = \{(i, j)\}$, where $i$ is any vertex in $\mathcal{R}_1$ and $j$ is any vertex in $\mathcal{N}_{t_1}$;
  for $k = 2 : r$ do
      $\mathcal{R}_k \leftarrow \mathcal{R}_{k-1} \cup \Delta(\mathcal{N}_{t_{k-1}})$;
      Assign $t_k$ to any value in $\mathcal{K} \setminus \bigcup_{h=1}^{k-1}\{t_h\}$;
      $S_B \leftarrow S_B \cup \{(i, j)\}$ for any $i \in \mathcal{R}_k$ and any $j \in \mathcal{N}_{t_k}$;
  end for
```

Algorithm 1 is illustrated in Figure 3-3. In particular, notice that at the end of this algorithm $\mathcal{N} = \bigcup_{h=1}^{r} \Delta(\mathcal{N}_{t_k})$, which implies that all unreachable states become reachable. Furthermore, notice that the set of bridging edges contains the minimum number edges required to ensure that all state vertices are reachable. In fact, it readily follows that the solutions to Case II in Section 3.3.2 can be characterized by the possible sets of bridging edges. Furthermore, the set of bridging edges only ensure Condition (a) in Theorem 5, which is not sufficient to ensure structural controllability in general. More specifically, to ensure structural controllability and, subsequently, to obtain a feasible edge-addition configuration, two types of edges are required: (i) a set of bridging edges, and (ii) edges that connect left-unmatched state vertices to right-unmatched vertices in some maximum matching associated with the system bipartite graph (recall Case I in Section 3.3.2). In what follows, we state necessary and sufficient conditions to obtain a feasible edge-addition configuration:

Theorem 6. *Let $G(\bar{A}, \bar{B})$ be a system digraph and $\mathcal{B}(\bar{A}, \bar{B})$ be its bipartite representation. Furthermore, let M be a maximum matching associated with $\mathcal{B}(\bar{A}, \bar{B})$ and $U_L(M) = \{v_i^l : i \in \{1, \ldots, n_l\}\}$ and $U_R(M) = \{v_i^r : i \in \{1, \ldots, n_r\}\}$ be the left- and right-unmatched vertices of M. Without loss of generality, let $U_L^{\mathcal{X}}(M) = \{v_i^l : i \in \{1, \ldots, n_r\}\}$ denotes the set of n_r left-unmatched state vertices of M. A set $\tilde{\mathcal{E}}$ is a feasible edge-addition configuration if and only if it contains the union of the following two sets:*

(a) S_B is the set of bridging edges; and

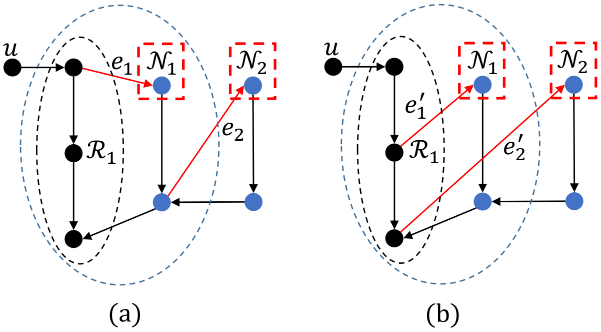


Figure 3-3: This figure provides an illustration of Algorithm 1. All vertices (blue or black), together with all black edges, form the initial system digraph $G(\bar{A}, \bar{B})$. The black vertices, except the input vertex u, constitute the set of reachable state vertices $\mathcal{R}_1$ (enclosed by the black dashed ellipsoid). Blue vertices constitute the set of unreachable state vertices $\mathcal{N}$. The unreachable state source SCCs, $\mathcal{N}_1$ and $\mathcal{N}_2$, are contained in red dashed squares. In Figure (a), we depict one possible result for Algorithm 1. In the initialization step, our algorithm initializes S_B as the set containing edge e_1 only. Subsequently, after e_1 is added to S_B, all the states reachable from $\mathcal{N}_1$ become reachable (we encircle these reachable states by a blue dashed ellipsoid in Figure (a)). Afterwards, in the FOR loop, edge e_2 in Figure (a) is added to S_B (in Step 5 of Algorithm 1), resulting in a digraph in which all vertices are reachable from the input node. An alternative output of Algorithm 1 is plotted in Figure (b). Notice that both in Figures (a) and (b), all vertices are reachable after adding two red edges. Therefore, $S_B = \{e_1, e_2\}$ and $S_B' = \{e_1', e_2'\}$ are two possible sets of bridging edges.

(b) $S_M = \{s^{-1}(\{v_i^l, v_i^r\})\colon v_i^l \in U_L^{\mathcal{X}}(M), v_i^r \in U_R(M),$ *and* $i = \{1, \ldots, n_r\}\}$, *for some maximum matching* M *associated with the system bipartite graph.*

◇

Proof. See Appendix A.2. □

From Theorem 6, we can readily obtain a lower-bound on the number of edges in a feasible edge-addition configuration.

Corollary 2. *The cardinality of an optimal edge-addition configuration* $\tilde{\mathcal{E}}^*$ *satisfies* $|\tilde{\mathcal{E}}^*| \geq \max\{n_r, r\}$*, where* n_r *is the number of right-unmatched vertices of any given maximum matching* M *associated with the system bipartite graph* $\mathcal{B}(\bar{A}, \bar{B})$*, and* r *is the number of unreachable state source SCCs in the DAG associated with the system digraph* $G(\bar{A}, \bar{B})$*.* ◇

Proof. See Appendix A.2. □

In particular, it is easy to verify that the equality in Corollary 2 is ensured when both special cases addressed in Section 3.3.2 are considered.

Although Theorem 6 characterizes feasible edge-addition configurations, we seek to find a feasible edge-addition configuration of minimum cardinality. To achieve this goal, we notice that it is preferable to obtain a maximum matching whose set of right-unmatched vertices are spread across different unreachable source SCCs. This is because the edges connecting left- to right-unmatched vertices in this particular maximum matching are useful to simultaneously satisfy both Conditions (a) and (b) in Theorem 6. To formalize this reasoning, we introduce the following concept.

Definition 5. *Let $G(\bar{A}, \bar{B})$ be the system digraph and M be a maximum matching associated with its bipartite representation $\mathcal{B}(\bar{A}, \bar{B})$. Furthermore, denote by $U_R(M)$ the set of right-unmatched vertices of M. An unreachable state source SCC of the DAG associated with the system digraph $G(\bar{A}, \bar{B})$ is said to be unreachable-assignable if it contains at least one right-unmatched vertex in $U_R(M)$.* ◇

Whether an unreachable state source SCC $\mathcal{S}$ is unreachable-assignable depends on the specific maximum matching M. In other words, given two sets $U_R(M_1)$ and $U_R(M_2)$ of right-unmatched vertices associated with two different maximum matchings M_1 and M_2, it is possible that $U_R(M_1)$ contains a vertex from $\mathcal{S}$ while $U_R(M_2)$ does not. We introduce the following definition to characterize the maximum number of possible unreachable-assignable state source SCCs.

Definition 6. *The* unreachable source assignability number *(USAN) of the system digraph $G(\bar{A}, \bar{B})$ is defined as the maximum number of unreachable-assignable state source SCCs among all the maximum matchings associated with the system bipartite graph $\mathcal{B}(\bar{A}, \bar{B})$.* ◇

Remark 5. *According to Definition 6, for every system digraph $G(\bar{A}, \bar{B})$, the USAN must be less or equal to the number of right-unmatched vertices associated with any maximum matching of the $\mathcal{B}(\bar{A}, \bar{B})$ and the total number of unreachable state source SCCs in $G(\bar{A}, \bar{B})$.* ◇

To find a maximum matching associated with the system bipartite graph that attains

```
Algorithm 2: Maximum matching attaining the USAN
  Input: A system digraph $G(\bar{A},\bar{B})$;
  Output: A maximum matching $M$ attaining the USAN;
    Partition the set of state vertices in the system digraph $G(\bar{A},\bar{B})$ based on their
      reachability. Obtain the set containing all the unreachable vertices of $G(\bar{A},\bar{B})$,
      denoted as $\mathcal{N}$, and its $\mathcal{N}$-induced subgraph, denoted as $G_u$.
    Obtain the source SCCs of $G_u$ and denote their vertex sets as $\mathcal{N}_1,\ldots,\mathcal{N}_r$,
      where $r$ is the total number of source SCCs in $G_u$;
    Define a vertex set $\mathcal{I}=\{\gamma_1,\ldots,\gamma_r\}$ comprising $r$ slack vertices. Construct a
      weighted bipartite graph $\mathcal{B}_w=\mathcal{B}(\mathcal{X}^+\cup\mathcal{U}^+\cup\mathcal{I},\mathcal{X}^-,\mathcal{E}_{\mathcal{X}^+,\mathcal{X}^-}\cup\mathcal{E}_{\mathcal{U}^+,\mathcal{X}^-}\cup\mathcal{E}_{\mathcal{I}})$,
      where $\mathcal{E}_{\mathcal{I}}=\bigcup_{i=1}^{r}\{\{\gamma_i,x_j^-\}\colon x_j\in\mathcal{N}_i\}$. The weights in $\mathcal{B}_w$ are as follows: every
      edge in $\mathcal{E}_{\mathcal{X}^+,\mathcal{X}^-}\cup\mathcal{E}_{\mathcal{U}^+,\mathcal{X}^-}$ is assigned to have unit weight, whereas every edge
      in $\mathcal{E}_{\mathcal{I}}$ has weight two;
    Let $M'$ be the minimum-weighted maximum matching of $\mathcal{B}_w$;
    Return $M=M'\setminus\mathcal{E}_{\mathcal{I}}$.
```

the USAN, one can naively enumerate all possible maximum matchings associated with $\mathcal{B}(\bar{A},\bar{B})$, but this approach incurs into a problem that is computationally $\sharp P$-complete[1] [158]. Instead of using an exhaustive search, it is possible to determine in *polynomial-time* a maximum matching attaining the USAN using the following algorithm.

Remark 6. *The proof of correctness of the algorithm described above is very similar to the proof of Theorem 11 in Section VI of [49].* $\diamond$

Essentially, in order to find a maximum matching attaining the USAN, we associate a slack vertex γ_i with each unreachable source SCC $\mathcal{N}_i$. We create additional edges from each slack vertex to every state vertex of its corresponding SCC. In other words, we let $\mathcal{E}_{\mathcal{I}}=\bigcup_{i=1}^{r}\{\{\gamma_i,x_j^-\}\colon x_j\in\mathcal{N}_i\}$. Next, we set the weights of edges $\mathcal{E}_{\mathcal{I}}$ higher than the weights of edges in $\mathcal{B}(\bar{A},\bar{B})$. With this particular selection of weights, the minimum-weighted maximum matching M' prefers selecting edges in $\mathcal{B}(\bar{A},\bar{B})$ to edges in $\mathcal{E}_{\mathcal{I}}$. In particular, edges are selected from $\mathcal{E}_{\mathcal{I}}$ if it helps to increase the matching. As a consequence, the vertices that are matched using edges in $\mathcal{E}_{\mathcal{I}}$ must correspond to right-unmatched vertices in the matching $M'\backslash\mathcal{E}_{\mathcal{I}}$. Furthermore, these right-unmatched vertices are spread across different unreachable source SCCs. Finally, due to maximality of matching, we can ensure that M achieves the USAN. To further illustrate the algorithm,

[1] The class of $\sharp P$-complete problems is a class of computationally equivalent counting problems that are at least as difficult as the *NP-complete* problems.

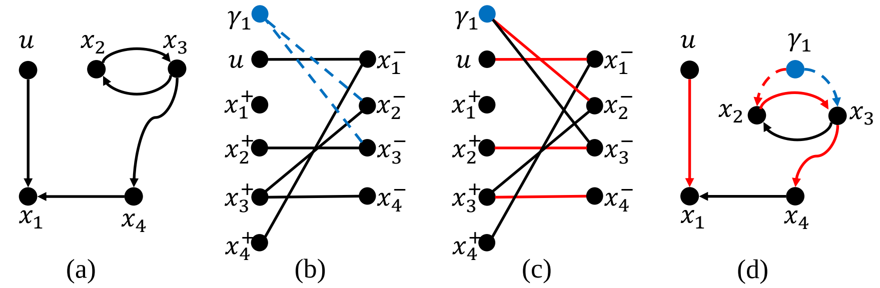


Figure 3-4: This figure presents an example illustrating Algorithm 2. The black vertices and edges in (a) form the initial system digraph $G(\bar{A}, \bar{B})$. In this case, $\mathcal{N} = \{x_2, x_3, x_4\}$ is the set of unreachable state vertices. Moreover, there is only one unreachable source SCC, whose vertex set is $\mathcal{N}_1 = \{x_2, x_3\}$. The black vertices and edges in (b) constitute the original system bipartite graph $\mathcal{B}(\bar{A}, \bar{B})$, while the blue vertex γ_1 represents a slack variable associated with $\mathcal{N}_1$. In addition, the blue dashed edges $\{\gamma_1, x_2\}$ and $\{\gamma_1, x_3\}$ together constitute $\mathcal{E}_{\mathcal{I}}$. The minimum-weighted maximum matching M' of $\mathcal{B}_w$ is depicted using red edges in (c). By removing $\{\gamma_1, x_2^-\} \in \mathcal{E}_{\mathcal{I}}$, we have that $M = \{\{u, x_1^-\}, \{x_2^+, x_3^-\}, \{x_3^+, x_4^-\}\}$ is a maximum matching of $\mathcal{B}(\bar{A}, \bar{B})$. In (d), we depict in red the edges from the system digraph $G(\bar{A}, \bar{B})$ associated with those in the maximum matching M. Notice that x_2 is a right-unmatched vertex of M and it is in $\mathcal{N}_1$; hence, M is a maximum matching attaining the USAN of $G(\bar{A}, \bar{B})$.

we present an example in Figure 3-4.

Remark 7. *Due to maximality, the USAN is unique for every system digraph $G(\bar{A}, \bar{B})$. Nonetheless, there may exist multiple maximum matchings that attains this value. Algorithm 2 obtains one particular solution.* ◇

Although the maximum matching that achieves the USAN can be efficiently obtained as described in Algorithm 2, this is not sufficient to obtain an optimal feasible edge-addition configuration. To illustrate this claim, let us consider the example depicted in Figure 3-5. In this case, the optimal feasible edge-addition configuration depends on the maximum matching achieving the USAN. Specifically, if all the left-unmatched vertices are unreachable state vertices, then, after fulfilling Condition (b) in Theorem 6, we should add extra edges to form a set of bridging edges to ensure Condition (a) in Theorem 6. This would result in a sub-optimal solution.

Since $\|B\|_0 \neq 0$, one can find a path rooted at an input vertex $u \in \mathcal{U}$ whose end vertex is some state vertex $x \in \mathcal{X}$. Thus, x^- is a left-unmatched vertex in the maximum matching containing the path. Consequently, it is always possible to obtain a maximum matching associated with $\mathcal{B}(\bar{A}, \bar{B})$ with at least one reachable left-unmatched state vertex – see

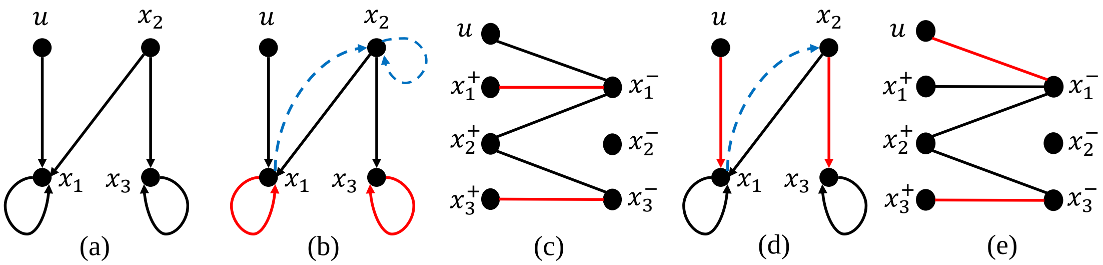


Figure 3-5: This figure presents two examples where different maximum matchings lead to sets of feasible edge-addition configurations with different cardinalities. The black vertices and edges in (a) form the initial system digraph $G(\bar{A}, \bar{B})$. The red edges in (c) and (e) constitute two different maximum matchings associated with $\mathcal{B}(\bar{A}, \bar{B})$. The red edges in (b) and (d) are direct graph representations of the edges determined by the maximum matchings in (b) and (d), respectively. The edge-set $\tilde{\mathcal{E}}_2 = \{(x_1, x_2)\}$ (depicted by blue dashed arrows in (d)) is a feasible edge-addition configuration, since the addition of (x_1, x_2) ensures both conditions in Theorem 5. In contrast, in Fig. (b) we also need to add edge (x_2, x_2) (in addition to (x_1, x_2)) to ensure that Theorem 6-(b) holds, which leads to a feasible edge-addition configuration given by $\tilde{\mathcal{E}}_1 = \{(x_1, x_2), (x_2, x_2)\}$. Thus, $\tilde{\mathcal{E}}_2$ is an optimal edge-addition configuration with cardinality 1 while $\tilde{\mathcal{E}}_1$ is not.

Proof of Theorem 7 in Appendix A.2 for more details. Moreover, when an edge is added from the reachable left-unmatched vertex to a right-unmatched state vertex in an unreachable source SCC, the set of reachable state vertices can be extended. We will use this fact to circumvent the sub-optimality issue mentioned above. In our next result, we characterize the relationship between the USAN and the optimal value to Problem 3:

Theorem 7. *Given the system digraph $G(\bar{A}, \bar{B})$ and its bipartite representation $\mathcal{B}(\bar{A}, \bar{B})$, if $\|\bar{B}\|_0 > 0$, then the cardinality of an optimal edge-addition configuration $p^* = |\tilde{\mathcal{E}}^*|$ satisfies*

$$p^* = n_r + r - q, \tag{3.3}$$

where n_r is the number of right-unmatched vertices in any maximum matching associated with $\mathcal{B}(\bar{A}, \bar{B})$, r is the number of unreachable source state SCCs in the DAG associated with $G(\bar{A}, \bar{B})$, and q is the USAN. ◇

Proof. See Appendix A.2. □

In fact, based on the constructive proof of Theorem 7 in Appendix A.2, we propose a procedure (described in Algorithm 8) to find an optimal edge-addition configuration in polynomial-time. Briefly, Algorithm 8 consists of the following four main steps: (*Step 1*) Decompose the system digraph based on the reachability of state vertices. (*Step 2*) Determine a maximum matching that achieves the USAN; if the obtained maximum matching admits no reachable left-unmatched vertex, then we alter the matching by finding a path rooted at certain input vertex. (*Step 3*) Based on the obtained maximum matching, in order to ensure both conditions in Theorem 6, select the edges from reachable left-unmatched vertices to right-unmatched vertices in unreachable source SCCs iteratively. (*Step 4*) If the system is still not structurally controllable, then add the smallest collection of edges ensuring that both conditions in Theorem 6 hold independently. The correctness and computational complexity of this procedure are described in the following result.

Theorem 8. *Given the system digraph $G(\bar{A},\bar{B}) = (\mathcal{X}\cup\mathcal{U}, \mathcal{E}_{\mathcal{X},\mathcal{X}}\cup\mathcal{E}_{\mathcal{U},\mathcal{X}})$, Algorithm 8 provides an optimal solution to Problem 3. Furthermore, the computational complexity of Algorithm 8 is $\mathcal{O}(|\mathcal{X}\cup\mathcal{U}|^3)$.* ◇

Proof. See Appendix A.2. □

Remark 8. *The computational complexity incurred by Algorithm 8 is comparable to that incurred by the algorithms required to solve the special cases described in Section 3.3.2. Specifically, the solution to Case I can be determined through the computation of a maximum matching, whose computational complexity is given by $\mathcal{O}(\sqrt{|\mathcal{X}\cup\mathcal{U}|}|\mathcal{E}_{\mathcal{X}^+,\mathcal{X}^-}\cup\mathcal{E}_{\mathcal{U}^+,\mathcal{X}^-}|)$ [154]. Alternatively, the solution to Case II can be obtained by determining the strongly connected components of the system digraph, which can be obtained by running a depth-first search algorithm twice [154] and incurring in $\mathcal{O}(|\mathcal{X}\cup\mathcal{U}|^2)$ computational complexity. A MATLAB implementation of Algorithm 8 can be found in [159].* ◇

3.4 Simulations

In this section, we illustrate the use of the main results of this paper. In particular, given a structurally uncontrollable system, we determine the minimum number of additional

Algorithm 3: Computing an optimal edge-addition configuration $\tilde{\mathcal{E}}^*$ to Problem 3

```
Input: The system digraph $G(\bar{A}, \bar{B})$;
Output: An optimal edge-addition configuration $\tilde{\mathcal{E}}^*$;
Step 1: System digraph decomposition
Obtain the set of all reachable (resp. unreachable) state vertices $\mathcal{R}_1$ (resp. $\mathcal{N}$ in $G(\bar{A}, \bar{B})$.
Step 2: Maximum matching attaining the USAN
Obtain a maximum matching $\bar{M}$ associated with $\mathcal{B}(\bar{A}, \bar{B})$ attaining the USAN $q$ using Algorithm 2;
if $U_L^{\mathcal{X}}(\bar{M}) \cap \mathcal{R}_1 = \emptyset$ then
    Find $v$ such that $v \in \mathcal{R}_1$ and $(u, v) \in \mathcal{E}_{\mathcal{U},\mathcal{X}} \setminus \bar{M}$;
    Find $\hat{v}$ such that $\{\hat{v}^+, v^-\} \in \bar{M}$;
    $M \leftarrow \big(\bar{M} \setminus \{\{\hat{v}^+, v^-\}\}\big) \cup \{\{u^+, v^-\}\}$;
else
    Set $M$ equal to $\bar{M}$;
end if
Step 3: Add edges to satisfy $(a)$ and $(b)$ in Theorem 6
Obtain the unique set of disjoint paths $\mathcal{P} = \bigcup_{i=1}^{q} \mathcal{P}_i$ in the matching $M$, where the starting vertex of each $\mathcal{P}_i$ is in some unreachable source SCC and the end vertex is a left-unmatched state vertex;
% We remark that the uniqueness of $\mathcal{P}$ is a direct consequence of $M$ being a matching.
Construct two sets of vertices $\mathcal{S} = \{s_1, \ldots, s_q\}$ and $\mathcal{T} = \{t_1, \ldots, t_q\}$ such that $s_i$ and $t_i$ are the starting and ending vertices of each path $\mathcal{P}_i$, respectively;
Let $\tilde{\mathcal{E}}^* \leftarrow \emptyset$ and $k \leftarrow 1$;
if $\mathcal{T} \cap \mathcal{R}_1 = \emptyset$ then
    Select a $t_0$ such that $t_0^+ \in U_L^{\mathcal{X}}(M)$ and $t_0 \in \mathcal{R}_1$;
    for $k \leq q$ do
        $\tilde{\mathcal{E}}^* \leftarrow \tilde{\mathcal{E}}^* \cup \{(t_{k-1}, s_k)\}$; $k \leftarrow k + 1$;
    end for
    $U_L^{\mathcal{X}}(M) \leftarrow U_L^{\mathcal{X}}(M) \setminus \{t_0^+, \ldots, t_{q-1}^+\}$;
else
    Find and apply a permutation of the $i$ indexes associated to the paths $\mathcal{P}_i$ s.t. $t_1 \in \mathcal{R}_1$ (accordingly, permute the elements in $\mathcal{S}$ and $\mathcal{T}$);
    for $k < q$ do
        $\tilde{\mathcal{E}}^* \leftarrow \tilde{\mathcal{E}}^* \cup \{(t_k, s_{k+1})\}$; $k \leftarrow k + 1$;
    end for
    $\tilde{\mathcal{E}}^* \leftarrow \tilde{\mathcal{E}}^* \cup \{(t_q, s_1)\}$; $U_L^{\mathcal{X}}(M) \leftarrow U_L^{\mathcal{X}}(M) \setminus \mathcal{T}$;
end if
$U_R(M) \leftarrow U_R(M) \setminus \mathcal{S}$;
```

```
Step 4: Add extra edges to satisfy Theorem 6
for $v_l^+ \in U_L^{\mathcal{X}}(M)$ do                         % to satisfy Theorem 6-(b)
  if $U_R(M) \neq \emptyset$ then
    $\tilde{\mathcal{E}}^* \leftarrow \tilde{\mathcal{E}}^* \cup \{(v_l, v_r)\}$, for some $v_r^- \in U_R(M)$;
    $U_L^{\mathcal{X}}(M) \leftarrow U_L^{\mathcal{X}}(M) \setminus v_l^+$; $U_R(M) \leftarrow U_R(M) \setminus v_r^-$;
  end if
end for
Construct a graph $G_{aug} = (\mathcal{X} \cup \mathcal{U}, \mathcal{E}_{\mathcal{X},\mathcal{X}} \cup \mathcal{E}_{\mathcal{U},\mathcal{X}} \cup \tilde{\mathcal{E}}^*)$. Let $C_i, i = 1, \ldots, \beta$,
be the vertex-sets of $\beta$ unreachable source SCCs in the DAG of $G_{aug}$.
Additionally, let $\mathcal{R}_{aug}$ be the set of all reachable vertices in $G_{aug}$;
for $i = 1 : \beta$ do                                        % to satisfy Theorem 6-(a)
  $\tilde{\mathcal{E}}^* \leftarrow \tilde{\mathcal{E}}^* \cup \{(v_i, z_i)\}$, for some $v_i \in \mathcal{R}_{aug}$, $z_i \in C_i$.
end for
```

edges required for ensuring structural controllability in a some artificial network models. First, in Section 3.4.1, we provide a pedagogical example capturing the outcome of the different steps of Algorithm 1. In Section 3.4.2, we evaluate the minimum number of edges required in the context of large-scale randomly generated networks.

3.4.1 Illustrative Example

Consider the pair $(\bar{A}, \bar{B})$, whose system digraph is depicted in Figure 3-6. Notice that the system is not structurally controllable since both conditions in Theorem 5 fail to hold. Therefore, additional edges are required to ensure structural controllability. Towards this goal, we invoke Algorithm 8 to obtain an optimal edge-addition configuration that solves Problem 2 given $(\bar{A}, \bar{B})$. In this algorithm, we need to decompose the system digraph $G(\bar{A}, \bar{B})$ according to the reachability of its state vertices. In particular, the set of reachable state vertices is given by $\mathcal{R}_1 = \{x_1, \ldots, x_4\}$, while the set of unreachable state vertices is $\mathcal{N} = \{x_5, \ldots, x_{10}\}$. Subsequently, we find the unreachable source SCCs, whose vertex sets are denoted by $\mathcal{N}_1, \mathcal{N}_2$, and $\mathcal{N}_3$ in Figure 3-6; hence, the set of states in unreachable source SCCs is $\{x_5, x_7, x_8, x_{10}\}$. Step 2 of Algorithm 8 computes a maximum matching $\bar{M}$ using Algorithm 2. In Figure 3-7-(a), we present in red such maximum matching, whose set of left-unmatched state vertices and right-unmatched vertices are $U_L^{\mathcal{X}}(\bar{M}) = \{x_2, x_9\}$ and $U_R(\bar{M}) = \{x_5, x_{10}\}$, respectively. Notice that x_5 and x_{10} belong to two different unreachable source SCCs; hence, the unreachable source assignability

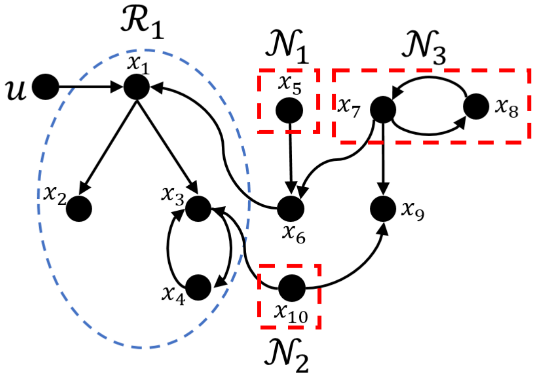


Figure 3-6: System digraph $G(\bar{A}, \bar{B})$ containing a single input vertex u and ten state vertices $\{x_1, \ldots, x_{10}\}$ (depicted in black dots). Black arrows correspond to the edges of $G(\bar{A}, \bar{B})$. The dashed blue ellipsoid contains all the reachable state vertices, i.e., $\mathcal{R}_1 = \{x_1, \ldots, x_4\}$, whereas each red dashed square contains an unreachable source SCC, whose vertex sets are $\mathcal{N}_1 = \{x_5\}$, $\mathcal{N}_2 = \{x_{10}\}$, and $\mathcal{N}_3 = \{x_7, x_8\}$, respectively.

number (USAN) equals two, i.e., $q = 2$. As a result, by invoking Theorem 7, it follows that an optimal edge-addition configuration consists of $p^* = 3$ edges.

Now, notice that x_2 is a reachable left-unmatched vertex, i.e., $x_2 \in U_L^{\mathcal{X}}(\bar{M}) \cap \mathcal{R}_1$. Thus, Step 2 of Algorithm 8 sets M equal to $\bar{M}$. To obtain an optimal edge-addition configuration $\tilde{\mathcal{E}}^*$, we should add an edge with tail in x_2 and head in some right-unmatched unreachable state vertex. According to M, we obtain $\mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_2$, where $\mathcal{P}_1 = \{x_5, x_6, x_1, x_2\}$ and $\mathcal{P}_2 = \{x_{10}, x_9\}$. From $\mathcal{P}$, the set $\mathcal{S} = \{s_1 = x_5, s_2 = x_{10}\}$ and $\mathcal{T} = \{t_1 = x_2, t_2 = x_9\}$ are constructed accordingly. As a result, Step 3 in Algorithm 8 adds the edge (x_2, x_{10}) to the edge-addition configuration $\tilde{\mathcal{E}}^*$. By selecting this edge, all vertices reachable from x_{10} become reachable. Subsequently, the algorithm adds (x_9, x_5) to $\tilde{\mathcal{E}}^*$, after which Condition (b) in Theorem 6 is satisfied, since $M \cup \tilde{\mathcal{E}}^*_{\mathcal{B}}$ is a maximum matching of $G(\bar{A} + \tilde{A}, \bar{B})$ without right-unmatched vertices, where $\tilde{\mathcal{E}}^*_{\mathcal{B}} = \{(x_2^-, x_{10}^+), (x_9^-, x_5^+)\}$ represents the bipartite representation of the edges in $\tilde{\mathcal{E}}^*$ in $G(\bar{A} + \tilde{A}, \bar{B})$.

Finally, it remains to ensure that every state vertex is reachable, i.e., that Condition (a) in Theorem 6 is satisfied by $G(\bar{A} + \tilde{A}, \bar{B})$. Towards this end, notice that the only remaining unreachable state source SCC is given by $\mathcal{N}_3 = \{x_7, x_8\}$. Consequently, it

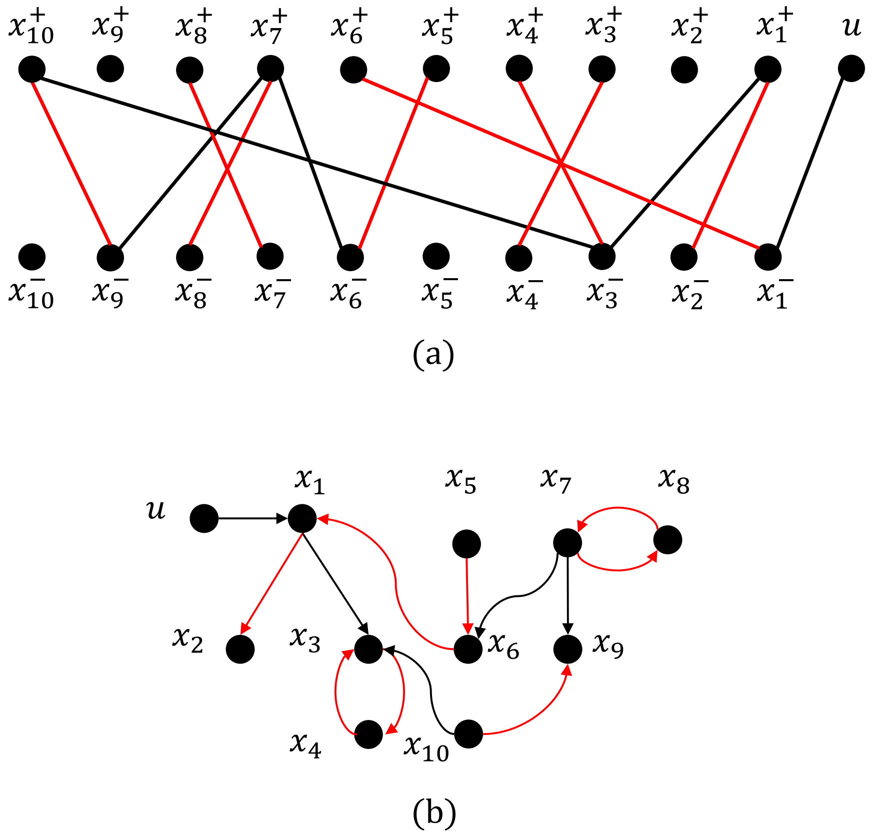


Figure 3-7: This figure shows a maximum matching $\bar{M}$ obtained using Step 2 in Algorithm 8. In (a), we depict the system bipartite graph associated with the pair $(\bar{A}, \bar{B})$, whose edges are depicted in black and red (edges in red are those in the maximum matching $\bar{M}$). In (b), we depict in red the edges from the system digraph $G(\bar{A}, \bar{B})$ associated with those in the maximum matching $\bar{M}$.

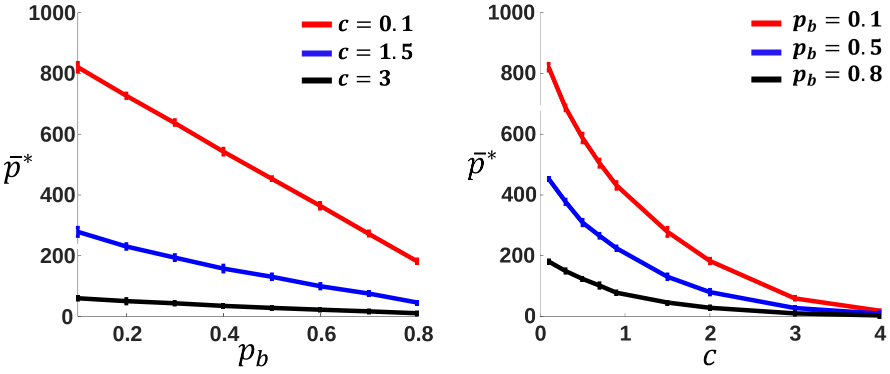
suffices to add (x_1, x_7) into $\tilde{\mathcal{E}}^*$ to ensure their reachability. However, there are multiple choices of edges to ensure the reachability of $\mathcal{N}_3$. More specifically, instead of adding (x_1, x_7) into $\tilde{\mathcal{E}}^*$, one can add any edge (x_i, x_j) with $i \in \{1, \ldots, 6, 10\}$ and $j \in \{7, 8\}$ as an alternative. In summary, an optimal edge-addition configuration, i.e., a solution to Problem 3, is given by $\tilde{\mathcal{E}}^* = \{(x_2, x_{10}), (x_9, x_5), (x_1, x_7)\}$, which contains $p^* = 3$ edges, as prescribed by Theorem 7.

3.4.2 Random Networks

In this section, we explore the minimum number of edges p^* contained in an optimal edge-addition configuration $\tilde{\mathcal{E}}^*$ required to ensure structural controllability of random networks. We assume that the structure of $\bar{A}$ is generated using an Erdős-Renyi model, i.e., $[\bar{A}]_{ij} = 1$ with probability $0 < p_a < 1$ for all i, j; 0 otherwise. In our simulations, the size of $\bar{A}$ is assumed to be $n = 1000$. We let $c \in \{0.1, 0.3, 0.5, 0.7, 0.9, 1.5, 2, 3, 4\}$ and define $p_a = \frac{c}{n}$ for every c accordingly. Thus, c represents the average sum of in-degree and out-degree of each vertex in the graph represented by A. Moreover, we assume $\bar{B}$ to be a random diagonal matrix with $p_b n$ entries equal to 1, and 0 otherwise, where $p_b \in (0, 1)$ represents the fraction of vertices to be set equal to 1. With this particular setup, we examine the value of p^* as we vary c and p_b, independently.

In Figure 3-8, we plot the empirical average of p^* (over 10 random realizations). Notice that p^* decreases as c or p_b increase. Intuitively, a larger value of c results in a denser state digraph. Thus, both conditions in Theorem 5 are more likely to be satisfied. In other words, the number of right-unmatched vertices associated with the maximum matching of the system bipartite graph and the number of unreachable state vertices are smaller as c increases. Furthermore, when p_b becomes close to one, almost every state vertex is actuated by an individual input. Thus, (a) in Theorem 5 holds with high probability. Since $p^* = n_r + r - q$, it follows that p^* decreases as c or p_b increase.

To emphasize the effect of varying p_b (respectively, c) on the minimum number of additional edges to ensure structural controllability, we plot in Figure 3-8-(a) (respectively, Figure 3-8-(b)) the evolution of p^* when c is fixed (respectively, p_b is fixed). In Figure 3-8-(a), we observe that for a reasonably small value of c (e.g., $c = 3$), the impact of p_b in the size of the optimal edge-addition configuration is almost negligible. Intuitively, as c increases towards $\log(n)$, the number of isolated vertices in the random subgraph induced by state vertices decreases. In particular, if $c \approx \log(n)$, then the state digraph presents a unique giant strongly connected component [160]. Subsequently, p^* is small even when there is only one state being actuated by an input. Indeed, in our experiment, $\bar{p}^* = 1.1$ when $c = 7$ and $p_b = 0.001$. In Figure 3-8-(b), we observe an almost exponential decrease of p^* with respect to c.



(a) (b)

Figure 3-8: In this figure, we plot the evolution of the average value of p^* as c and p_b vary. In (a), we fix the value of c and show the evolution of p^* versus p_b, when p_b ranges from 0.1 to 0.8 with step size 0.1. The red, blue, and black lines correspond to $c = 0.1$, $c = 1.5$, and $c = 3$, respectively. In (b), we plot the evolution of $\bar{p}^*$ when c varies in the interval $c \in \{0.1, 0.3, 0.5, 0.7, 0.9, 1.5, 2, 3, 4\}$, while fixing p_b. The red, blue, and black lines show the value of $\bar{p}^*$ when $p_b = 0.1$, $p_b = 0.5$, and $p_b = 0.8$, respectively. In both figures, the error bars represent the standard deviation of p^*.

Chapter 4

Topology Design in Symmetric Linear Structural Systems

In the previous chapter, we have designed an efficient algorithm to add a set of edges with minimum cardinality in the system digraph to render a structurally controllable system. Our results is build on graph-theoretical necessary and sufficient conditions for (asymmetric) structural controllability of a structural pair $(\bar{A}, \bar{B})$—see Theorem 5 for more details. Noticing the similarity between this condition and the graph-theoretical condition in the case when the state matrix is captured by an undirected graph (Theorem 1), we conjecture that it is possible to leverage the framework developed by Theorem 7 to add minimum number of *undirected* edges to render a symmetrically structured system. To solve this problem, we will proceed as follows. First, we first provide a rigorous statement of the minimum-cost edge selection problem under consideration in Section 4.1. In Section 4.2, we provide thorough analysis of the computation complexity of the minimum-cost edge selection problem and identify a few instances that are solvable in polynomial-time. Finally, we present illustrative examples for our algorithms in Section 4.3.

4.1 Problem Statement

Before introducing our problem of interest, let us recall some definitions in structural systems theory – interested readers are referred to Subsection 2.1.2 for more details. Since structural controllability problems are defined using the sparsity pattern of the system, it is natural to consider graph representations. Given a symmetrically structured matrix $\bar{A}$, we associate it with a directed graph $G(\bar{A}) = (\mathcal{X}, \mathcal{E}(\bar{A}))$, which we refer to as the *state digraph*, where $\mathcal{X} = \{x_i\}_{i=1}^n$ is the set of state vertices, and $\mathcal{E}(\bar{A}) = \{(x_j, x_i) : [\bar{A}]_{ij} = \star\}$ is the set of directed edges[1]. To capture the symmetrical parameter dependencies, it is also useful to associate with $\bar{A}$ an undirected graph $\mathcal{G}(\bar{A}) = (\mathcal{X}, \mathcal{E}_u(\bar{A}))$, where $\mathcal{E}_u(\bar{A}) = \{\{x_i, x_j\} \colon [\bar{A}]_{ij} = \star, i \leq j\}$ is the set of undirected edges. Similarly, we associate with the structural pair $(\bar{A}, \bar{B})$ a directed graph $G(\bar{A}, \bar{B}) = (\mathcal{X} \cup \mathcal{U}, \mathcal{E}(\bar{A}) \cup \mathcal{E}_{\mathcal{U},\mathcal{X}})$, which we refer to the *system digraph*, where $\mathcal{U} = \{u_i\}_{i=1}^m$ is the set of input vertices and $\mathcal{E}_{\mathcal{U},\mathcal{X}} = \{(u_j, x_i) \colon [\bar{B}]_{ij} = \star\}$ is the set of edges from input vertices to state vertices. Due to the symmetry of A, we also associate with $(\bar{A}, \bar{B})$ a mixed graph, referred to as the *system mixed graph*, $\mathcal{G}(\bar{A}, \bar{B}) = \{\mathcal{X} \cup \mathcal{U}, \mathcal{E}_u(\bar{A}), \mathcal{E}_{\mathcal{U},\mathcal{X}}\}$ containing undirected edges between state vertices and directed edges from input vertices to state vertices. Given a target set $\mathcal{T} \subseteq [n]$, we say a state vertex $x_i \in \mathcal{X}$ is a *target vertex* if $i \in \mathcal{T}$, and let $\mathcal{X}_{\mathcal{T}} \subseteq \mathcal{X}$ denote the set of target vertices.

In this chapter, we consider a few design problems aiming to render a structurally target controllable system:

Problem 4 (Minimum-Cost Edge Selection for Structural Target Controllability)**.** *Consider a structural pair $(\bar{A}, \bar{B})$, where $\bar{A} = \{0, \star\}^{n \times n}$ is symmetrically structured and $[\bar{A}]_{ij} = \star$ for all $i \leq j$. Let $\mathcal{G}(\bar{A}, \bar{B}) = (\mathcal{X} \cup \mathcal{U}, \mathcal{E}_u(\bar{A}), \mathcal{E}_{\mathcal{U},\mathcal{X}})$ be the mixed graph representation of $(\bar{A}, \bar{B})$. Consider a target set $\mathcal{T} \subseteq [n]$, and a function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ that assigns a non-negative cost to each undirected edge e in $\mathcal{X} \times \mathcal{X}$. Find,*

$$\bar{A}^\star = \arg \min_{\hat{A} \in \{0,\star\}^{n \times n}} \sum_{e \in \mathcal{E}_u(\hat{A})} c(e),$$

[1] We denote directed edges and undirected edges using parentheses (x_i, x_j) and curly brackets $\{x_i, x_j\}$, respectively

such that $(\hat{A}, \bar{B})$ is structurally target controllable with respect to $\mathcal{T}$ and $\hat{A}$ is symmetrically structured.

Notice that structural controllability is a special case of structural target controllability when $\mathcal{T} = [n]$, thus all solutions to Problem 4 are applicable to design problems concerning structral controllablity.

We also consider the particular problem of finding the sparsest state matrix to ensure structural target controllability, as stated below:

Problem 5 (Sparsest State Matrix Design for Structural Target Controllability)**.** *Consider a structural pair $(\bar{A}, \bar{B})$, where $\bar{A} \in \{0, \star\}^{n\times n}$ is symmetrically structured. Let $\mathcal{G}(\bar{A}, \bar{B}) = (\mathcal{X} \cup \mathcal{U}, \mathcal{E}_u(\bar{A}), \mathcal{E}_{\mathcal{U},\mathcal{X}})$ be the mixed graph representation of $(\bar{A}, \bar{B})$. Consider a target set $\mathcal{T} \subseteq [n]$, and a cost function $c : \mathcal{X} \times \mathcal{X} \to \{1, \infty\}$, where*

$$c(e) = \begin{cases} 1, & \text{if } e \in \mathcal{E}_u(\bar{A}), \\ \infty, & \text{otherwise.} \end{cases} \tag{4.1}$$

Find,

$$\bar{A}^{\star} = \arg \min_{\hat{A}\in\{0,\star\}^{n\times n}} \sum_{e\in\mathcal{E}_u(\hat{A})} c(e),$$

such that $(\hat{A}, \bar{B})$ is structurally target controllable with respect to $\mathcal{T}$ and $\hat{A}$ is symmetrically structured.

In addition to these problems, when a structural pair $(\bar{A}.\bar{B})$ is not structurally controllable, one may consider the problem of adding a few edges in order to obtain a (target) controllable system, as stated below:

Problem 6 (Minimum Edge Addition for Structural Target Controllability)**.** *Consider a structural pair $(\bar{A}, \bar{B})$, where $\bar{A} \in \{0, \star\}^{n\times n}$ is symmetrically structured. Let $\mathcal{G}(\bar{A}, \bar{B}) = (\mathcal{X} \cup \mathcal{U}, \mathcal{E}_u(\bar{A}), \mathcal{E}_{\mathcal{U},\mathcal{X}})$ be the mixed graph representation of $(\bar{A}, \bar{B})$. Consider a target set $\mathcal{T} \subseteq [n]$, and a cost function $c : \mathcal{X} \times \mathcal{X} \to \{0, 1\}$,*

$$c(e) = \begin{cases} 0, & \text{if } e \in \mathcal{E}_u(\bar{A}), \\ 1, & \text{otherwise.} \end{cases} \tag{4.2}$$

Find,

$$\bar{A}^{\star} = \arg \min_{\hat{A}\in\{0,\star\}^{n\times n}} \sum_{e\in\mathcal{E}_u(\hat{A})} c(e)$$

such that $(\hat{A}, \bar{B})$ is structurally target controllable with respect to $\mathcal{T}$ and $\hat{A}$ is symmetrically structured.

Notice that both Problem 5 and Problem 6 are special cases of Problem 4, in which the cost functions are specially adapted to different scenarios. In order to solve these problems, in the next section, we introduce a few concepts that are crucial in deriving our results.

4.2 Minimum-cost Edge Selection for Structural Target Controllability

In this section, we leverage the graph-theoretical conditions provided by Theorem 2 (see Chapter 2 for more details) to solve several network design problems. More specifically, given a symmetrically structured pair, we aim to find a set of edges to render a structural target controllable system incurring in a minimum total cost. We present a thorough analysis on the computational complexity of this problem under various assumptions on the cost function and the topology of the system graph. We first show that Problem 4 is NP-hard in general (see Theorems 9 and 10). Nonetheless, we identify a few instances of the problem that are polynomial solvable (see Theorems 12 and 14). Moreover, we provide polynomial-time algorithms to obtain an optimal solution to each identified solvable case (see Algorithms 4 and 6).

4.2.1 NP-Hardness of the Minimum-cost Edge Selection Problem

We first show that the Problem 4 is, in general, NP-hard. A conventional approach to prove NP-hardness is to reduce a known NP-complete problem to an instance of the problem of interest. Following this general principle, we design our instance as follows. First, we consider a specific cost function: $c(e) \in \{1, \infty\}, \forall e \in \mathcal{X} \times \mathcal{X}$, i.e., some of the edges cannot be selected for the design. With this cost function, we aim to seek minimum number of undirected edges with a unit cost such that both conditions in Theorem 2

Cost functions and assumptions	Complexity
$c(e) \in \{1, \infty\}, \forall e \in \mathcal{X} \times \mathcal{X}$	NP-hard
$c(e) \in \{1, \infty\}, \forall e \in \mathcal{X} \times \mathcal{X}$ and Assumption 2[2]	Polynomial
$c(e) \in \{0, 1\}, \forall e \in \mathcal{X} \times \mathcal{X}$	Polynomial

Table 4.1: Special cases of Problem 4 and their computational complexity.

are satisfied. Furthermore, we assume that Condition-2) in Theorem 2 holds in our graph instance $G(\bar{A}, \bar{B})$; hence, we have $|\mathcal{N}(\mathcal{S})| \geq |\mathcal{S}|, \forall \mathcal{S} \subseteq \mathcal{X}_{\mathcal{T}} = \{x_i \in \mathcal{X} \colon i \in \mathcal{T}\}$ in $G(\bar{A}, \bar{B})$. Thus, in order to solve Problem 4 under these assumptions, it remains to ensure Condition-1) in Theorem 2. In order words, to solve Problem 4 with the above cost function and graph instance, we need to find the minimum number of undirected edges need to be added such that all the state vertices indexed by the target set are reachable. Next, we reduce the min-set-cover problem [154] to this instance of Problem 4.

Definition 7 (Min-set-cover Problem)**.** *Let $\mathcal{X} = \{x_i\}_{i=1}^{n}$ be a set of n elements. Let $\mathcal{S}_j \subseteq \mathcal{X}$ for all $j \in [m]$. A set cover of $\mathcal{X}$ is a set $\mathcal{I} \subseteq [m]$ such that $\bigcup_{j \in \mathcal{I}} \mathcal{S}_j \supseteq \mathcal{X}$. Assume that $\bigcup_{j=1}^{m} \mathcal{S}_j \supseteq \mathcal{X}$, find a set cover $\mathcal{I}$ such that $|\mathcal{I}|$ is minimized.*

The similarity between Min-set-cover problem and the constructed instance of the Problem 4 lies in the fact that both problems are related with reachability of vertices in graphs. Intuited by this idea, we have the following theorem:

Theorem 9. *The minimum-cost edge selection to achieve structural target controllability (Problem 4) is* NP-hard.

Proof. See Appendix A.3. □

Although Problem 4 is NP-hard in general, this does not imply that every instance, i.e., problems with specific cost functions and graph topologies, is NP-hard. In what follows, we identify a few cases of Problem 4 and show how the complexities of obtaining an optimal solution may differ (see Table 4.1 for a summary).

[2] Assumption 2 is defined in Section 4.2.2.

4.2.2 Solution to the Sparsest State Matrix Design Problem

As the first variant of the Problem 4, we consider the design of a sparse state matrix that renders a structurally target controllable system. In other words, given a structurally target controllable pair $(\bar{A}, \bar{B})$, Problem 5 seeks the sparsest symmetrically structured state matrix $\bar{A}^{\star}$ that preserves structural target controllability with respect to $\mathcal{T}$. In this case, we show that the problem under consideration is NP-hard:

Theorem 10. *The Sparsest State Matrix Design Problem (Problem 5) is NP-hard.*

Proof. See Appendix A.3. □

As stated in the above theorem, Problem 5 is computationally challenging regardless of the additional assumption. Thus, we consider relaxing the problem by imposing additional assumptions on the topology of the state graph, motivated by a class of brain network model [41], as follows:

Assumption 2. *The structural pair $(\bar{A}, \bar{B})$, where $\bar{A} \in \{0, \star\}^{n \times n}$ is symmetrically structured, satisfies the following conditions:*

$$\begin{aligned} &[\bar{A}]_{ii} = 0, \forall i \in [n], \\ &\bar{B} \in \{0, \star\}^{n \times 1}, ||\bar{B}||_0 = 1, \end{aligned} \tag{4.3}$$

Furthermore, $\mathcal{T} = [n]$.

Under the above assumption, $(\bar{A}, \bar{B})$ represents a single-input system with only one actuated state. The assumption $\mathcal{T} = [n]$ is necessary because, otherwise, Problem 5 is still NP-hard, as shown in the proof of Theorem 9. Since $\mathcal{T} = [n]$, it follows that structural target controllability is equivalent to structural controllability. As a result, we aim to find the sparsest state matrix to render a structurally controllable single-input system. Hereafter, we show that this problem is polynomially solvable. To show this, we exploit the special structure of the system digraph. More specifically, we show in Lemma 4 that structural controllability can be equivalently characterized by another set of graph-theoretic conditions. Using these refined conditions, we will recast the problem

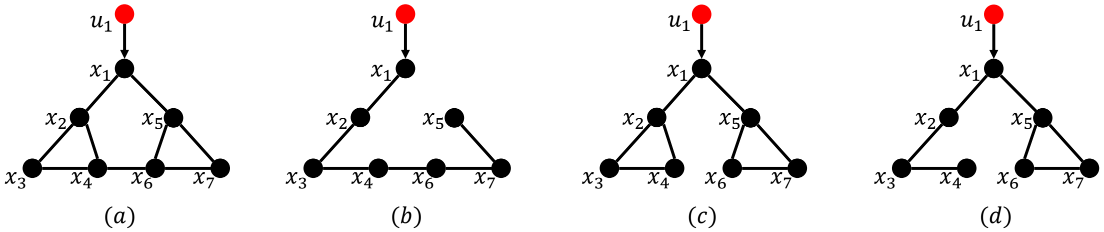


Figure 4-1: Illustrations of Theorem 11. Consider symmetrically structured pairs $(\bar{A}_1, \bar{B})$ and $(\bar{A}_2, \bar{B})$ whose mixed graph representations are depicted in (a) and (c), respectively. In each subfigure, the black and red vertices are state and input vertices, respectively. For the structural pair $(\bar{A}_1, \bar{B})$, the second condition in Theorem 11 is satisfied, and an optimal solution is depicted in (b). For the structural pair $(\bar{A}_2, \bar{B})$, the first condition in Theorem 11 is satisfied, and an optimal solution is shown in (d).

such that effective graph-theoretic algorithms can be applied.

Before stating Lemma 4, we introduce several relevant notions. Given a symmetrically structured matrix $\bar{A}$, a set of undirected edges $\mathcal{M} \subseteq \mathcal{E}_u(\bar{A})$ is said to be an *undirected matching* if no two edges in $\mathcal{M}$ has a common vertex and $\mathcal{M}$ is a *maximum undirected matching* if $\mathcal{M}$ has the largest cardinality among all feasible undirected matchings. Given an undirected matching $\mathcal{M}$, a vertex is called *unmatched* if it is not incident to any edge in $\mathcal{M}$. Since there is only one state vertex actuated by an input in $G(\bar{A}, \bar{B})$, without loss of generality, we assume that the state vertex x_1 is the only input-actuated vertex hereafter.

Lemma 4. *Consider a structural pair $(\bar{A}, \bar{B})$ satisfying Assumption 2. The pair $(\bar{A}, \bar{B})$ is structurally controllable, if and only if, the following two conditions hold simultaneously:*

1. *either $G_{\mathcal{X}}$ or $G_{\mathcal{X}\setminus\{x_1\}}$, induced subgraphs of $G(\bar{A})$, can be covered by vertex-disjoint directed cycles of length of at least 2;*

2. *$G_{\mathcal{X}}$ is strongly connected.*

Proof. See Appendix A.3. □

By Lemma 4, a feasible solution to Problem 4 must contain edges to ensure both conditions in Lemma 4 simultaneously. In the following theorem, we characterize these feasible solutions and identify a condition for optimality.

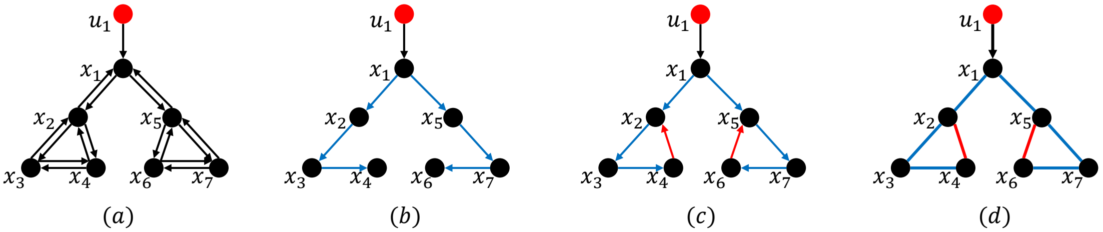


Figure 4-2: Non-optimality of applying the algorithm in [58] to Problem 5. Consider a structural pair $(\bar{A}, \bar{B})$ and associate it with a digraph $G(\bar{A}, \bar{B}) = (\mathcal{X} \cup \mathcal{U}, \mathcal{E}(\bar{A}) \cup \mathcal{E}_{\mathcal{U},\mathcal{X}})$, depicted in subfigure (a). In order to obtain a set of a minimum total number of directed edges in $G(\bar{A})$ to ensure structural controllability, we execute the algorithm in [58]. At the first step, the algorithm selects a set of edges to form a minimum spanning tree, depicted by the blue edges in (b); at the second step, it returns the directed edges to be added such that the Condition-2) in Theorem 1 is satisfied, depicted by the red edges in (c). We denote by the set $\mathcal{E}$ the solution returned by the algorithm in [58], and depict the mixed graph $\mathcal{G} = (\mathcal{X} \cup \mathcal{U}, \{\{x_i, x_j\} \colon (x_i, x_j) \text{ or } (x_j, x_i) \in \mathcal{E}\}, \mathcal{E}_{\mathcal{U},\mathcal{X}})$ in (d). Comparing with Figure 4-1, we see that $\mathcal{G}$ is not an optimal solution to Problem 5.

Theorem 11. *Consider a structural pair $(\bar{A}, \bar{B})$ satisfying Assumption 2. Let*

$$\bar{A}_a = \begin{bmatrix} \bar{A} & \bar{B} \\ \bar{B}^\top & 0 \end{bmatrix},$$

and define $\mathcal{G}(\bar{A}_a) = (\mathcal{X} \cup \{u_1\}, \mathcal{E}_u(\bar{A}) \cup \{\{u_1, x_1\}\})$. Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be maximum undirected matchings in $\mathcal{G}(\bar{A})$ and $\mathcal{G}(\bar{A}_a)$, respectively. Consider a set $\mathcal{E}_{u_1}$ (resp., $\mathcal{E}_{u_2}$) such that $G(\mathcal{X}, \{(x_i, x_j) \colon \{x_i, x_j\} \in \mathcal{E}_{u_1}\})$ (resp., $G(\mathcal{X} \setminus \{x_1\}, \{(x_i, x_j) \colon \{x_i, x_j\} \in \mathcal{E}_{u_2}\})$) can be covered by vertex-disjoint directed cycles and is strongly connected. Then,

1. *if $|\mathcal{M}_1| = |\mathcal{M}_2|$, then $|\mathcal{E}_{u_1}| = 2|\mathcal{X}| - 2|\mathcal{M}_1| - 1$ and $\mathcal{E}_{u_1}$ is an optimal solution to Problem 5;*

2. *if $|\mathcal{M}_1| \neq |\mathcal{M}_2|$, then $|\mathcal{E}_{u_2}| = 2|\mathcal{X}| - 2|\mathcal{M}_1| - 2$ and $\mathcal{E}_{u_2}$ is an optimal solution to Problem 5.*

Proof. See Appendix A.3. □

Based on Theorem 11, a naive solution to Problem 4 is to first find a set of undirected edges to ensure Condition-1) in Lemma 4 (each state vertex being covered by vertex-disjoint directed cycles) and then condense each derived vertex-disjoint cycles into a

Algorithm 4: Solution to Problem 5 under Assumption 2

```
Input: The system digraph $G(\bar{A},\bar{B}) = (\mathcal{X} \cup \mathcal{U}, \mathcal{E}(\bar{A}) \cup \mathcal{E}_{\mathcal{U},\mathcal{X}})$ of a structural pair
    $(\bar{A},\bar{B})$, where $\bar{A}$ is symmetrically structured, and the undirected state graph
    $\mathcal{G}(\bar{A}) = (\mathcal{X}, \mathcal{E}_u(\bar{A}))$;
Output: Optimal solution to Problem 4, $\mathcal{E}_u(\bar{A}^*)$;
    Step 1: Find undirected edges to ensure Lemma 4-1)
Define $\bar{A}_a$ and undirected graph $\mathcal{G}(\bar{A}_a)$ as in Theorem 11;
Find maximum undirected matchings in $\mathcal{G}(\bar{A})$ and $\mathcal{G}(\bar{A}_a)$, respectively. Denote
    them by $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively;
if $|\mathcal{M}_1| = |\mathcal{M}_2|$ then
    $\mathcal{E}_u \leftarrow \mathcal{M}_1$, $\mathcal{V} \leftarrow \mathcal{X}$, $\mathcal{E} \leftarrow \mathcal{E}(\bar{A})$;
else
    $\mathcal{E}_u \leftarrow \mathcal{M}_2$, $\mathcal{V} \leftarrow \mathcal{X} \setminus \{x_1\}$, $\mathcal{E} \leftarrow \mathcal{E}(\bar{A}) \setminus \{(x_j,x_1),(x_1,x_j) \colon [\bar{A}]_{1j} = \star\}$;
end if
Define a cost function $c'(e) \colon \mathcal{X} \times \mathcal{X} \to \{0,1,\infty\}$, as follows:

    $$c'(e) = \begin{cases} 0, & \text{if } e \in \{(x_i,x_j) \in \mathcal{E} \colon \{x_i,x_j\} \in \mathcal{E}_u\}, \\ 1, & \text{if } e \in \mathcal{E} \setminus \{(x_i,x_j) \in \mathcal{E} \colon \{x_i,x_j\} \in \mathcal{E}_u\}, \\ \infty, & \text{otherwise.} \end{cases}$$

Find a minimum weighted perfect matching $\mathcal{M}_b$ in the bipartite graph
    $\mathcal{B}(\mathcal{V},\mathcal{V},\mathcal{E})$ with weight $c'(e)$;
Let $\mathcal{E}'_u \leftarrow \{\{x_i,x_j\} \in \mathcal{E}_u(\bar{A}) \colon (x_i,x_j) \in \mathcal{M}_b\}$;
    Step 2: Find undirected edges to ensure Lemma 4-2)
Let $\{\mathcal{C}_i\}_{i=1}^{k}$ be the set of cycles in digraph $G(\mathcal{V},\mathcal{M}_b)$;
if $|\mathcal{M}_1| = |\mathcal{M}_2|$ then
    Let $\tilde{\mathcal{V}} \leftarrow \{v_i\}_{i=1}^{k}$;
    Let $\tilde{\mathcal{E}}_u \leftarrow \{\{v_i,v_j\} \colon \{(x_{i'},x_{j'}) \in \mathcal{E}(\bar{A}) \colon x_{i'} \in \mathcal{V}_{\mathcal{C}_i}, x_{j'} \in \mathcal{V}_{\mathcal{C}_j}\} \neq \emptyset\}$;
else
    $\tilde{\mathcal{V}} \leftarrow \{v_i\}_{i=1}^{k+1}$, $\tilde{\mathcal{E}}_u \leftarrow \{\{v_i,v_j\} \colon \{(x_{i'},x_{j'}) \in \mathcal{E}(\bar{A}) \colon x_{i'} \in \mathcal{V}_{\mathcal{C}_i}, x_{j'} \in \mathcal{V}_{\mathcal{C}_j}\} \neq$
    $\emptyset\} \cup \{\{v_j,v_{k+1}\} \colon \{x_{j'} \in \mathcal{V}_{\mathcal{C}_j} \colon [\bar{A}]_{1j'} = \star\} \neq \emptyset\}$;
end if
Find a minimum spanning tree $\mathcal{M}_t$ in $\mathcal{G}(\tilde{\mathcal{V}},\tilde{\mathcal{E}}_u)$;
Let $[\bar{A}^*]_{ij} = \star$, if $\{x_i,x_j\} \in \mathcal{E}'_u \cup \mathcal{M}_t$, $[\bar{A}^*]_{ij} = 0$ otherwise;
Return $\mathcal{E}_u(\bar{A}^*)$.
```

condensed node and recast paths among vertices in different vertex-disjoint cycles as edges among the corresponding condensed nodes. The problem of adding more undirected edges to ensure Condition-2) in Lemma 4 is, therefore, equivalent to finding a minimum spanning tree in the condensed graph, as illustrated by Algorithm 4. Indeed, using Theorem 11, we prove in the following theorem that such an iterative approach is optimal.

Theorem 12. *Under Assumption 2, Algorithm 4 returns an optimal solution to Problem 5 in $\mathcal{O}(|\mathcal{X}|^3)$.*

Proof. See Appendix A.3. □

Remark that this polynomial-time solution heavily relies on the fact that the system is symmetrically structured. Without this constraint, the problem is NP-hard, as shown in [58]. In addition, notice that the authors in [58] propose a 2-approximation algorithm to the problem of selecting a minimum total number of directed edges ensuring structural controllability. However, the algorithm cannot be applied to solve Problem 5 in our case, as we illustrate in Figure 2.

4.2.3 Solution to the Minimum Undirected Edge Addition Problem

We now proceed to address the Minimum-cost Edge Addition Problem, i.e., Problem 6. Since the necessary and sufficient conditions for structural target controllability share a similar form with the ones of structural controllability, we adopt a similar approach to the one proposed in Chapter 3 to solve this problem. More specifically, on the one hand, suppose that the topology of the system digraph ensures that Condition-1) in Theorem 2 holds, then it remains to add edges to ensure all target vertices are matched in the view of Remark 2. On the other hand, suppose that Condition-2) in Theorem 2 holds, then it suffices to add undirected edges to ensure that all target vertices are reachable (from the inputs). This is equivalent to ensuring all connected components containing target vertices that are reachable from inputs. As a result, in order to add a minimum number of undirected edges, it is beneficial to add undirected edges to connected components that contain right-unmatched target vertices. To formalize this argument, we define the following notion:

Definition 8 (Target-SCC)**.** *Let $G(\bar{A}, \bar{B})$ be the system digraph of a structural pair, and $\mathcal{T}$ be the target set. An SCC in $G(\bar{A}, \bar{B})$ is called a T-SCC if its vertex set contains at least one target vertex from $\mathcal{X}_{\mathcal{T}} = \{x_i \in \mathcal{X} \colon i \in \mathcal{T}\}$. We say a T-SCC is reachable if there exists a path from an input vertex $u_i \in \mathcal{U}$ to a vertex in the T-SCC, and unreachable otherwise.*

However, since we are adding *undirected* edges, we cannot simply apply Algorithm 8. In our case, adding an undirected edge $\{x_i, x_j\}$ can be viewed as adding a pair of directed edges (x_i, x_j) and (x_j, x_i) if $x_i \neq x_j$; or a self-loop (x_i, x_i), otherwise. Thus, it is possible that adding one undirected edge results in two right-unmatched vertices to become right-matched. This poses new challenges on Problem 6, since the addition of undirected edges requires a more careful analysis than Problem 3 (see Figures 4-3 and 4-4 for examples).

Since it is beneficial to add undirected edges to ensure reachability of unreachable T-SCC and reduce the number of right-unmatched target vertices, we characterize in the following lemma how many unmatched vertices can be reduced through the added edge.

Lemma 5. *Consider a structural pair $(\bar{A}, \bar{B})$, where $\bar{A}$ is symmetrically structured. Define $\mathcal{B}_1 = (\mathcal{X} \cup \mathcal{U}, \mathcal{X}, \mathcal{E}_{\mathcal{X}\cup\mathcal{U},\mathcal{X}})$, where $\mathcal{E}_{\mathcal{X}\cup\mathcal{U},\mathcal{X}} = \mathcal{E}(\bar{A}) \cup \mathcal{E}_{\mathcal{U},\mathcal{X}}$. Let $\mathcal{M}$ be a maximum matching in $\mathcal{B}_1$. Given an undirected edge $\{x_i, x_j\}$, let $\mathcal{B}_2 = (\mathcal{X} \cup \mathcal{U}, \mathcal{X}, \mathcal{E}_{\mathcal{X}\cup\mathcal{U},\mathcal{X}} \cup \{(x_i, x_j), (x_j, x_i)\})$. Let r_1 and r_2 be the number of right-unmatched vertices in $\mathcal{B}_1$ and $\mathcal{B}_2$, respectively. Then,*

1. *$r_2 \geq r_1 - 2$ for any $\{x_i, x_j\}$;*
2. *if both x_i and x_j are right-unmatched with respect to $\mathcal{M}$ in $\mathcal{B}_1$, then there exists a matching $\mathcal{M}'$ in $\mathcal{B}_2$ such that $|\mathcal{M}'| = |\mathcal{M}| + 2$ with x_i and x_j being right-matched.*

Proof. See Appendix A.3. □

Item-1) in Lemma 5 shows that adding an undirected edge can reduce at most two right-

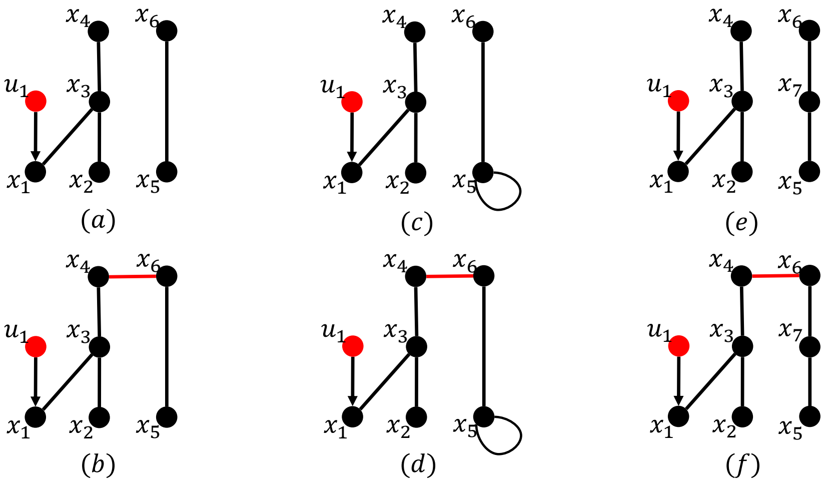


Figure 4-3: Illustrations on the Class-0, 1 and 2 of unreachable T-SCC. In each subfigure, the red and black vertices are input and state vertices, respectively. In (a), let $\mathcal{T} = \{i\}_{i=1}^{6}$. $G_{\{x_5,x_6\}}$ is a Class-0 T-SCC because the maximum number of right-unmatched target vertices that can be reduced by adding an edge making $G_{\{x_5,x_6\}}$ reachable is 0 (as shown by the red edge in (b)). In (c), let $\mathcal{T} = \{i\}_{i=1}^{6}$. $G_{\{x_5,x_6\}}$ is a Class-1 T-SCC because the maximum number of right-unmatched target vertices that can be reduced by adding an edge, making $G_{\{x_5,x_6\}}$ reachable, is 1 (as shown by the red edge in (d)). In (e), let $\mathcal{T} = \{i\}_{i=1}^{7}$. $G_{\{x_5,x_6,x_7\}}$ is a Class-2 T-SCC because we can reduce the total number of right-unmatched target vertices by 2 after adding an edge, making $G_{\{x_5,x_6,x_7\}}$ reachable (as shown by the red edge in (f)).

unmatched vertices in the overall system mixed graph. Meanwhile, it is also possible that a T-SCC cannot reduce any right-unmatched vertex in spite of any added edge, as depicted in Figure 4-3. Thus, we need to characterize how many right-unmatched vertices can be reduced by adding an undirected edge to a T-SCC:

Definition 9 (Class-η Unreachable T-SCC)**.** *Let $G_{\mathcal{S}} = (\mathcal{S}, (\mathcal{S} \times \mathcal{S}) \cap \mathcal{E}(\bar{A}))$ be an unreachable T-SCC in the system digraph $G(\bar{A}, \bar{B})$. Let $\mathcal{X}_{\mathcal{T}}$ be the set of target vertices in $\mathcal{S}$. Construct two bipartite graphs: (i) $\mathcal{B}(\mathcal{S}, \mathcal{X}_{\mathcal{T}}, \mathcal{E}_{\mathcal{S},\mathcal{X}_{\mathcal{T}}})$, and (ii) $\mathcal{B}'(\mathcal{S} \cup \{x_0\}, \mathcal{X}_{\mathcal{T}} \cup \{x_0\}, \mathcal{E}_{\mathcal{S},\mathcal{X}_{\mathcal{T}}} \cup \{(x_0, x_j), (x_j, x_0) \colon x_j \in \mathcal{S}\})$, where x_0 is an auxiliary vertex. Let r and r' be the right-unmatched vertices with respect to a maximum matching in $\mathcal{B}$ and $\mathcal{B}'$. Then an unreachable T-SCC is in Class-η if $r - r' = \eta$.*

Remark 9. *By Definition 9, an unreachable T-SCC is in Class-2 if it has at least one right-unmatched target vertex. If an unreachable T-SCC has no right-unmatched target vertex, then it is in either Class-0 or Class-1.*

According to Lemma 5, an unreachable T-SCC can only be in Class-2, 1 or 0. Meanwhile,

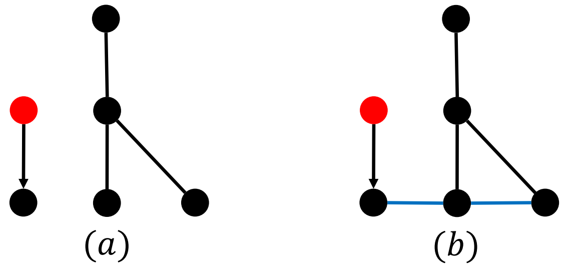


Figure 4-4: Illustrations of Theorem 13. In each subfigure, the red and black vertices are input and state vertices, respectively. We consider a structural pair $(\bar{A}, \bar{B})$, where $\bar{A} \in \{0, \star\}^{5\times 5}$ is symmetrically structured, and depict its mixed graph representation in subfigure (a). Let the target set be $\mathcal{T} = \{1, 2, \cdots, 5\}$. There are $q_2 = 1$ unreachable Class-2 T-SCCs, and there are 2 right-unmatched target vertices with respect to a maximum matching in the corresponding bipartite graph of $G(\bar{A}, \bar{B})$. However, there is no reachable right-unmatched target vertex, which implies $t = 1$ in the lower bound (4.4) of Theorem 13. Therefore, we need to add at least $\text{ceil}(1+\max(\frac{2-0-(2\times 1-1)}{2}, 0)) = 2$ edges to ensure structural target controllability. In subfigure (b), the blue undirected edges are the newly added edges which ensure structural target controllability, constituting an optimal solution to Problem 6.

we can design an algorithm to determine which class a T-SCC is in, using Definition 9, as follows. First, we create an auxiliary vertex and connect it with every target vertices within this T-SCC using undirected edges. Then, we construct the bipartite graph associated with the T-SCC, as well as that of the T-SCC together with the auxiliary vertex and edges. Finally, we compute the maximum matching in these two bipartite graphs and compare their number of right-unmatched vertices. By doing so, we can not only find out how many right-unmatched vertices can be reduced through an added edge, but we can also obtain a set of vertices in the T-SCC that enables the reduction, i.e., the vertices matching with the artificial vertex in the bipartite matching. We refer to this set as the *feature set*. The above discussion is formalized in Algorithm 5.

Although the right-unmatched vertices may be different with respect to different maximum matchings in a same bipartite graph, the number of them is invariant among different maximum matchings. As a result, it suffices to only consider a set of right-unmatched vertices with respect to a particular maximum matching when running Algorithm 5.

After defining the T-SCCs, we shift our focus to characterize the feasible solutions to the problem. Clearly, the total number of undirected edges needed to ensure structural

Algorithm 5: Classification of the i-th unreachable T-SCC into Class-0, 1 or 2.

```
Input: The $i$th unreachable T-SCC $G(\mathcal{X}_i, \mathcal{E}_i)$, target vertex set $\mathcal{X}_{\mathcal{T}i} \subseteq \mathcal{X}_i$;
Output: The classifier $\eta_i$ and a feature set $\mathcal{X}_i'$;
Find a Maximum Bipartite Matching $\mathcal{M}_b$ in $\mathcal{B}(\mathcal{X}_i, \mathcal{X}_{\mathcal{T}i}, \mathcal{E}_i)$. Let $\mathcal{X}_i''$ be the set of right-unmatched target vertices;
if $|\mathcal{X}_i''| \neq \emptyset$ then
    $\eta_i \leftarrow 2, \mathcal{X}_i' \leftarrow \mathcal{X}_i''$;
else
    Add a slack target vertex $x_0$;
    $\tilde{\mathcal{X}}_i \leftarrow \mathcal{X}_i \cup \{x_0\}, \mathcal{X}_{\mathcal{T}i} \leftarrow \mathcal{X}_{\mathcal{T}i} \cup \{x_0\}$;
    Find a Maximum Bipartite Matching $\tilde{\mathcal{M}}_b$ in
  $\mathcal{B}(\tilde{\mathcal{X}}_i, \mathcal{X}_{\mathcal{T}i}, \mathcal{E}_i \cup \{(x_0, x_j), (x_j, x_0) \colon x_j \in \mathcal{X}_i\})$.
    Let $\tilde{\mathcal{X}}_i''$ be the set of right-unmatched vertices in $\mathcal{X}_{\mathcal{T}i}$;
    if $|\tilde{\mathcal{X}}_i''| = \emptyset$ then
        $\eta_i \leftarrow 1, \mathcal{X}_i' \leftarrow \{x_j \colon (x_j, x_0) \text{ or } (x_0, x_j) \in \tilde{\mathcal{M}}_b\}$;
    else
        $\eta_i \leftarrow 0, \mathcal{X}_i' \leftarrow \mathcal{X}_i$;
    end if
end if
Return $\eta_i$ and $\mathcal{X}_i'$.
```

target controllability must be larger or equal to the total number of unreachable T-SCCs. Similarly, it is also no less than the number of undirected edges needed to make all the right-unmatched target vertices matched. As it is beneficial to add edges that serve both purposes, the optimal solution should leverage this fact and add edges between different T-SCCs. Using this idea, we compute a lower bound on the number of edges in any feasible solution.

Theorem 13. *Consider a structural pair $(\bar{A}, \bar{B})$, where $\bar{A}$ is symmetrically structured, and a target set $\mathcal{T}$. Let the cost function $c(e)$ be defined as in (4.2). Let $\mathcal{E}_u(\hat{A})$ be a feasible solution to Problem 6, then*

$$\sum_{e \in \mathcal{E}_u(\hat{A})} c(e) \geq \mathrm{ceil}(\ell + \max(\frac{r - q_1 - (2q_2 - t)}{2}, 0)), \tag{4.4}$$

where ℓ and r are the total number of unreachable T-SCCs, and the number of right-unmatched target vertices, respectively. q_1 and q_2 are the total number of unreachable Class-1 T-SCCs, and unreachable Class-2 T-SCCs, respectively. In particular, $t = 1$ if there is no reachable right-unmatched target vertex and there exists an unreachable Class-2 T-SCC, and $t = 0$ otherwise. The function $\mathrm{ceil} : \mathbb{R} \to \mathbb{N}$, *is defined as* $\mathrm{ceil}(q) =$

$\min\{p \in \mathbb{N} \colon p \geq q\}$.

Proof. See Appendix A.3. □

This theorem characterizes the theoretical minimum number of edges in any feasible solutions. In particular, we notice that the number is larger than both ℓ and $r/2$, which is consistent with our previous analysis. Meanwhile, the minus terms are involved due to consideration of edges that satisfy both conditions in Theorem 2. With the help of this theorem, it remains to construct a feasible solution whose cardinality equals the theoretical lower bound. To do this, we notice that different classes of T-SCC have different power in reducing the overall number of right-unmatched vertices through adding an undirected edge. Subsequently, a reasonable approach is to add edges 'greedily' to ensure reachability of Class-2 T-SCCs, followed by Class-1, and Class-0 T-SCCs. Based on this intuition, we propose Algorithm 6 to obtain a solution to Problem 6.

Essentially, Algorithm 6 follows four steps: (i) We first iterate over each Class-2 T-SCC following the order of decreasing total number of right-unmatched target vertices in it. For each Class-2 T-SCC, we add an undirected edge between a target vertex in it and a vertex outside such that the total number of unreachable T-SCCs is reduced by one and the total number of right-unmatched target vertices is reduced maximally. (ii) For each Class-1 T-SCC, we add an undirected edge between a vertex in it and a vertex outside such that the total number of unreachable T-SCC is reduced by one and the total number of right-unmatched target vertices is reduced maximally. (iii) For each Class-0 T-SCC, we add an undirected edge between a vertex in it and a reachable vertex such that this T-SCC becomes reachable. (iv) As such, we have made all the unreachable T-SCCs reachable. In this step, we add undirected edges such that all the remained right-unmatched target vertices are made right-matched. We state the complexity of Algorithm 6 in the following Theorem.

Theorem 14. *Algorithm 6 gives an optimal solution to Problem 6 in* $\mathcal{O}(|\mathcal{X}|^3)$.

Proof. See Appendix A.3. □

Algorithm 6: Solution to Problem 6

```
Input: The digraph $G(\bar{A}, \bar{B})$, a set of target vertices $\mathcal{X}_{\mathcal{T}}$, and undirected edge cost function $c(e) \colon \mathcal{X} \times \mathcal{X} \to \{0, 1\}$;
Output: Optimal solution $\mathcal{E}_u(\bar{A}^{\star})$;
Initialization:
$\mathcal{E}_{\mathcal{T}} \leftarrow ((\mathcal{U} \cup \mathcal{X}) \times \mathcal{X}_{\mathcal{T}}) \cap (\mathcal{E}(\bar{A}) \cup \mathcal{E}_{\mathcal{U},\mathcal{X}})$, $\mathcal{E}_u \leftarrow \emptyset, \check{\mathcal{E}}_u \leftarrow \emptyset$;
Let $\{\mathcal{X}_i\}_{i=1}^{q_2}, \{\mathcal{X}_i\}_{i=q_2+1}^{q_2+q_1}$, and $\{\mathcal{X}_i\}_{i=q_2+q_1+1}^{q_2+q_1+q_0}$ be the set of vertices in Class-2, 1, and 0 unreachable T-SCCs in $G(\bar{A}, \bar{B})$, respectively. Let $\ell = q_2 + q_1 + q_0$. Let $\mathcal{M}$ be a matching in $\mathcal{B}(\mathcal{X} \cup \mathcal{U}, \mathcal{X}_{\mathcal{T}}, \mathcal{E}_{\mathcal{T}})$. Let $\mathcal{S}$ (respectively, $\tilde{\mathcal{S}}$) be the set of reachable target vertices which are right-unmatched (respectively, matched) with respect to $\mathcal{M}$. Denote by $\mathcal{X}_i''$ the set of right-unmatched target vertices in $\mathcal{X}_i, \forall i \in [q_2]$. Rearrange the index of vertex sets in $\{\mathcal{X}_i\}_{i=1}^{q_2}$ such that $|\mathcal{X}_i''| \geq |\mathcal{X}_{i+1}''|, \forall i \in [(q_2 - 1)]$. Let $\mathcal{X}_i'$ be the feature set of $\mathcal{X}_i, \forall i \in [\ell]$;
Step 1: Iterate over Class-2 T-SCC
for each $\mathcal{X}_j \in \{\mathcal{X}_i\}_{i=1}^{q_2}$ do
  if $\mathcal{S} \neq \emptyset$ then
    Let $e = \{x_{j'}, x_k\}$, where $x_{j'} \in \mathcal{X}_j''$ and $x_k \in \mathcal{S}$;
    $\mathcal{E}_u \leftarrow \mathcal{E}_u \cup \{e\}, \mathcal{S} \leftarrow (\mathcal{X}_j'' \setminus \{x_{j'}\}) \cup (\mathcal{S} \setminus \{x_k\})$;
    $\tilde{\mathcal{S}} \leftarrow (\tilde{\mathcal{S}} \cup \mathcal{X}_j) \setminus \mathcal{S}$;
  else if $\mathcal{X}_j'' \neq \emptyset$ then
    if $j < q_2$ then
      Let $e = \{x_{j'}, x_k\}$, where $x_{j'} \in \mathcal{X}_j''$ and $x_k \in \mathcal{X}_{j+1}''$;
      $\mathcal{E}_u \leftarrow \mathcal{E}_u \cup \{e\}$;
      $\mathcal{X}_{j+1}'' \leftarrow (\mathcal{X}_j'' \setminus \{x_{j'}\}) \cup (\mathcal{X}_{j+1}'' \setminus \{x_k\})$;
      $\mathcal{X}_{j+1} \leftarrow \mathcal{X}_{j+1} \cup \mathcal{X}_j$;
    else
      Find a maximum matching $\tilde{\mathcal{M}}$ in $\mathcal{B}(\mathcal{U} \cup \mathcal{X}, \tilde{\mathcal{S}}, \mathcal{E}_{\mathcal{U},\mathcal{X}} \cup \{(x_i, x_j) \colon \{x_i, x_j\} \in \mathcal{E}_u(\bar{A}) \cup \mathcal{E}_u, x_j \in \tilde{\mathcal{S}}\})$;
      if $\hat{\mathcal{S}} = \{x_i \in \tilde{\mathcal{S}} \colon (u_{i'}, x_i) \in \tilde{\mathcal{M}}, u_{i'} \in \mathcal{U}\} \neq \emptyset$ then
        Let $e = \{x_{j'}, x_i\}$, where $x_{j'} \in \mathcal{X}_{q_2}''$ and $x_i \in \hat{\mathcal{S}}$;
      else
        Let $e = \{x_{j'}, x_{i'}\}$, where $x_{j'} \in \mathcal{X}_{q_2}''$ and $x_{i'} \in \{x_{i'} \colon (x_i, x_{i'}) \in \mathcal{M}\}$;
      end if
      $\mathcal{E}_u \leftarrow \mathcal{E}_u \cup \{e\}$, $\mathcal{S} \leftarrow \mathcal{X}_{q_2}'' \setminus \{x_{j'}\}$;
      $\tilde{\mathcal{S}} \leftarrow (\tilde{\mathcal{S}} \cup \mathcal{X}_{q_2}) \setminus \mathcal{S}$;
    end if
  else
    Let $e = \{x_{j'}, x_k\}$, where $x_{j'} \in \mathcal{X}_j$ and $x_k \in \tilde{\mathcal{S}}$;
    $\mathcal{E}_u \leftarrow \mathcal{E}_u \cup \{e\}, \tilde{\mathcal{S}} \leftarrow \tilde{\mathcal{S}} \cup \mathcal{X}_j$;
  end if
end for
```

```
Step 2: Iterate over Class-1 T-SCC
for each $\mathcal{X}_j \in \{\mathcal{X}_i\}_{i=q_2+1}^{q_2+q_1}$ do
    if $\mathcal{S} \neq \emptyset$ then
        Let $e = \{x_{j'}, x_k\}$, where $x_{j'} \in \mathcal{X}'_j$ and $x_k \in \mathcal{S}$;
        $\mathcal{E}_u \leftarrow \mathcal{E}_u \cup \{e\}, \mathcal{S} \leftarrow \mathcal{S} \setminus \{x_k\}$;
        $\tilde{\mathcal{S}} \leftarrow (\tilde{\mathcal{S}} \cup \mathcal{X}_i) \setminus \mathcal{S}$;
    else
        Let $e = \{x_{j'}, x_k\}$, where $x_{j'} \in \mathcal{X}_j$ and $x_k \in \tilde{\mathcal{S}}$;
        $\mathcal{E}_u \leftarrow \mathcal{E}_u \cup \{e\}, \tilde{\mathcal{S}} \leftarrow \tilde{\mathcal{S}} \cup \mathcal{X}_j$;
    end if
end for
Step 3: Iterate over Class-0 T-SCC
for each $\mathcal{X}_j \in \{\mathcal{X}_i\}_{i=q_2+q_1+1}^{q_2+q_1+q_0}$ do
    Let $e = \{x_{j'}, x_k\}$, where $x_{j'} \in \mathcal{X}_j$ and $x_k \in \tilde{\mathcal{S}}$;
    $\mathcal{E}_u \leftarrow \mathcal{E}_u \cup \{e\}, \tilde{\mathcal{S}} \leftarrow \tilde{\mathcal{S}} \cup \mathcal{X}_j$;
end for
Step 4: Matching remaining right-unmatched target vertices
if $\mathcal{S} \neq \emptyset$ then
    if $|\mathcal{S}| = 2k$, for some $k \in \mathbb{Z}$ then
        Partition $\mathcal{S}$ into subsets $\mathcal{R} = \{x_{\ell_i}\}_{i=1}^{k}$ and $\mathcal{Q} = \{x_{\gamma_i}\}_{i=1}^{k}$ s.t. $\mathcal{Q} \cup \mathcal{R} = \mathcal{S}$. Let $\mathcal{E}_u \leftarrow \mathcal{E}_u \cup \{x_{\ell_i}, x_{\gamma_i}\}_{i=1}^{k}$;
    else
        $k \leftarrow \{k \in \mathbb{Z}\colon |\mathcal{S}| = 2k+1\}$;
        Pick a $x_p \in \mathcal{S}$, and then partition $\mathcal{S} \setminus \{x_p\}$ into subsets $\mathcal{R} = \{x_{\ell_i}\}_{i=1}^{k}$ and $\mathcal{Q} = \{x_{\gamma_i}\}_{i=1}^{k}$ such that $\mathcal{R} \cup \mathcal{Q} = \mathcal{S} \setminus \{x_p\}$. Then, $\mathcal{E}_u \leftarrow \mathcal{E}_u \cup \{\{x_{\ell_i}, x_{\gamma_i}\}\}_{i=1}^{k} \cup \{\{x_p, x_p\}\}$;
    end if
end if
Let $[\bar{A}^\star]_{ij} = \star$ if $\{x_i, x_j\} \in \mathcal{E}_u(\bar{A}) \cup \mathcal{E}_u$ and $[\bar{A}^\star]_{ij} = 0$ otherwise;
Return $\mathcal{E}_u(\bar{A}^\star)$.
```

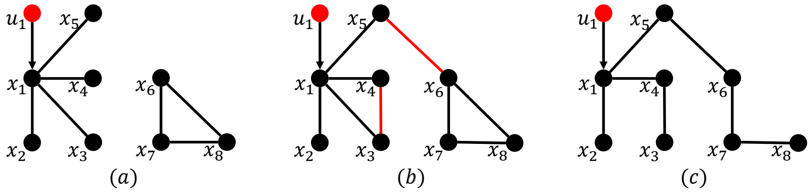


Figure 4-5: In each subfigure, the red vertex is the input vertex and black vertices are state vertices. The subfigure (a) is the mixed graph representation $\mathcal{G}(\bar{A}, \bar{B})$ of $(\bar{A}, \bar{B})$; The subfigure (b) is the mixed graph representation $\mathcal{G}(\bar{A}_1^*, \bar{B})$ of $(\bar{A}_1^*, \bar{B})$, where the red edges is the 'newly added' edges compared with (a); The subfigure (c) is the mixed graph representation $\mathcal{G}(\bar{A}_2^*, \bar{B})$ of $(\bar{A}_2^*, \bar{B})$.

4.3 Illustrative Examples

In this section, we consider a few examples to illustrate how to use Algorithms 4 and 6 to solve Problem 5 and 6, respectively. To illustrate these algorithms, we consider a structural pair $(\bar{A}, \bar{B})$, where $\bar{A} \in \{0, \star\}^{8\times 8}$ is symmetrically structured and $\bar{B} \in \{0, \star\}^{8\times 1}$. The mixed graph representation $\mathcal{G}(\bar{A}, \bar{B})$ is depicted Figure 4-5-(a). Let $\mathcal{X} = \{x_i\}_{i=1}^{8}$ and $\mathcal{U} = \{u_1\}$ be the sets of state and input vertices, respectively. Let the target set be $\mathcal{T} = \{1, 2, \cdots, 8\}$. Since there exists a set $\mathcal{S} = \{x_2, x_3, x_4, x_5\}$ such that $\mathcal{N}(\mathcal{S}) = \{x_1\}$, by Theorem 1, $(\bar{A}, \bar{B})$ is not structurally controllable.

We first consider the minimum cost edge addition problem (Problem 6), i.e., adding undirected edges to achieve structural target controllability with minimum total cost. We define the cost function $c_1(e), \forall e \in \mathcal{X} \times \mathcal{X}$, as follows.

$$c_1(e) = \begin{cases} 0, & \text{for } e \in \mathcal{E}_u(\bar{A}), \\ 1, & \text{otherwise.} \end{cases}$$

We use Algorithm 6 to solve this problem. Notice that in Figure 4-5-(a), $\mathcal{D}_{\mathcal{X}_1}$, where $\mathcal{X}_1 = \{x_6, x_7, x_8\}$, is a Class-1 unreachable T-SCC. With respect to a maximum matching $\mathcal{M}$ in $\mathcal{B}(\mathcal{X} \cup \mathcal{U}, \mathcal{X}, \mathcal{E}(\bar{A}) \cup \mathcal{E}_{\mathcal{U},\mathcal{X}})$, we have x_3, x_4, and x_5 are right-unmatched target vertices. By Algorithm 6, we first add an edge $e_1 = \{x_5, x_6\}$ into the system mixed graph such that, in $\mathcal{G}(\mathcal{X} \cup \mathcal{U}, \mathcal{E}_u(\bar{A}) \cup \{e_1\}, \mathcal{E}_{\mathcal{U},\mathcal{X}})$, $\forall x_i \in \mathcal{X}_1$ is input-reachable and total number of right-unmatched target vertices is reduced by 1 in the induced bipartite graph. Then we add $e_2 = \{x_2, x_3\}$ into the system mixed graph such that all the reachable right-unmatched vertices are matched and the total number of right-unmatched target vertices will be reduced by 2 in the induced bipartite graph. Let $\bar{A}_1^*$ be the symmetrically structured matrix returned by Algorithm 6. Through the proof of Theorem 14, we have $\mathcal{E}_u(\bar{A}_1^*)$ is the optimal solution of Problem 6.

Consequently, after applying Algorithm 6, the resulting structural system $(\bar{A}_1^*, \bar{B})$ is structurally controllable. Next, we consider the sparsest state matrix design problem (Problem 5) for $(\bar{A}_1^*, \bar{B})$. Let the new cost function be

$$c_2(e) = \begin{cases} 1, & \text{for } e \in \mathcal{E}_u(\bar{A}_1^*), \\ \infty, & \text{otherwise.} \end{cases}$$

We observe that $(\bar{A}_1^*, \bar{B})$ and target set $\mathcal{T}$ satisfy Assumption 2; hence, we can use Algorithm 4 to solve this problem. We can check that Condition-1) in Theorem 11 is satisfied. Let $\bar{A}_2^*$ be the symmetrically structured matrix returned by Algorithm 4. Through the proof of Theorem 11, we have $\mathcal{E}_u(\bar{A}_2^*)$ is the optimal solution to Problem 5.

Chapter 5

Structural Stabilizability and Network Resilience

While controllability is concerned about the ability of a system to steer all its states arbitrarily, in certain applications, it is not of major concern. Instead, stabilizability is a less restrictive system property since it only requires that the system states can be steered to the origin asymptotically by injecting proper control signals. Nonetheless, assessing whether a system is stabilizable requires exact parameters of the system.

In this chapter, we consider characterizing the stabilizability of a system from a topological perspective. Firstly, we derive a graph-theoretic necessary and sufficient condition for structural stabilizability of undirected networks. Secondly, we propose computationally efficient methods to determine the generic dimension of controllable subspace and the maximum stabilizable subspace of an undirected network system. Thirdly, we formulate the optimal actuator-disabling attack problem, where the attacker disables a limited number of actuators such that the maximum stabilizable subspace is minimized. Fourthly, we aim to solve the optimal recovery problem, where a defender activates a limited number of new actuators such that the dimension of the stabilizable subspace is maximized. We show the NP-hardness of this NP-hard, and we propose a $(1-1/e)$ approximation algorithm. Finally, we provide graph-theoretic characterizations for structural stabilizability in (arbitrary) linear structural systems.

5.1 Problem Formulations

We consider networks whose interconnection between states are captured by a symmetric linear time-invariant (LTI) system, described by

$$\dot{x} = Ax + Bu, \tag{5.1}$$

where $x \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$ are state vector and input vector, respectively. We refer to matrices $A = A^\top \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ as the state matrix and input matrix, respectively. Hereafter, we use the pair (A, B) to represent the system (5.1).

In order to infer the properties of a system modeled by (5.1) from its structure, and to ease presentation of this chapter, we recall and reintroduce some necessary concepts on structured matrices – also see Subsection 2.1.1 in Chapter 2.

Definition 10 (Structured and Symmetrically Structured Matrices)**.** *A matrix $\bar{M} \in \{0, \star\}^{n\times m}$ is called a* structured matrix*, if $[\bar{M}]_{ij}$, the (i, j)-th entry of $\bar{M}$, is either a fixed zero or an independent free parameter, denoted by $\star$. In particular, a matrix $\bar{M} \in \{0, \star\}^{n\times n}$ is* symmetrically structured*, if the value of the free parameter associated with $[\bar{M}]_{ji}$ is constrained to be the same as the value of the free parameter associated with $[\bar{M}]_{ij}$, for all i and j.*

Similar to the definition of structural controllability (Definition 1), we define *structural stabilizability* as follows:

Definition 11 (Structural Stabilizability)**.** *A structural pair $(\bar{A}, \bar{B})$ is said to be structurally stabilizable if there exists a stabilizable numerical realization $(\tilde{A}, \tilde{B})$.*

Remark 10. *Stabilizability is not a generic property [34], yet the structural stabilizability of $(\bar{A}, \bar{B})$ implies the existence of a numerical realization $(\tilde{A}, \tilde{B})$ such that $(\tilde{A}, \tilde{B})$ is stabilizable. In other words, it is a necessary condition for the stabilizability of any realization $(\tilde{A}, \tilde{B})$ of a structural pair $(\bar{A}, \bar{B})$.*

In the next two subsections, we will be focusing on two different main threads: (i) analysis, and (ii) design. We first formulate the problem of characterizing structural stabilizability using *only* the structural pattern of a pair, as stated below:

Problem 7. *Given a continuous-time linear time-invariant pair (A, B), we denote by $(\bar{A}, \bar{B})$ the structural pattern of (A, B), where $\bar{A} \in \{0, \star\}^{n \times n}$ is symmetrically structured. Find a necessary and sufficient condition such that $(\bar{A}, \bar{B})$ is structurally stabilizable.*

In addition to the above problem, we also consider how "unstabilizable" a system is, when a system is not stabilizable. To characterize the "unstabilizability", we propose using the dimension of the stabilizable subspace of a system, which can be stated as follows:

Definition 12 (Stabilizable Subspace [161])**.** *Given a pair (A, B), where $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times m}$, a set $S \subseteq \mathbb{R}^n$ is said to be the stabilizable subspace of (A, B) if for $\forall x(0) \in S$, there exists a control input $u(t) \in \mathbb{R}^m$, for $t \geq 0$, such that*

$$\lim_{t \to \infty} x(t) = \mathbf{0}.$$

As a special case, if a pair (A, B) is stabilizable, then $S = \mathbb{R}^n$. Moreover, we aim to determine the maximum dimension of stabilizable subspace, denoted by m-dim$(\bar{A}, \bar{B})$, among all numerical realizations of $(\bar{A}, \bar{B})$. Formally, we can state this problem as follows.

Problem 8. *Given a structural pair $(\bar{A}, \bar{B})$, where $\bar{A}$ is symmetrically structured, find* m-dim$(\bar{A}, \bar{B})$.

Upon these problems that concern mainly with the analysis of structural stabilizability, we can now focus on the design aspect of these problems in the following paragraphs. Since stabilizability plays a key role on network security – see, for example [64], in this thesis, we also consider network resilient problems. In this context, we assume that there exists an attacker who aims to minimize the maximum dimension of the stabilizable subspace by removing a certain amount of actuation capabilities, i.e., inputs. Formally, we consider the following version of the problem:

Problem 9 (Optimal Actuator-disabling Attack Problem)**.** *Consider a structural pair $(\bar{A}, \bar{B})$, where $\bar{A} \in \{0, \star\}^{n \times n}$ is symmetrically structured, and $\bar{B} \in \{0, \star\}^{n \times m}$ is a struc-*

tured matrix. Let the set Ω be $\Omega = [m]$, where $[m] := \{1, 2, \cdots, m\}$. Given a budget $k \in \mathbb{N}$, find

$$\begin{aligned} \mathcal{J}^* = \arg \min_{\mathcal{J} \subseteq \Omega} \text{m–dim}(\bar{A}, \bar{B}(\Omega \setminus \mathcal{J})) \\ \text{s.t. } |\mathcal{J}| \leq k, \end{aligned} \tag{5.2}$$

where $\bar{B}(\mathcal{I}) \in \{0, \star\}^{n \times |\mathcal{I}|}$ is a matrix formed by the columns of $\bar{B}$ indexed by $\mathcal{I}$, for some $\mathcal{I} \subseteq \Omega$.

In other words, the Problem 9 concerns about finding an optimal strategy to attack the stabilizability of a network using a fixed budget. Meanwhile, it is also of interest to consider the perspective of a system's designer (or, defender) that is concerned with the resilience of the network, i.e., how to maximize the dimension of stabilizable subspace by adding actuation capabilities (i.e., inputs) to the system:

Problem 10 (Optimal Recovery Problem)**.** *Consider a structural pair $(\bar{A}, \bar{B})$, where $\bar{A} \in \{0, \star\}^{n \times n}$ is symmetrically structured and $\bar{B} \in \{0, \star\}^{n \times m}$ is structured. Let $\mathcal{U}_{can}$, where $|\mathcal{U}_{can}| = m'$, be the set of candidate inputs that can be added to the system, and let $\bar{B}_{\mathcal{U}_{can}} \in \{0, \star\}^{n \times m'}$ be the structured matrix characterizing the interconnection between new inputs and the states in the system. Given a budget $k \in \mathbb{N}$, find*

$$\begin{aligned} \mathcal{J}^* = \arg \max_{\mathcal{J} \subseteq [m']} \text{m-dim}(\bar{A}, [\bar{B}, \bar{B}_{\mathcal{U}_{can}}(\mathcal{J})]) \\ \text{s.t. } |\mathcal{J}| \leq k, \end{aligned} \tag{5.3}$$

where $\bar{B}_{\mathcal{U}_{can}}(\mathcal{J}) \in \{0, \star\}^{n \times |\mathcal{J}|}$ is a structured matrix formed by the columns in $\bar{B}_{\mathcal{U}_{can}}$ indexed by $\mathcal{J}$, and $[\bar{B}, \bar{B}_{\mathcal{U}_{can}}(\mathcal{J})]$ is the concatenation of $\bar{B}$ and $\bar{B}_{\mathcal{U}_{can}}(\mathcal{J})$.

By the duality between stabilizability and detectability [162], all the results obtained on stabilizability in this chapter can be readily used to characterize detectability. All relevant notions and preliminaries needed to present solutions to Problems 7–10 have been introduced in Subsection 2.1.1, Subsection 2.1.2, and Subsection 2.2, respectively.

5.2 Analysis of Structural Stabilizability

In what follows, we have two subsections where we address Problems 7 and 8, respectively. In Section 5.2.1, we obtain a theorem (i.e., Theorem 15) that characterizes the solutions to Problem 7, whereas in Section 5.2.2, Theorem 16 gives a characterization of the maximum dimension of stabilizable subspace, which can be leveraged to provide solutions to Problem 8.

5.2.1 Graph-theoretic Conditions on Structural Stabilizability

Since stabilizability is a property that concerns about the stability of the uncontrollable part of (A, B), in order to obtain a graph-theoretic condition, we first characterize the controllable and uncontrollable parts from the structural information contained in the pair $(\bar{A}, \bar{B})$.

As characterized by Lemma 3 in Chapter 2, if a (symmetric) structural pair $(\bar{A}, \bar{B})$ is irreducible, then every non-zero mode of a numerical realization $(\tilde{A}, \tilde{B})$ is controllable. Furthermore, the above claim holds for almost all numerical realizations. In other words, the irreducibility of $(\bar{A}, \bar{B})$ guarantees that all the non-zero modes of $(\tilde{A}, \tilde{B})$ are controllable generically. Subsequently, we can claim that, from a contrapositive point of view, given an irreducible pair $(\bar{A}, \bar{B})$, if for any numerical realization $(\tilde{A}, \tilde{B})$ there exists an uncontrollable eigenvalue, then that uncontrollable eigenvalue is 0. This implies that $(\tilde{A}, \tilde{B})$ is not stabilizable. Therefore, if a pair $(\bar{A}, \bar{B})$ is irreducible but not structurally controllable, then $(\bar{A}, \bar{B})$ is not structurally stabilizable. Hence, we have the following lemma.

Lemma 6. *Given an irreducible structural pair $(\bar{A}, \bar{B})$, where $\bar{A} \in \{0, \star\}^{n \times n}$ is symmetrically structured, then $(\bar{A}, \bar{B})$ is structurally stabilizable if and only if $(\bar{A}, \bar{B})$ is structurally controllable.*

Proof. See Appendix A.4. □

While Lemma 6 provides us a condition for structural stabilizability when $(\bar{A}, \bar{B})$ is irreducible, we should also consider the case when $(\bar{A}, \bar{B})$ is reducible. By the definition of reducibility, $(\bar{A}, \bar{B})$ can be permuted to the form of (2.3). In order for $(\bar{A}, \bar{B})$ to be structurally stabilizable, it is required that there exists a numerical realization $\tilde{A}_{22}$

whose eigenvalues of are all negative. Summarizing these two arguments, it is equivalent to say that whether there exists a negative definite numerical realization $\tilde{A}_{22}$ determines whether the structural pair is stabilizable. Consequently, it is important to determine when the above claim is true, as follows.

Lemma 7. *Given a reducible structural pair $(\bar{A}, \bar{B})$, where $\bar{A} \in \{0, \star\}^{n \times n}$ is in the form of* (2.3). *Then there exists a numerical realization $\tilde{A}_{22}$ which is negative definite if and only if the diagonal entries of $\bar{A}_{22}$ are all $\star$-entries.*

Proof. See Appendix A.4. □

Combining Lemmas 6 and 7, we have an algebraic condition for structurally stabilizability. In what follows, we present graph-theoretic interpretation of these conditions.

Theorem 15. *Consider a structural pair $(\bar{A}, \bar{B})$, where $\bar{A}$ is symmetrically structured. Let $G(\bar{A}, \bar{B}) = (\mathcal{X} \cup \mathcal{U}, \mathcal{E}_{\mathcal{X},\mathcal{X}} \cup \mathcal{E}_{\mathcal{U},\mathcal{X}})$ be the digraph associated with $(\bar{A}, \bar{B})$, and $\mathcal{X}_r \subseteq \mathcal{X}$ and $\mathcal{X}_u \subseteq \mathcal{X}$ be the subset of state vertices which are input-reachable and input-unreachable, respectively. The $(\bar{A}, \bar{B})$ is structurally stabilizable if and only if the following two conditions hold simultaneously in $G(\bar{A}, \bar{B})$:*

1. *the vertex x_i has a self-loop, $\forall x_i \in \mathcal{X}_u$;*
2. *$|\mathcal{N}(\mathcal{S})| \geq |\mathcal{S}|$, $\forall \mathcal{S} \subseteq \mathcal{X}_r$.*

Proof. See Appendix A.4. □

Essentially, to ensure structural stabilizability, two conditions should hold simultaneously: (i) every unreachable state vertex should have a self-loop, and (ii) the reachable part of the system should be structurally controllable – see Theorem 2 for detailed description on graph-conditions for symmetric structural controllability. Next, we utilize Theorem 15 to characterize the maximum dimension of the stabilizable subspace.

5.2.2 Maximum Dimension of Stabilizable Subspace

Similar to the previous subsection, we will first consider the case when $(\bar{A}, \bar{B})$ is irreducible, then extend the solution approach to the general case.

By Lemma 6, when $(\bar{A}, \bar{B})$ is irreducible, the $(\bar{A}, \bar{B})$ is structurally controllable if and only if it is structurally stabilizable. This motivates us to consider the relationship between controllable subspace and stabilizable subspace. Moreover, it is shown in [36] that the maximum dimension of controllable subspace is equal to the generic dimension of controllable subspace of a structural pair without symmetric parameter constraints. We may suspect that equality also holds when symmetric parameter dependency is considered. Motivated by this intuition, we first study the generic dimension of the controllable subspace, and then extend the derived results to obtain a solution of Problem 8.

Given a structured pair $(\bar{A}, \bar{B})$, where $\bar{A}$ is symmetrically structured, if there exists a proper variety $V \subset \mathbb{R}^{n_{\bar{A}}+n_{\bar{B}}}$, such that $\operatorname{rank}(Q(\tilde{A}, \tilde{B})) = k$ when $[\mathbf{p}_{\tilde{A}}, \mathbf{p}_{\tilde{B}}] \in V^c$, then we say the *generic dimension* [36] of controllable subspace of $(\bar{A}, \bar{B})$, denoted as d_c, is k. For almost all numerical realizations $(\tilde{A}, \tilde{B})$ with $[\mathbf{p}_{\tilde{A}}, \mathbf{p}_{\tilde{B}}] \in \mathbb{R}^{n_{\bar{A}}+n_{\bar{B}}}$ (except for a proper variety, e.g., $[\mathbf{p}_{\tilde{A}}, \mathbf{p}_{\tilde{B}}] \in V$), the dimension of controllable subspace is d_c.

We characterize the generic dimension of controllable subspace of a structural pair involving a symmetrically structured matrix by the following lemma.

Lemma 8. *Given an irreducible structural pair $(\bar{A}, \bar{B})$, where $\bar{A} \in \{0, \star\}^{n\times n}$ is symmetrically structured and $\bar{B} \in \{0, \star\}^{n\times m}$ is structured, the generic dimension of controllable subspace equals to the term rank of $[\bar{A}, \bar{B}]$, i.e., the concatenation of matrices $\bar{A}$ and $\bar{B}$.*

Proof. See Appendix A.4. □

When $(\bar{A}, \bar{B})$ is reducible, we can permute $(\bar{A}, \bar{B})$ to obtain the form in (2.3). By Definition 12 and Theorem 15, the maximum dimension of the stabilizable subspace should be the sum of the generic dimension of controllable subspace and the maximum number of negative eigenvalues over all the numerical realizations of the uncontrollable part. This can be formalized in the following result.

Theorem 16. *Consider a structural pair $(\bar{A}, \bar{B})$, where $\bar{A} \in \{0, \star\}^{n \times n}$ is symmetrically structured. Then,*

1. *if $(\bar{A}, \bar{B})$ is irreducible, then the maximum dimension of stabilizable subspace of $(\bar{A}, \bar{B})$ equals to the generic dimension of controllable subspace of $(\bar{A}, \bar{B})$;*
2. *if $(\bar{A}, \bar{B})$ is reducible, then we permute the matrix $\bar{A}$ into the form* (2.3)*. The m-dim$(\bar{A}, \bar{B})$ is larger or equal to t–rank$([\bar{A}_{11}, \bar{B}_1]) + k$, where k is the total number of $\star$-entries in the diagonal of $\bar{A}_{22}$.*

Proof. See Appendix A.4. □

Remark 11. *In the form* (2.3)*, the index of columns of $\bar{A}_{11}$ are corresponding to input-reachable state vertices in $G(\bar{A}, \bar{B})$, and the index of columns of $\bar{A}_{22}$ are corresponding to the input-unreachable state vertices in $G(\bar{A}, \bar{B})$. The input-reachable/unreachable vertices can be identified by running a depth-first search [163]. Besides, the term-rank of $([\bar{A}_{11}, \bar{B}_1])$ can be obtained by finding a maximum bipartite matching in $\mathcal{B}(\bar{A}, \bar{B})$. Thus, the maximum stabilizable subspace can be determined in polynomial time $\mathcal{O}(n^3)$.*

5.3 Optimal Actuator-Attack and Recovery Problems

In this section, equipped with the results from Section 5.2, we show the NP-hardness of Problem 9 and Problem 10 in Theorem 17 and Theorem 19, respectively. Then, we introduced a greedy algorithm to solve Problem 10 – see Algorithm 7. Besides, we show that Algorithm 7 achieves a $(1 - 1/e)$ approximation guarantee to the optimal solution of Problem 10, which is formally captured in Theorem 20.

5.3.1 Computational Complexity of the Optimal Actuator-disabling Attack Problem

Suppose that there is no self-loop in the system digraph of a structural pair $(\bar{A}, \bar{B})$ and the Condition-2) in Theorem 15 is satisfied. Then, we will show that Problem 9 is equivalent to minimizing the number of input-reachable states by removing a limited number of inputs. This problem shares the similarities with Min-k-Union problem described next.

Definition 13 (Min-k-Union Problem [164])**.** *Given a universe $\mathcal{U}_{\mathcal{S}} = \{\mathcal{S}_\ell\}_{\ell=1}^p$ and an integer $k \in \mathbb{Z}^+$, find*

$$\begin{aligned} \mathcal{L}^* = \arg \min_{\mathcal{L}=\{\ell_i\}_{i=1}^k} & \,|\bigcup_{i=1}^k \mathcal{S}_{\ell_i}| \\ \text{s.t.} \quad & \mathcal{L} \subseteq [p]. \end{aligned} \tag{5.4}$$

Therefore, we aim at selecting a limited number of sets whose union is minimized, leading to the following result.

Theorem 17. *The Optimal Actuator-disabling Attack Problem (Problem 9) is NP-hard.*

Proof. See Appendix A.4. □

Although the problem is NP-hard, that does not imply that all instances of the problem are equally difficult. As a consequence, we now propose to characterize the approximability of Problem 9. In particular, we first consider a subclass of instances of Problem 9, which satisfy the following assumption.

Assumption 3. *The symmetrically structured matrix $\bar{A} \in \{0, \star\}^{n\times n}$ is such that for any $\mathcal{S} \subseteq \mathcal{X}$, where $\mathcal{X}$ is the set of state vertices in the state digraph $G(\bar{A})$, $|\mathcal{N}(\mathcal{S})| \geq |\mathcal{S}|$.*

Assumption 3 ensures that in the bipartite graph associated with $G(\bar{A})$, there is no right-unmatched vertex with respect to any maximum matching, i.e., the Condition-2) in Theorem 15 is always satisfied. We then have the following theorem.

Theorem 18. *Under Assumption 3, denote by m_1 the total number of sets (i.e., $\{\mathcal{S}_i\}_{i=1}^{m_1}$) in an instance of Min-k-Union problem, and m_2 the total number of candidate inputs in an instance of Problem 9. Additionally, let $\rho : \mathbb{Z} \to \mathbb{R}$. Then, there exists a $\rho(m_1)$-approximation algorithm for Min-k-Union problem if and only if there exists a $\rho(m_2)$-approximation algorithm for Problem 9.*

Proof. See Appendix A.4. □

As a result of Theorem 18, any approximation algorithm solving Min-k-Union problem can be adapted to solve Problem 9 with approximation guarantees.

5.3.2 Solution to the Optimal Recovery Problem

To investigate the computation complexity of obtaining a solution to Problem 10, we take a similar strategy to that used in the previous section, i.e., we first consider the following special instance: the pair $(\bar{A}, \bar{B})$ satisfies the Assumption 1. In this case, we will show that Problem 10 is equivalent to adding a limited number of actuators to maximize the total number of input-reachable state vertices, which is similar to the Max-k-Union problem, stated as follows.

Definition 14 (Max-k-Union Problem [165]). *Given a universe $\mathcal{U}_{\mathcal{S}} = \{\mathcal{S}_\ell\}_{\ell=1}^{p}$ and an integer $k \in \mathbb{Z}^+$, find*

$$\begin{aligned} \mathcal{L}^* = \arg \max_{\mathcal{L}=\{\ell_i\}_{i=1}^{k}} & |\bigcup_{i=1}^{k} \mathcal{S}_{\ell_i}| \\ \text{s.t. } & \mathcal{L} \subseteq [p]. \end{aligned} \tag{5.5}$$

Thus, we obtain the following theorem.

Theorem 19. *The Optimal Recovery Problem (Problem 10) is NP-hard.*

Proof. See Appendix A.4. □

A natural approximation solution to optimal design problems is through greedy algorithms [166]. Although greedy algorithms may not provide an optimal solution, under specific objective functions of the problem, a suboptimal solution with suboptimally guarantees can be provided. Specifically, a particular class of problem with such properties is called submodularity function problems, defined as follows.

Definition 15 (Submodular function [166]). *Let Ω be a nonempty finite set. A set function $f\colon 2^\Omega \to \mathbb{R}$, where 2^Ω denotes the power set of Ω, is a* submodular function *if for every $\mathcal{J}_1, \mathcal{J}_2 \subseteq \Omega$ with $\mathcal{J}_1 \subseteq \mathcal{J}_2$ and every $i \in \Omega \setminus \mathcal{J}_2$, we have $f(\mathcal{J}_2 \cup \{i\}) - f(\mathcal{J}_2) \leq f(\mathcal{J}_1 \cup \{i\}) - f(\mathcal{J}_1)$.*

Algorithm 7: $(1 - 1/e)$ approximation solution to Problem 10

```
Input: The pair $(\bar{A}, \bar{B})$, $\bar{B}_{\mathcal{U}_{can}} \in \{0, \star\}^{n \times m'}$, and the budget $k$;
Output: Suboptimal solution $\mathcal{J}$;
Initialize $\mathcal{J} \leftarrow \emptyset$, $\mathcal{L} \leftarrow [m']$;
% $\mathcal{L}$ is the set of indices of new actuators in $\mathcal{U}_{can}$ that can be added to
the system.
for iteration $i \in [k]$ do
    for each $j \in \mathcal{L}$ do
        $d_j \leftarrow \text{m-dim}(\bar{A}, [\bar{B}, \bar{B}_{can}(\mathcal{J} \cup \{j\})])$;
    end for
    $\mathcal{I} \leftarrow \{i \colon d_i = \max\{d_j\}_{j=1}^{|\mathcal{L}|}\}$;
    Pick a $j \in \mathcal{I}$;
    $\mathcal{J} \leftarrow \mathcal{J} \cup \{j\}$;
    $\mathcal{L} \leftarrow \mathcal{L} \setminus \{j\}$;
end for
Return $\mathcal{J}$
```

The greedy algorithm [166] achieves a $(1 - 1/e)$-factor approximation to the optimal solution provided that the objective function is submodular. In this paper, we show that the objective function in Problem 10 is submodular; hence, the greedy algorithm provides a constant factor guarantee to the optimal solutions.

Theorem 20. *Algorithm 7 returns a $(1 - 1/e)$-approximation of the optimal solution to Problem 10.*

Proof. See Appendix A.4. □

Remark 12. *In [167], the authors argue that insofar there is no constant factor approximation to the Min-k-Union problem. Thus, together with Theorem 18, we cannot use the greedy algorithm to approximate Problem 9 with guarantee.*

5.4 Illustrative Examples

In this section, we present examples to illustrate our results on structural stabilizability and approximation solution to Problem 10.

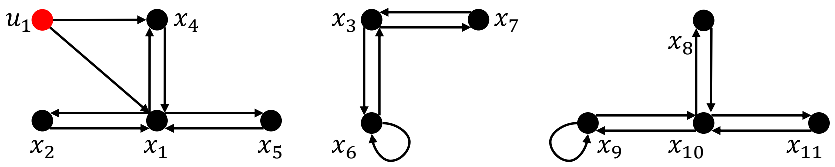


Figure 5-1: In this figure, we depict the structure of $G(\bar{A}, \bar{B})$. The red vertex labeled by u_1 and black vertices labeled by $x_1, \ldots, x_{11}$ are the input vertex and state vertices, respectively. The black arrows represent the edges from input vertex to state vertices, as well as edges between state vertices.

5.4.1 Examples on Maximum Dimension of Stabilizable Subspace

We consider a structural pair $(\bar{A}, \bar{B})$, where $\bar{A} \in \{0, \star\}^{11\times 11}$ is symmetrically structured and $\bar{B} \in \{0, \star\}^{11\times 1}$ is structured.

$$\bar{A} = \begin{bmatrix} 0 & a_{12} & 0 & a_{14} & a_{15} & 0 & 0 & 0 & 0 & 0 & 0 \\ a_{12} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & a_{36} & a_{37} & 0 & 0 & 0 & 0 \\ a_{14} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ a_{15} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & a_{36} & 0 & 0 & a_{66} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & a_{37} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & a_{810} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & a_{99} & a_{910} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & a_{810} & a_{910} & 0 & a_{1011} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & a_{1011} & 0 \end{bmatrix}, \bar{B} = \begin{bmatrix} b_{11} \\ 0 \\ 0 \\ b_{41} \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \tag{5.6}$$

We depict the digraph representation of the structural pair $(\bar{A}, \bar{B})$, denoted by $G(\bar{A}, \bar{B})$, in Figure 5-1. Since x_3 and x_7 are unreachable vertices and they do not have self-loops, the pair $(\bar{A}, \bar{B})$ is not structurally stabilizable due to Theorem 15. Furthermore, the total number of right-matched (with respect to any maximum matching in the associated bipartite graph $\mathcal{B}(\bar{A}, \bar{B})$) reachable vertices is 3, and the total number of unreachable vertices with self-loop is 2. Therefore, by invoking Theorem 16, we conclude that the maximum stabilizable subspace is $3 + 2 = 5$.

5.4.2 Examples on the Optimal Recovery Problem

Now, we present an example to illustrate the use of Algorithm 7. Consider again the structural pair $(\bar{A}, \bar{B})$ specified in (5.6). As noted in the last subsection, the $(\bar{A}, \bar{B})$ is not structurally stabilizable. We let $\mathcal{U}_{can} = \{u_i\}_{i=2}^{7}$ be the set of candidate actuators that can be added into the system and associate it with the structured matrix $\bar{B}_{\mathcal{U}_{can}} \in \{0, \star\}^{11\times 6}$, of which nonzero entries are captured by the red edges of the

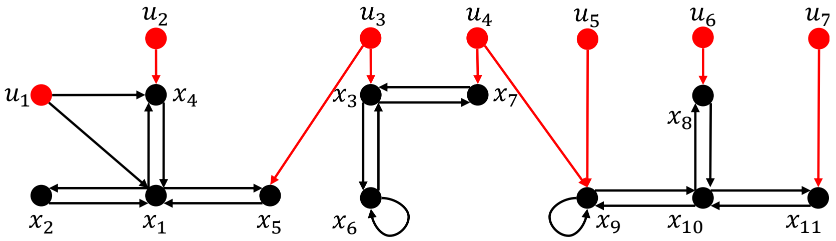


Figure 5-2: In this figure, we depict the digraph $G(\bar{A}, [\bar{B}, \bar{B}_{\mathcal{U}_{can}}])$. We use red and black vertices to represent input vertices and state vertices, respectively. The black and red arrows represent are the edges in $\mathcal{E}_{\mathcal{X},\mathcal{X}} \cup \mathcal{E}_{\{u_1\},\mathcal{X}}$ and edges in $\mathcal{E}_{\mathcal{U}_{can},\mathcal{X}}$, respectively.

digraph $G(\bar{A}, [\bar{B}, \bar{B}_{\mathcal{U}_{can}}])$ depicted in Figure 5-2. We have obtained in the last subsection that m-dim$(\bar{A}, \bar{B})$ is 5. Suppose we have a budget $k = 3$, then Problem 10 consists in adding 3 actuators from $\mathcal{U}_{can}$ into the system such that the maximum stabilizable subspace is maximized. In the first iteration of Algorithm 7, u_4 is selected because m-dim$(\bar{A}, [\bar{B}, \bar{B}_{\mathcal{U}_{can}}(\{4\})])$ − m-dim$(\bar{A}, \bar{B}) = 4 \geq$ m-dim$(\bar{A}, [\bar{B}, \bar{B}_{\mathcal{U}_{can}}(\{i\})])$ − m-dim$(\bar{A}, \bar{B}), \forall u_i \in \mathcal{U}_{can}$. Similarly, in the second iteration, u_3 is selected by Algorithm 7. This results that m-dim$(\bar{A}, [\bar{B}, \bar{B}_{\mathcal{U}_{can}}(\{3, 4\})]) = 10$. Finally, u_7 is selected and m-dim$(\bar{A}, [\bar{B}, \bar{B}_{\mathcal{U}_{can}}(\{3, 4, 7\})]) = 11$. Since the maximum possible stabilizable subspace is always less than or equal to the total number of states, in this example, Algorithm 7 returns an optimal solution to Problem 10.

5.5 General Structural Stabilizability

In Section 5.2, we have used graph theory to characterize necessary and sufficient condition for structural stabilizability – see Theorem 15. However, such a condition is valid only when the underlying system is captured by an undirected graph, i.e., the state matrix A is symmetric. In this section, we aim to extend this result to general structural systems. More specifically, our goal in this section is to solve the following problem:

Problem 11. *Given a continuous-time linear time-invariant system* $\dot{\mathbf{x}} = A\mathbf{x} + B\mathbf{u}$*, we denote by* $(\bar{A}, \bar{B})$ *the structural pattern of* (A, B)*. Let* $G(\bar{A}, \bar{B})$ *be the digraph representation of the structural pair. Find necessary and sufficient conditions in* $G(\bar{A}, \bar{B})$ *such that* $(\bar{A}, \bar{B})$ *is structurally stabilizable.*

5.5.1 Characterizing General Structural Stabilizability

To solve the above problems, we first recall the definition of structural stabilizability – see Definition 11. As discussed in Lemma 6 earlier this chapter, structural stabilizability is equivalent to structural controllability when the structrual pair $(\bar{A}, \bar{B})$ is irreducible. Therefore, in order to characterize structural stabilizability, we consider permuting the pair $(\bar{A}, \bar{B})$ into the form (2.3) and analyze the graph-unreachable part of the structural system. In other words, according to the definition of stabilizability, we have to find conditions on when there exists a numerical realization $\tilde{A}_{22}$ such that all its eigenvalues are contained in the open left-half-plane. This motivates us to introduce the following definition.

Definition 16 (Structural Hurwitz-stability)**.** *A structured matrix $\bar{A}$ is called structurally Hurwitz-stable if there exists a numerical realization $\tilde{A}$ of $\bar{A}$ such that all eigenvalues of $\tilde{A}$ are contained in the open left-half-plane of $\mathbb{C}$.*

Next, we present a few sufficient and/or necessary conditions on when a structured matrix is structurally Hurwitz-stable.

Lemma 9. *Let $\bar{A}$ be a structured matrix, and $G(\bar{A})$ be the digraph representation of $\bar{A}$. If every vertex x_i has a self-loop, then $\bar{A}$ is structurally Hurwitz-stable.*

Proof. If every vertex x_i has a self-loop in $G(\bar{A})$, then $[\bar{A}]_{ii}$ is a $\star$-entry for all $i \in [n]$. In this case, we consider the following assignment on the $\star$-entries of $\bar{A}$:

$$[\tilde{A}]_{ij} = \begin{cases} -1, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases} \tag{5.7}$$

Thus, $\tilde{A} = -I$ is a numerical realization of $\bar{A}$ and all eigenvalues of $\tilde{A}$ are strictly less than 0. □

In the above construction of a numerical realization of $\bar{A}$, we have set all off-diagonal $\star$-entries to zero and all diagonal entries to negative real numbers. Hereafter, we show that it is possible to find a stable numerical realization of $\bar{A}$ with all its $\star$-entries are not equal to 0.

Lemma 10. *Let $\bar{A}$ be a structured matrix and all conditions in Lemma 9 holds in $G(\bar{A})$. Then there exists a numerical realization of $\bar{A}$ with all its $\star$-entries are not equal to* 0 *such that $\tilde{A}$ is (strictly) Hurwitz-stable.*

Proof. Let d_{ij} be the independent parameter of $[\bar{A}]_{ij}$ (provided that $[\bar{A}]_{ij}$ is a $\star$-entry. Since every vertex in $G(\bar{A})$ has a self-loop, it is possible to assign values to d_{ij} such that $d_{ii} < -\sum_{j\neq i} |d_{ij}|$ for all $i, j \in [n]$. Subsequently, using Gershgorin's disk theorem [168], by setting $[\tilde{A}]_{ij} = d_{ij}$ for all $[\bar{A}]_{ij} = \star$ and $[\tilde{A}]_{ij} = 0$ otherwise, the matrix $\tilde{A}$ is Hurwitz-stable. □

The above lemmas provide sufficient conditions on structural stabilizability. However, such a condition is not necessary. Consider the following example where

$$\bar{A} = \begin{bmatrix} \star & \star \\ \star & 0 \end{bmatrix}.$$

In the digraph representation of $\bar{A}$, only the first vertex has a self-loop, violating the assumption in both Lemma 9 and 10. However, consider the following numerical realization:

$$\tilde{A} = \begin{bmatrix} -10 & 3 \\ -3 & 0 \end{bmatrix}.$$

It is easy to see that $\tilde{A}$ is Hurwitz-stable since both of its eigenvalues are strictly less than 0. To partly strengthen the above graph-theoretical condition, we next present a lemma that characterizes a necessary condition for structural Hurwitz-stability.

Lemma 11. *Let $\bar{A}$ be a structured matrix, if there exists a (strictly) Hurwitz-stable numerical realization $\tilde{A}$, then $G(\bar{A})$ must contain at least one self-loop.*

Proof. We proof this lemma by contradiction. Let us suppose that $G(\bar{A})$ does not contain any self-loop. Following this assumption, all diagonal elements of $\bar{A}$ are fixed-zeros. Let $\tilde{A}$ be an arbitrary numerical realization of $\bar{A}$, and denote by λ_i the i-th eigenvalue of

$\tilde{A}$. Since $\tilde{A}$ is asymmetric, we can write $\lambda_i = \sigma_i + j\omega_i$ for all $i \in [n]$. In order words, we denote by σ_i and ω_i the real and imaginary part of the eigenvalue λ_i. Subsequently, we have that

$$\begin{aligned} \mathrm{Tr}(\tilde{A}) &= \sum_{i}^{n} \lambda_i \\ &= \sum_{i=1}^{n} \sigma_i + j\omega_i = \sum_{i=1}^{n} \sigma_i < 0, \end{aligned}$$

where the last equality is due to the fact that solutions to characteristic polynomials come in pairs. However, since all diagonal elements of $\tilde{A}$ are zero, $\mathrm{Tr}(\tilde{A})$ is equal to 0, which is a contraction with the above derivation. □

Part II

Networked System Analysis via Measure Theory

Chapter 6

Bounds on the Spectral Radius of Digraphs

In the first part of the thesis, we have used tools from structural systems theory and graph theory to characterize properties in symmetrically structured linear systems. In this chapter, we analyze global system properties using, solely, local structural information. This chapter is organized as follows: In Section 6.1, we introduce certain notions from algebraic graph theory used in our derivations. In Section 6.2, we relate the spectral moments of a digraph to subgraph counts and introduce the truncated K-moment problem from functional analysis, which we then use to upper and lower bound the spectral radius using subgraph counts. In Section 6.3, we propose a refine approach to find more accurate bounds on spectral radius by analyzing the skew-symmetric part of the adjacency matrix. We numerically validate the quality of our bounds using randomly generated directed graphs, as well as real networks in Section 6.4.

6.1 Adjacency Matrix and Digraph Isomorphism

In the rest of this chapter, we adopt notations introduced in Subsection 2.1.2. However, we assume that the digraph under consideration is simple. Additionally, a subgraph

G_s is called a *bidirected edge* if $\mathcal{V}_s = \{i, j\}$ and $\mathcal{E}_s = \{(i, j), (j, i)\}$, where $i, j \in \mathcal{V}$. A subgraph G_s is called a *directed triangle* if $\mathcal{V}_s = \{i, j, k\}$ and $\mathcal{E}_s = \{(i, j), (j, k), (k, i),$ where $i, j, k \in \mathcal{V}$.

A digraph G can be represented by an *adjacency matrix* $A \in \mathbb{R}^{n\times n}$, whose entries are defined as $[A]_{ij} = 1$ if $(j, i) \in \mathcal{E}$; $[A]_{ij} = 0$ otherwise. Particularly, if the graph is undirected, then $A = A^\top$ and all its eigenvalues are real. When the digraph is simple, all the diagonal entries of A are zero. In what follows, we use $\lambda_1, \dots, \lambda_n$ to denote the eigenvalues of A. The eigenvalue spectrum of A is denoted by $\texttt{spec}(A) = \{\lambda_i\}_{i=1}^n$. Moreover, the real part (respectively, imaginary part) of λ_i is denoted as σ_i (respectively, ω_i). Without loss of generality, we assume $|\lambda_1(A)| \leq \cdots \leq |\lambda_n(A)|$. The *spectral radius* of A is defined as $|\lambda_n(A)|$. Furthermore, we denote by $\omega_{\max}(A) = \max_i |\omega_i|$.

Two directed subgraphs $G_s, G_h \subseteq G$ are said to be *isomorphic* [22], denoted by $G_s \simeq G_h$, if there exist a bijection $f : \mathcal{V}_s \to \mathcal{V}_h$ such that $(u, v) \in \mathcal{E}_s$ if and only if $(f(u), f(v)) \in \mathcal{E}_h$ for all $u, v \in \mathcal{V}_s$. When G_s and G_h are non-isomorphic, we write $G_s \not\simeq G_h$. In particular, when $\mathcal{V}_s = \mathcal{V}_h$, the bijection f is called an *automorphism* and the two directed subgraphs G_s and G_h are said to be *automorphic*, denoted by $G_s \overset{a}{\simeq} G_h$. Consequently, an automorphism is an equivalence relation on the set of directed subgraphs of the same order, i.e., it classifies all possible directed subgraphs into equivalent classes. Based on these notions, we define the *isomorphic group* (respectively, *automorphic group*) of a directed subgraph $G_s \subseteq G$ by $\texttt{Iso}(G_s, G) = \{G_h \subseteq G : G_h \simeq G_s\}$ (respectively, $\texttt{Auto}(G_s, G) = \{G_h \subseteq G : G_h \overset{a}{\simeq} G_s\}$). Given a directed subgraph $G_s \subseteq G$, the *count* of G_s is defined by

$$\texttt{Count}(G_s, G) = \frac{|\texttt{Iso}(G_s, G)|}{|\texttt{Auto}(G_s, G)|}.$$

Finally, let Ξ_s be the set of weakly-connected digraphs of order s. We denote by $\Omega_s \subseteq \Xi_s$, the set of non-isomorphic strongly-connected digraphs of order s.

6.2 Analyzing Spectral Radius Using Subgraph Counts

In the next subsections, we will establish a connection between the spectral moments of G and the counts of certain subgraphs. Later on, in Subsection 6.2.3, we will exploit recent results regarding the existence of measures with a given sequence of moments to derive upper and lower bounds on the spectral radius of the graph in terms of these subgraph counts (presented in Subsection 6.2.4). These bounds will be further refined in Subsection 6.3.

6.2.1 From Subgraphs to Closed Walks

The eigenvalues of the adjacency matrix of a digraph are closely related to the walks within the digraph, as stated in the following lemma:

Lemma 12 ([169]). *Let A be the adjacency matrix of a simple digraph G. Given a positive integer k, $\mathrm{Tr}(A^k)$ is equal to the total number of closed walks of length k in G.*

Hereafter, we derive a relationship between the $\mathrm{Tr}(A^k)$ and the counts of subgraphs of different sizes. To illustrate the idea behind our approach with a simple case, let us decompose $\mathrm{Tr}(A^2)$ (i.e., $k = 2$), as follows:

$$\mathrm{Tr}(A^2) = \sum_{i=1}^{n} [A^2]_{ii} = \sum_{i=1}^{n} \sum_{j=1}^{n} [A]_{ij} [A]_{ji} = \sum_{i,j:\ (i,j),(j,i) \in \mathcal{E}} 1, \tag{6.1}$$

Note that the last term is counting (twice) the number of bidirected-edge subgraphs, e.g., pairs of vertices connected by two directed edges with reciprocal directions. For clarity, let us also consider the case $k = 3$. In this case, we can decompose the trace as,

$$\mathrm{Tr}(A^3) = \sum_{i=1}^{n} [A^3]_{ii} = \sum_{i,j,k:\ (i,j),(j,k),(k,i) \in \mathcal{E}} 1. \tag{6.2}$$

Therefore, $\mathrm{Tr}(A^3)$ is equal to (three times) the number of directed triangles in G.

More generally, for given $k \in \mathbb{N}$, we prove the following theorem:

Theorem 21. *Consider a (simple) digraph G with adjacency matrix A. For all $\widehat{G} \in \Omega_s$*

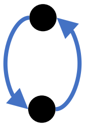
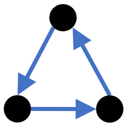
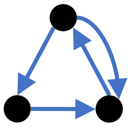
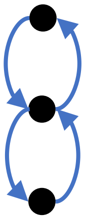
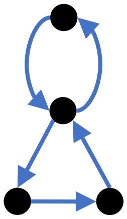
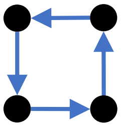
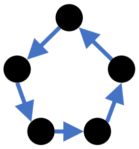
							
Tr(A^2)	2	0	0	0	0	0	0
Tr(A^3)	0	3	0	0	0	0	0
Tr(A^4)	2	0	0	4	0	4	0
Tr(A^5)	0	0	5	0	5	0	5

Figure 6-1: This table shows the values of $\eta(\widehat{G}, k)$, defined in Theorem 21, for $k \leq 5$. The value k indexes the powers of A in the rows, while the columns of the table are indexed by all non-isomorphic strongly-connected subgraphs of order at most 5 involved in the computation of the traces up to the fifth power. For example, from the second row of the table, we infer that $\text{Tr}(A^3)$ equals 3 times the number of directed triangles (second column in the table).

and all positive integers k, we define $\eta(\widehat{G}, k)$ as the number of closed walks of length k in $\widehat{G}$ visiting all the edges of $\widehat{G}$ at least once. Then, the following holds

$$\text{Tr}(A^k) = \sum_{s=2}^{k} \sum_{\widehat{G} \in \Omega_s} \eta(\widehat{G}, k) \, \texttt{Count}(\widehat{G}, G). \tag{6.3}$$

Proof. See Appendix A.5. □

Based on Theorem 21, we can fill a table with the values of $\eta(\widehat{G}, k)$ for different values of k (see Figure 6-1). The rows in this table are indexed by those subgraphs involved in the computation of the traces up to the fifth power. The coefficients in this table can then be used to compute $\text{Tr}(A^k)$ for $k \leq 5$, as a linear combination of the counts of the subgraphs plotted in the table. For example, from the first row of the table, we infer that $\text{Tr}(A^2)$ is equal to two times the count of bidirected-edge subgraphs. Similarly, from the second row, we infer that $\text{Tr}(A^3)$ equals three times the count of directed triangles.

6.2.2 From Subgraph Counts to Spectral Moments

In this subsection, we derive a relationship between closed walks in G and the power-sums of the eigenvalues in A. To achieve this goal, we first introduce some notions from probability theory. Let μ be a measure on $\mathbb{R}^n$. The support of μ, denoted as $\mathrm{Supp}(\mu)$, is defined as the smallest closed set $C \subseteq \mathbb{R}^n$ such that $\mu(\mathbb{R}^n \setminus C) = 0$, [170]. The measure μ is called r-atomic if $|\,\mathrm{Supp}(\mu)| = r$, i.e., a discrete set of cardinality r. The k-th moment of an $\mathbb{R}$-valued random variable x is defined as $\mathbb{E}[x^k] = \int_{\mathbb{R}} x^k d\mu_x$, where μ_x is the corresponding probability measure of x. Given an $\mathbb{R}^n$-valued random variable $\mathbf{x}$, and an n-dimensional vector of integers $\boldsymbol{\alpha} \in \mathbb{N}^n$, we let $\mathbf{x}^{\boldsymbol{\alpha}} = \prod_{i=1}^n x_i^{\boldsymbol{\alpha}_i}$. Subsequently, the $\boldsymbol{\alpha}$-moment of $\mathbf{x}$ is defined as $\mathbb{E}[\mathbf{x}^{\boldsymbol{\alpha}}] = \int_{\mathbb{R}^n} \prod_{i=1}^n x_i^{\boldsymbol{\alpha}_i} d\boldsymbol{\mu}$, where $\boldsymbol{\mu}$ is the probability measure of $\mathbf{x} = [x_1, \dots, x_n]^\top$. Moreover, the *order* of $\boldsymbol{\alpha}$ is defined by $|\boldsymbol{\alpha}| = \sum_{i=1}^n \alpha_i$.

Given a digraph G, we define the *spectral measure* of its adjacency matrix A as the following two-dimensional probability density:

$$\mu_A(x, y) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - \sigma_i)\delta(y - \omega_i), \tag{6.4}$$

where $\delta(\cdot)$ is the Dirac's delta measure, i.e., the probability measure on $\mathbb{R}$ that assigns unit mass to the origin, and zero elsewhere. In other words, the spectral measure μ_A is a discrete probability measure on $\mathbb{R}^2$ assigning a mass $1/n$ to each one of the n points in the set $\{(\sigma_i, \omega_i)\}_{i=1}^n$. Furthermore, we define the $\boldsymbol{\alpha}$-spectral moments of G, where $\boldsymbol{\alpha} = [a, b]^\top \in \mathbb{N}^2$, as the $\boldsymbol{\alpha}$-moment of the spectral measure μ_A, given by

$$m_{\boldsymbol{\alpha}}(A) = \int_{\mathbb{R}^2} x^a y^b d\mu_A(x, y). \tag{6.5}$$

We also write $m_{ab}(A)$ as an abbreviation of $m_{\boldsymbol{\alpha}}(A)$. As demonstrated in [105], the spectral moments of an undirected graph can be computed as a linear combination of the counts of certain non-isomorphic subgraphs. Hereafter, we derive a similar relationship between the spectral moments of a digraph G and the counts of certain (directed) subgraphs contained in G. To achieve this goal, we start by deriving a closed-form expression of the $\boldsymbol{\alpha}$-spectral moments, as stated in the following lemma.

Lemma 13. *Given a directed graph G with adjacency matrix A, it holds that*

$$\operatorname{Tr}\left(A^k\right) = \sum_{s=0}^{\lfloor k/2 \rfloor} \binom{k}{2s} (-1)^s \, n \, m_{2s,k-2s}(A), \ \textit{for all } k \in \mathbb{N}. \tag{6.6}$$

Proof. From (6.4) and (6.5), the $\boldsymbol{\alpha}$-moment of the spectral measure for $\boldsymbol{\alpha} = [a,b]^\top$ equals

$$\begin{aligned} m_{ab}(A) &= \int_{\mathbb{R}} \int_{\mathbb{R}} x^a y^b \frac{1}{n} \sum_{i=1}^{n} \delta(x-\sigma_i)\,\delta(y-\omega_i)\,dxdy \\ &= \frac{1}{n} \sum_{i=1}^{n} \left[\int x^a \delta(x-\sigma_i)\,dx\right] \left[\int y^b \delta(y-\omega_i)\,dy\right] \\ &= \frac{1}{n} \sum_{i=1}^{n} \sigma_i^a \omega_i^b, \end{aligned}$$

where σ_i and ω_i are the real and imaginary part of the i-th eigenvalue of A, respectively. Since $\operatorname{Tr}(A^k)$ equals the sum of the k-th powers of the eigenvalues of A, we have that

$$\begin{aligned} \operatorname{Tr}\left(A^k\right) &= \sum_{i=1}^{n} (\sigma_i + j\omega_i)^k = \sum_{i=1}^{n} \sum_{r=0}^{k} \binom{k}{r} j^r \omega_i^r \sigma_i^{k-r} \\ &= \sum_{i=1}^{n} \left(\sum_{s=0}^{\lfloor k/2 \rfloor} \binom{k}{2s} (-1)^s \omega_i^{2s} \sigma_i^{k-2s} + j \sum_{s=0}^{\lfloor k/2 \rfloor} \binom{k}{2s+1} (-1)^s \omega_i^{2s+1} \sigma_i^{k-2s+1} \right) \\ &= \sum_{s=0}^{\lfloor k/2 \rfloor} \binom{k}{2s} (-1)^s \sum_{i=1}^{n} \omega_i^{2s} \sigma_i^{k-2s} + j \sum_{s=0}^{\lfloor k/2 \rfloor} \binom{k}{2s+1} (-1)^s \sum_{i=1}^{n} \omega_i^{2s+1} \sigma_i^{k-2s+1} \\ &= \sum_{s=0}^{\lfloor k/2 \rfloor} \binom{k}{2s} (-1)^s \, n \, m_{2s,k-2s}(A), \end{aligned}$$

Notice that the imaginary term vanishes in the last equality, since $\operatorname{Tr}\left(A^k\right)$ is a purely real quantity. □

Combining Theorem 21 and (6.6), we have that

$$\sum_{s=2}^{k} \sum_{\widehat{G} \in \Omega_s} \eta(\widehat{G}, k) \, \texttt{Count}(\widehat{G}, G) = \sum_{s=0}^{\lfloor k/2 \rfloor} \binom{k}{2s} (-1)^s \, n \, m_{2s,k-2s}(A), \tag{6.7}$$

for all $k \in \mathbb{N}$. This expression allows us to directly relate the moments of the spectral measure of A to the counts of certain subgraphs in G.

6.2.3 The K-moment Problem

In many practical applications, such as the analysis of large-scale social networks, we do not have access to the whole topology of the graph G. Therefore, it is not possible to explicitly compute the eigenvalues of A. However, it may be possible to retrieve local structural information in the form of subgraph counts by crawling the network. Since in this situation it is not possible to exactly compute all the eigenvalues of A, it would be interesting to have tools allowing us to infer spectral information, such as bounds on eigenvalues, from the counts of small subgraphs in G. This is the main aim of this paper.

As we will show below, the counts of certain subgraphs can be used to constraint the moments of the spectral measure, which can then be used to find bounds on the spectral radius. In particular, from the counts of certain subgraphs of order less or equal to k, we can write down an equality constraint for linear combinations of spectral moments using (6.7). However, it may be possible to find many different spectral measures (with different supports) satisfying the linear constraints in (6.7). In what follows, we will exploit recent results in the multidimensional moment problem [108] to compute outer and inner bounds on the set of all all possible spectral supports. This result will directly provide us with upper and lower bounds on the spectral radius of A.

To explain our approach, we first need to introduce the K-moment problem [108] and related notions. A sequence $\mathbf{y} = \{y_{\boldsymbol{\alpha}}\}$ indexed by $\boldsymbol{\alpha} \in \mathbb{N}^n$ is called a *multi-sequence*. We will use multi-sequences to index the moments of $\mathbb{R}^n$-valued random variables. In particular, given a $\mathbb{R}^2$-valued random variable $\mathbf{x} \sim \mu$ and an index $\boldsymbol{\alpha} = [a, b]^\top \in \mathbb{N}^2$, we will use the notation $y_{\boldsymbol{\alpha}} = y_{ab}$ to denote the $\boldsymbol{\alpha}$-moment of μ, i.e., $y_{ab} = \mathbb{E}[\mathbf{x}^{[a,b]^\top}] = \int_{\mathbb{R}^2} x^a y^b d\mu(x, y)$.

Definition 17. *Let K be a closed subset of $\mathbb{R}^n$. Let $\mathbf{y}_{n,\infty} = \{y_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathbb{N}^n}$ be an infinite real* *multi-sequence. A measure μ on $\mathbb{R}^n$ is said to be a K-representing measure for $\mathbf{y}_{n,\infty}$ if*

$$y_{\boldsymbol{\alpha}} = \int_{\mathbb{R}^n} \mathbf{x}^{\boldsymbol{\alpha}} d\mu(\mathbf{x}), \text{ for all } \boldsymbol{\alpha} \in \mathbb{N}^n, \tag{6.8}$$

and

$$\mathtt{Supp}(\mu) \subseteq K. \tag{6.9}$$

If $\mathbf{y}_{n,\infty}$ has a K-representing measure, we say that $\mathbf{y}_{n,\infty}$ is K-feasible. Similarly, a finite real multi-sequence $\mathbf{y}_{n,2r} = \{y_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{N}^n, |\boldsymbol{\alpha}|\leq 2r}$ is said to be K-feasible if there exists a measure μ with $\mathtt{Supp}(\mu) \subseteq K$ such that (6.8) holds for all $\boldsymbol{\alpha} \in \mathbb{N}^n_{2r}$.

In this paper, we are interested in the case when K is characterized by polynomial inequalities, as stated below.

Definition 18. *A set $K \subseteq \mathbb{R}^n$ is called a* semi-algebraic *set if there exist m polynomials $g_i : \mathbb{R}^n \to \mathbb{R}$ such that*

$$K = \{\mathbf{x} \in \mathbb{R}^n \colon g_i(\mathbf{x}) \geq 0 \text{ for all } i \in [m]\}. \tag{6.10}$$

A necessary and sufficient condition to determine whether a finite multi-sequence is K-feasible, restricted to the case when K is both semi-algebraic and compact, can be stated in terms of linear matrix inequalities involving *moment matrices* and *localizing matrices*, defined below.

Definition 19. [108] *Let $\mathbf{y}_{n,2r} = \{y_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{N}^n_{2r}}$ be a finite real multi-sequence. The* moment matrix *of $\mathbf{y}_{n,2r}$, denoted by $M_r(\mathbf{y}_{n,2r})$, is defined as the real matrix indexed by $\mathbb{N}^n_r$ and having the entries*

$$[M_r(\mathbf{y}_{n,2r})]_{\boldsymbol{\alpha},\boldsymbol{\beta}} = y_{\boldsymbol{\alpha}+\boldsymbol{\beta}}, \tag{6.11}$$

for all $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{N}^n_r$.

In this paper, we consider a particular order while indexing the entries of the moment matrix, as described below. Consider $\mathbf{x} = [x_1, \ldots, x_n]^\top$, and let

$$\mathcal{M} = \{1, x_1, \ldots, x_n, x_1^2, x_1x_2, \ldots, x_n^2, \ldots, x_1^r, x_1^{r-1}x_2, \ldots, x_n^r\}$$

be the set of monomials with degree up to r, written in degree-lexicographic order. The cardinality[1] of $\mathcal{M}$ is given by $\binom{n+r}{n}$. Given an $\mathbb{R}^n$-valued random variable $\mathbf{x}$, suppose that $\mathbf{y}_{n,2r} = \{y_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathbb{N}^n_{2r}}$ is a moment sequence of $\mathbf{x}$, i.e., $y_{\boldsymbol{\alpha}} = \mathbb{E}[\mathbf{x}^{\boldsymbol{\alpha}}]$ for all $\boldsymbol{\alpha} \in \mathbb{N}^n_{2r}$. Then, according to Definition 21, the moment matrix of $\mathbf{y}_{n,2r}$ is expressed entry-wise by (6.11). In this case, we have

$$[M_r(\mathbf{y}_{n,2r})]_{\boldsymbol{\alpha},\boldsymbol{\beta}} = y_{\boldsymbol{\alpha}+\boldsymbol{\beta}} = \mathbb{E}[\mathbf{x}^{\boldsymbol{\alpha}}\mathbf{x}^{\boldsymbol{\beta}}],$$

for all $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{N}^n_r$. The right-hand side of the above equality can be viewed as taking the expectation of the product between the $\boldsymbol{\alpha}$-th and the $\boldsymbol{\beta}$-th monomial in $\mathcal{M}$. We use degree-lexicographic ordering to locate these moments inside the moment matrix. Consequently, the exponent of the monomials in $\mathcal{M}$ index the columns and rows in $M_r(\mathbf{y}_{n,2r})$, as shown in the example below.

Example 1. *Let* $n = 2$, $r = 1$, *and* $\mathbf{y}_{2,2} = \{y_{00}, y_{01}, y_{10}, y_{11}, y_{02}, y_{20}\}$. *Suppose* $\boldsymbol{\alpha} = [0, 1]^\top$ *and* $\boldsymbol{\beta} = [1, 0]^\top$, *then* $[M_1(\mathbf{y}_{2,2})]_{\boldsymbol{\alpha},\boldsymbol{\beta}} = y_{11}$. *Moreover, according to Definition 21, the moment matrix of* $\mathbf{y}_{2,2}$ *is:*

$$\begin{aligned} M_1(\mathbf{y}_{2,2}) &= \begin{bmatrix} \mathbb{E}\left[\mathbf{x}^{[00]^\intercal}\mathbf{x}^{[00]^\intercal}\right] & \mathbb{E}\left[\mathbf{x}^{[00]^\intercal}\mathbf{x}^{[10]^\intercal}\right] & \mathbb{E}\left[\mathbf{x}^{[00]^\intercal}\mathbf{x}^{[01]^\intercal}\right] \\ \mathbb{E}\left[\mathbf{x}^{[10]^\intercal}\mathbf{x}^{[00]^\intercal}\right] & \mathbb{E}\left[\mathbf{x}^{[10]^\intercal}\mathbf{x}^{[10]^\intercal}\right] & \mathbb{E}\left[\mathbf{x}^{[10]^\intercal}\mathbf{x}^{[01]^\intercal}\right] \\ \mathbb{E}\left[\mathbf{x}^{[01]^\intercal}\mathbf{x}^{[00]^\intercal}\right] & \mathbb{E}\left[\mathbf{x}^{[01]^\intercal}\mathbf{x}^{[10]^\intercal}\right] & \mathbb{E}\left[\mathbf{x}^{[01]^\intercal}\mathbf{x}^{[01]^\intercal}\right] \end{bmatrix} \\ &= \begin{bmatrix} y_{00} & y_{10} & y_{01} \\ y_{10} & y_{20} & y_{11} \\ y_{01} & y_{11} & y_{02} \end{bmatrix}. \end{aligned}$$

The *localizing matrix* of a multi-sequence $\mathbf{y}_{n,2r}$ with respect to a polynomial $g : \mathbb{R}^n \to \mathbb{R}$ is defined as follows:

Definition 20. *Consider a multivariate polynomial of degree* v, $g(\mathbf{x}) = \sum_{\boldsymbol{\gamma} \in \mathbb{N}^n_v} u_{\boldsymbol{\gamma}} \mathbf{x}^{\boldsymbol{\gamma}}$, *and a finite multi-sequence* $\mathbf{y}_{n,2r} = \{y_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathbb{N}^n_{2r}}$. *The localizing matrix of* $\mathbf{y}_{n,2r}$ *with respect*

[1] The cardinality of the set $\mathcal{M}$ can be derived by a star-and-bar argument in combinatorial mathematics, see for example [170].

to g, denoted by $L_r(g, \mathbf{y}_{n,2r})$, is defined by the real matrix[2]*:*

$$[L_r(g, \mathbf{y}_{n,2r})]_{\boldsymbol{\alpha},\boldsymbol{\beta}} = \sum_{\boldsymbol{\gamma} \in \mathbb{N}_v^n} u_{\boldsymbol{\gamma}} y_{\boldsymbol{\gamma}+\boldsymbol{\alpha}+\boldsymbol{\beta}}, \tag{6.12}$$

for all $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{N}_r^n$.

Example 2. *Consider Example 1 with $n = 2$ and $r = 1$. Suppose that $g(\mathbf{x}) = a - x_1 + x_2^2$, then $\mathbf{u} = \{u_{00}, u_{10}, u_{02}\}$ with $u_{00} = a, u_{10} = -1, u_{02} = 1$. Subsequently, according to (6.12), $L_1(g, \mathbf{y}_{2,2})$ equals*

$$L_1(g, \mathbf{y}_{2,2}) = \begin{bmatrix} ay_{00} - y_{10} + y_{02} & ay_{10} - y_{20} + y_{12} & ay_{01} - y_{11} + y_{03} \\ ay_{10} - y_{20} + y_{12} & ay_{20} - y_{30} + y_{02} & ay_{11} - y_{21} + y_{13} \\ ay_{01} - y_{11} + y_{03} & ay_{11} - y_{21} + y_{13} & ay_{02} - y_{12} + y_{04} \end{bmatrix}.$$

Hereafter, whenever clear from the context, we adopt the short-handed notation M_r to represent $M_r(\mathbf{y}_{n,2r})$, and $L_r(g)$ to represent $L_r(g, \mathbf{y}_{n,2r})$.

A necessary and sufficient condition for a finite multi-sequence $\mathbf{y} = \{y_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathbb{N}_r^n}$ being K-feasible is stated below.

Theorem 22. [108] *Let $K \subseteq \mathbb{R}^n$ be a semi-algebraic set defined by (6.10) and $v = \max_j \lceil \frac{deg(g_j)}{2} \rceil$. Given a finite multi-sequence $\mathbf{y}_{n,2r} = \{y_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathbb{N}_{2r}^n}$, there exists a $\texttt{rank}(M_{r-v})$-atomic K-representing measure for $\mathbf{y}_{n,2r}$, if and only if,*

$$\begin{aligned} &M_r(\mathbf{y}_{n,2r}) \succeq 0, \text{ and } L_{r-v}(g_j, \mathbf{y}_{n,2r}) \succeq 0, \text{ for all } j \in [m], \\ &\texttt{rank}(M_r(\mathbf{y}_{n,2r})) = \texttt{rank}(M_{r-v}(\mathbf{y}_{n,2r})). \end{aligned} \tag{6.13}$$

In addition to this theorem, we present a corollary that is useful in the development of our framework.

Corollary 3. *Let $K \subseteq \mathbb{R}^n$ be a semi-algebraic set defined as in (6.10) and $v = \max_j \lceil \frac{deg(g_j)}{2} \rceil$. Given a finite multi-sequence $\mathbf{y}_{n,2r} = \{y_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathbb{N}_{2r}^n}$, if $\mathbf{y}_{n,2r}$ is K-feasible,*

[2] As described above, the elements of this matrix are ordered using degree-lexicographic ordering of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

then

$$M_r(\mathbf{y}_{n,2r}) \succeq 0, \text{ and } L_{r-v}(g_j\mathbf{y}_{n,2r}) \succeq 0, \text{ for all } j \in [m]. \tag{6.14}$$

Based on Theorem 22, one can verify whether a given multi-sequence is K-feasible by verifying the positive semidefiniteness of finitely many matrices. In the next subsection, we make use of Theorem 22 to provide upper and lower bounds on spectral radius of a directed graph given counts of subgraphs contained in G up to order r.

6.2.4 Lower Bounds using the K-moment Problem

In this subsection, we aim to obtain upper and lower bounds for the spectral radius of A by leveraging the connection between subgraph counts and the spectral moments of G, as shown in (6.7). To obtain a lower bound on the spectral radius, we use the theory behind the K-moment problem to characterize all K-feasible multi-sequences, $\mathbf{y}_{2,d} = \{y_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{N}_d^2}$, for particular choices of K. Following this idea, we next present necessary conditions for the existence of a *spectral* measure supported on K.

As shown in (6.7), the moments of a (spectral) measure must obey linear constraints imposed by the counts of certain subgraphs in G. In other words, if a multi-sequence $\mathbf{y}_{2,d}$ is a feasible spectral moment sequence, then there exists a spectral measure μ_A such that $y_{\boldsymbol{\alpha}} = \mathbb{E}_{\mu_A}[\mathbf{x}^{\boldsymbol{\alpha}}]$, for all $\boldsymbol{\alpha} \in \mathbb{N}_d^2$ (see Definition 17). Furthermore, according to (6.6), the entries of the sequence $\mathbf{y}_{2,d}$ must satisfy the following linear constraints:

$$\sum_{s=2}^{k} \sum_{\widehat{G}\in\Omega_s} \eta(\widehat{G}, k)\, \mathtt{Count}(\widehat{G}, G) = n \sum_{s=0}^{\lfloor k/2 \rfloor} \binom{k}{2s} (-1)^s\, y_{k-2s,2s}, \tag{6.15}$$

for $k \in [d]$, where the left-hand side is a function of the counts of certain subgraphs of order up to d.

In addition to the above linear constraint, we notice that $\{\lambda_i\}_{i=1}^n$ are the eigenvalues of an adjacency matrix and the eigenvalue spectrum of A is symmetric with respect to the real-axis in the complex plane. Therefore, the moments of a spectral measure must satisfy:

$$y_{ab} = 0 \text{ for } b \text{ odd.} \tag{6.16}$$

Furthermore, when a and b are both even, we have that $m_{ab}(A) = \frac{1}{n}\sum_{i=1}^{n} \sigma_i^a \omega_i^b \geq 0$. Therefore, the moments of a spectral measure must also satisfy:

$$y_{ab} \geq 0 \text{ for } a \text{ and } b \text{ even.} \tag{6.17}$$

Let us define $r = \lfloor \frac{d}{2} \rfloor$. In order to ensure that $\mathbf{y}_{2,d}$ is a feasible spectral moment sequence, the moment matrix defined by

$$[M_r]_{\boldsymbol{\alpha}\boldsymbol{\beta}} = y_{\boldsymbol{\alpha}+\boldsymbol{\beta}}, \text{ for } \boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{N}_r^2 \tag{6.18}$$

must be positive semidefinite according to Corollary 3. Furthermore, since A is entry-wise non-negative, the spectral radius of A equals λ_n according to Perron-Frobenius theory [168]. This also implies that $\omega_i \leq \rho$ for all $i \in [n]$ and $\rho = \lambda_n$. Consequently, the support of the spectral measure of A is contained in the square

$$S = \{\mathbf{x} \in \mathbb{R}^2 \colon x_1 \in [-\rho, \rho], x_2 \in [-\rho, \rho]\}.$$

Let $\mathbf{x} = [x_1, x_2]^\top$ and define the polynomials $g_1(\mathbf{x}) = \rho - x_1$, $g_2(\mathbf{x}) = x_1 + \rho$, $g_3(\mathbf{x}) = \rho - x_2$, and $g_4(\mathbf{x}) = x_2 + \rho$. The set S can be defined by

$$S = \{\mathbf{x} \in \mathbb{R}^2 \colon g_i(\mathbf{x}) \geq 0, \text{ for } i \in [4]\},$$

which is both compact and semi-algebraic. According to Corollary 3, the localizing matrices of $\mathbf{y}_{2,d}$ with respect to $\{g_i\}_{i\in[4]}$ must be positive semidefinite. These matrices are given, entry-wise, by

$$[L_r(g_1)]_{\boldsymbol{\alpha}\boldsymbol{\beta}} = \rho y_{\boldsymbol{\alpha}+\boldsymbol{\beta}} - y_{\boldsymbol{\alpha}+\boldsymbol{\beta}+[1,0]^\top}, \tag{6.19}$$

$$[L_r(g_2)]_{\boldsymbol{\alpha}\boldsymbol{\beta}} = \rho y_{\boldsymbol{\alpha}+\boldsymbol{\beta}} + y_{\boldsymbol{\alpha}+\boldsymbol{\beta}+[1,0]^\top}, \tag{6.20}$$

$$[L_r(g_3)]_{\boldsymbol{\alpha\beta}} = \rho y_{\boldsymbol{\alpha}+\boldsymbol{\beta}} - y_{\boldsymbol{\alpha}+\boldsymbol{\beta}+[0,1]^\top}, \tag{6.21}$$

$$[L_r(g_4)]_{\boldsymbol{\alpha\beta}} = \rho y_{\boldsymbol{\alpha}+\boldsymbol{\beta}} + y_{\boldsymbol{\alpha}+\boldsymbol{\beta}+[0,1]^\top}, \tag{6.22}$$

for $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{N}_r^2$. Therefore, the moment sequence $\mathbf{y}_{2,d}$ of the spectral measure of a matrix with spectral radius ρ must satisfy (6.15)–(6.17), and the moment and localizing matrices defined in (6.18)–(6.22) must be positive semidefinite.

Remark 13. *Notice that, since $|\lambda_i| \leq \rho$ for all $i \in [n]$, the support of the spectral measure is also contained in the circle*

$$S_c = \{[x, y]^\top \in \mathbb{R}^2 \colon x^2 + y^2 \leq \rho^2\}.$$

Defining $g_c = \rho^2 - x^2 - y^2$, we have that $S_c = \{[x, y]^\top \in \mathbb{R}^2 \colon g_c([x, y]^\top) \geq 0\}$. Therefore, the localizing matrix with respect to g_c of the moment sequence $\mathbf{y}_{2,d}$, given by

$$[L_r(g_c)]_{\boldsymbol{\alpha\beta}} = \rho^2 y_{\boldsymbol{\alpha}+\boldsymbol{\beta}} - y_{\boldsymbol{\alpha}+\boldsymbol{\beta}+[2,0]^\top} - y_{\boldsymbol{\alpha}+\boldsymbol{\beta}+[0,2]^\top}, \tag{6.23}$$

must satisfy $L_{r-1}(g_c) \succeq 0$ for $\mathbf{y}_{2,d}$ to be a valid moment sequence of the spectral measure of a matrix with spectral radius ρ (see Corollary 1).

In what follows, we propose to find a lower bound on the spectral radius of A by solving a semidefinite program aiming to minimize the value of the parameter ρ in (6.19)–(6.23) while satisfying all the constraints described above. Subsequently, the solution to this semidefinite program renders a lower bound on the spectral radius of A, denoted by λ_n, as shown in the following theorem.

Theorem 23. *Let r be an arbitrary positive integer and $d = 2r + 1$. Denote by $\underline{\rho}_r^\star$ the solution of the following semidefinite program:*

$$\begin{aligned} &\underset{\rho, \mathbf{y}_{2,d}}{\text{minimize}}\ \rho \\ &\text{subject to}\ \ (6.15)\text{–}(6.17), \\ &\quad M_r \succeq 0, \\ &\quad L_r(g_i) \succeq 0,\ \textit{for all } i \in [4], \end{aligned} \tag{6.24}$$

where M_r and $L_r(g_i)$ are defined in (6.18)–(6.22)*. Then, $\underline{\rho}_r^\star \leq \lambda_n$ for all $r \in \mathbb{N}$. Furthermore, $\underline{\rho}_r^\star$ is a non-decreasing function of $r \in \mathbb{N}$.*

Proof. See Appendix A.5. □

Theorem 23 allows us to compute a family of lower bound, parameterized by r, on the spectral radius of a digraph from counts of subgraphs up to order $d = 2r + 1$. In what follows, we provide a similar result to obtain a family of upper bounds on the spectral radius of A.

6.2.5 Upper Bounds using the K-moment Problem

From Perron-Frobenius theory [168], we know that the spectral radius of A is equal to the largest (nonnegative) real eigenvalue of A, denoted by λ_n. Hence, the set of eigenvalues $\lambda_1, \ldots, \lambda_{n-1}$ must be contained inside a circle of radius λ_n, denoted by S_{λ_n}. In other words, if we define an auxiliary atomic density with $n-1$ atoms located on the positions of the eigenvalues $\lambda_1, \ldots, \lambda_{n-1}$, the multi-sequence of moments of this auxiliary density must be S_{λ_n}-feasible. Furthermore, we can consider a circle of radius ρ, denoted by S_ρ, and find the maximum value of ρ for which the multi-sequence of moments of the auxiliary density is S_ρ-feasible. This optimal value of ρ will provide us with an upper bound on the spectral radius λ_n. In what follows, we elaborate on the details behind this approach.

We start our derivation with the following observation:

$$\sum_{s=2}^{k} \sum_{\widehat{G} \in \Omega_s} \eta(\widehat{G}, k)\, \texttt{Count}(\widehat{G}, G) = \lambda_n^k + \sum_{i=1}^{n-1} \lambda_i^k, \tag{6.25}$$

for all $k \in \mathbb{N}$. Let us introduce the following auxiliary atomic measure

$$\tilde{\mu}_A(x, y) = \frac{1}{n-1} \sum_{i=1}^{n-1} \delta(x - \sigma_i)\delta(y - \omega_i). \tag{6.26}$$

We denote by $\tilde{m}_{\boldsymbol{\alpha}}$ the $\boldsymbol{\alpha}$-moment of $\tilde{\mu}_A$. In what follows, we use the theory behind the K-

moment problem to derive necessary conditions that must be satisfied for all K-feasible multi-sequences, $\mathbf{z}_{2,d} = \{z_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{N}_d^2}$, for particular choices of K. In our derivations, we make use of the following lemma:

Lemma 14. *Given a directed graph G with adjacency matrix A, it holds that*

$$\operatorname{Tr}(A^k) = \lambda_n^k + (n-1)\sum_{s=0}^{\lfloor k/2\rfloor}\binom{k}{2s}(-1)^s\,\tilde{m}_{k-2s,2s},\ \textit{for all } k\in\mathbb{N}. \tag{6.27}$$

Proof. From (6.26), the $\boldsymbol{\alpha}$-moment of $\tilde{\mu}_A$ for $\boldsymbol{\alpha} = [a,b]^\top$ equals

$$\begin{aligned}\tilde{m}_{ab} &= \frac{1}{n-1}\int_{\mathbb{R}}\int_{\mathbb{R}} x^a y^b \sum_{i=1}^{n-1}\delta\left(x-\sigma_i\right)\delta\left(y-\omega_i\right)dxdy\\ &= \frac{1}{n-1}\sum_{i=1}^{n-1}\sigma_i^a\omega_i^b.\end{aligned} \tag{6.28}$$

From the proof of Lemma 3, we have that $m_{ab}(A) = \frac{1}{n}\sum_{i=1}^n \sigma_i^a\omega_i^b$. Combining this with (6.28), we have that

$$\tilde{m}_{ab} = \begin{cases}\dfrac{n}{n-1}m_{ab}, & \text{if } b>0,\\ \dfrac{nm_{ab}-\sigma_n^a}{n-1}, & \text{if } b=0.\end{cases} \tag{6.29}$$

Leveraging the connection between $m_{ab}(A)$ and $\operatorname{Tr}(A^k)$ (see (6.6)), we have

$$\begin{aligned}\operatorname{Tr}(A^k) &= n\sum_{s=0}^{\lfloor k/2\rfloor}\binom{k}{2s}(-1)^s\,m_{k-2s,2s}\left(A\right)\\ &= (n-1)\sum_{s=1}^{\lfloor k/2\rfloor}\binom{k}{2s}(-1)^s\,\tilde{m}_{k-2s,2s} + (n-1)\tilde{m}_{k,0} + \sigma_n^k\\ &= (n-1)\sum_{s=0}^{\lfloor k/2\rfloor}\binom{k}{2s}(-1)^s\,\tilde{m}_{k-2s,2s} + \sigma_n^k.\end{aligned}$$

Furthermore, according to Perron-Frobenius theory, we have that $\lambda_n = \sigma_n$. Thus, we obtain that

$$\operatorname{Tr}(A^k) = (n-1)\sum_{s=0}^{\lfloor k/2\rfloor}\binom{k}{2s}(-1)^s\,\tilde{m}_{k-2s,2s} + \lambda_n^k, \tag{6.30}$$

for all $k\in\mathbb{N}$. □

If $\mathbf{z}_{2,d}$ is the moment multi-sequence for $\tilde{\mu}_A$, then $z_{\boldsymbol{\alpha}} = \mathbb{E}_{\tilde{\mu}_A}[\mathbf{x}^{\boldsymbol{\alpha}}]$, for all $\boldsymbol{\alpha} \in \mathbb{N}_d^2$ (see Definition 17). Furthermore, according to Lemma 14 and Theorem 21, the entries of the sequence $\mathbf{z}_{2,d}$ must satisfy the following linear constraint:

$$\sum_{s=2}^{k} \sum_{\widehat{G} \in \Omega_s} \eta(\widehat{G}, k)\, \texttt{Count}(\widehat{G}, G) = (n-1) \sum_{s=0}^{\lfloor k/2 \rfloor} \binom{k}{2s} (-1)^s\, z_{k-2s,2s} + \rho^k, \tag{6.31}$$

for $k \in [d]$. Moreover, similar to (6.16) and (6.17), we also have that

$$z_{ab} = 0 \text{ for } b \text{ odd}, \tag{6.32}$$

$$z_{ab} \geq 0 \text{ for } a \text{ and } b \text{ even}. \tag{6.33}$$

Notice that the support of $\tilde{\mu}_A(x, y)$ is contained in the square $S = [-\lambda_n, \lambda_n]^2$. Thus, the moment and localizing matrices corresponding to $\mathbf{z}_{2,d}$ have the same form as those in (6.18)–(6.22) after substituting $y_{\boldsymbol{\alpha}}$ by $z_{\boldsymbol{\alpha}}$. As a result, we obtain the following moment and localizing matrices:

$$[\tilde{M_r}]_{\boldsymbol{\alpha\beta}} = z_{\boldsymbol{\alpha}+\boldsymbol{\beta}}, \tag{6.34}$$

$$[\tilde{L}_r(g_1)]_{\boldsymbol{\alpha\beta}} = \rho z_{\boldsymbol{\alpha}+\boldsymbol{\beta}} - z_{\boldsymbol{\alpha}+\boldsymbol{\beta}+[1,0]^\top}, \tag{6.35}$$

$$[\tilde{L}_r(g_2)]_{\boldsymbol{\alpha\beta}} = \rho z_{\boldsymbol{\alpha}+\boldsymbol{\beta}} + z_{\boldsymbol{\alpha}+\boldsymbol{\beta}+[1,0]^\top}, \tag{6.36}$$

$$[\tilde{L}_r(g_3)]_{\boldsymbol{\alpha\beta}} = \rho z_{\boldsymbol{\alpha}+\boldsymbol{\beta}} - z_{\boldsymbol{\alpha}+\boldsymbol{\beta}+[0,1]^\top}, \tag{6.37}$$

$$[\tilde{L}_r(g_4)]_{\boldsymbol{\alpha\beta}} = \rho z_{\boldsymbol{\alpha}+\boldsymbol{\beta}} + z_{\boldsymbol{\alpha}+\boldsymbol{\beta}+[0,1]^\top}, \tag{6.38}$$

for $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{N}_r^2$. As required by Corollary 3, the moment matrix (6.34) and localizing matrices (6.35)–(6.38) must be positive semidefinite. As a result, for $\rho = \lambda_n$, the moment sequence $\mathbf{z}_{2,d}$ of the auxiliary spectral measure $\tilde{\mu}_A$ must satisfy (6.31)–(6.33) and the moment and localizing matrices in (6.34)–(6.38) must be positive semidefinite.

In what follows, we find an upper bound on the spectral radius by solving a semidefinite program whose objective is to maximize the value of the parameter ρ in (6.31)–(6.38), while satisfying all the aforementioned constraints, as described in the following theorem.

Theorem 24. *Let r be an arbitrary positive integer and $d = 2r + 1$. Denote by $\overline{\rho}_r^\star$ the solution of the following semidefinite program:*

$$\begin{aligned} &\underset{\rho, \mathbf{z}_{2,d}}{\text{maximize}}\ \rho \\ &\text{subject to (6.31)–(6.33)}, \\ &\tilde{M}_r \succeq 0, \\ &\tilde{L}_r(g_i) \succeq 0, \ \textit{for all } i \in [4], \end{aligned} \tag{6.39}$$

where $\tilde{M}_r$ and $\tilde{L}_r(g_i)$ are defined in (6.34)–(6.38). Then, $\overline{\rho}_r^\star \geq \lambda_n$ for all $r \in \mathbb{N}$. Furthermore, $\overline{\rho}_r^\star$ is a non-increasing function of $r \in \mathbb{N}$.

Proof. See Appendix A.5. □

Using Theorem 23 and Theorem 24, we can compute lower and upper bounds on the spectral radius of a directed graph using counts of subgraphs in G. Furthermore, these bounds become tighter as the order of subgraphs under consideration increases.

6.2.6 Illustration and Discussion

To demonstrate the performance of these bounds, we apply our methodology to a directed graph modeling the connections between $n = 1,574$ different airports within the United States [171]. Assuming we are able to count the number of all subgraphs of order up to 6, the upper bound on the spectral radius obtained via Theorem 24 equals $\overline{\rho}_3^\star = 99.2906$, whereas the actual spectral radius equals $\lambda_n = 99.1183$. However, when we only have access to the counts of subgraphs of small order, our approach can lead to loose bounds. For example, considering a realization of the Erdős-Rényi random directed graph with $n = 100$ vertices and $\mathbb{P}((i, j) \in \mathcal{E}) = 0.15$ for all $i, j \in \mathcal{V}$, we obtain a spectral radius of $\lambda_n = 14.5431$. In this case, when the counts of subgraphs of order up to 4 are available, the lower bound obtained using Theorem 23 is $\underline{\rho}_2^\star = 5.5$. This bound is loose for the following two reasons: First, although Theorem 23 and Theorem 24

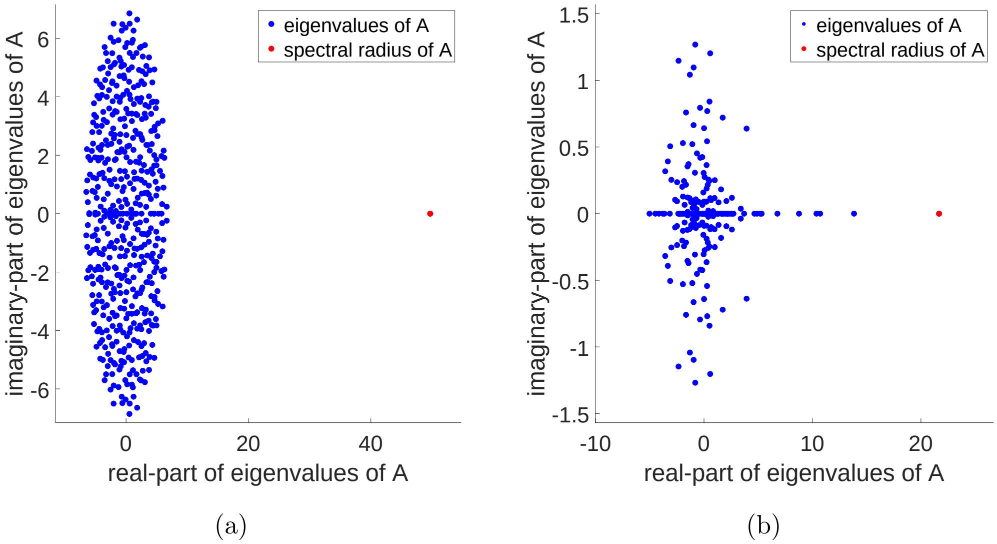


Figure 6-2: In (a), we plot the complex eigenvalues of A for an Erdős-Rényi random directed graph with $n = 500$ vertices and edge probability 0.1.The spectral radius of A is $\lambda_n \approx 50$, whereas $\omega_{\max} < 7$. In (b), we plot the complex eigenvalues of A for a real social network from Google+ [171]. The spectral radius of $\lambda_n \approx 21$, whereas $\omega_{\max} < 1.5$.

provide lower and upper bounds on the spectral radius, the moments of the optimal solutions may not correspond to an n-atomic measure, since Corollary 3 does not provide a sufficient condition to guarantee the existence of an n-atomic measure. Secondly, and more importantly, we have assumed that $\texttt{spec}(A)$ is contained in the square $[-\lambda_n, \lambda_n]^2$. However, the support of μ_A is contained in $[-\lambda_n, \lambda_n] \times [-\omega_{\max}, \omega_{\max}]$, where $\omega_{\max}$ can be much smaller than λ_n in some real digraphs, leading to loose bounds (see Figure 6-2). In the following Section, we propose a refinement of our technique in order to overcome this issue by finding better bounds on ω_{max}.

6.3 Refined Moment-based Bounds

In this section, we introduce a refined moment-based framework to improve the quality of our bounds on the spectral radius. The main idea behind this approach is to obtain an upper bound on $\omega_{\max}$. To achieve this goal, we will study the spectral measure of the matrix $A - A^\top$. As we discuss below, the largest imaginary part among the eigenvalues of $A - A^\top$ upper-bounds ω_{max} of A. We then relate the spectral moments of $A - A^\top$ to the counts of certain subgraphs in G. Finally, we will resort to the K-moment problem to provide an upper bound on $\omega_{\max}$. This upper bound will be further used to provide refined upper and lower bounds on the spectral radius of A.

In order to provide an upper bound on $\omega_{\max}$ of A, we first present a connection between the eigenvalues of the (imaginary) matrix $A_I = j(A - A^\top)$ and those of A. Notice that the matrix $A - A^\top$ is skew-symmetric; hence, its eigenvalues are a collection of purely imaginary conjugate pairs. Hence, the spectrum of A_I is purely real and symmetric around the imaginary axis. From [94], we have that

$$\omega_{\max} \leq \frac{1}{2}\max\{\mathbf{v}^* A_I \mathbf{v} : \mathbf{v}^*\mathbf{v} = 1, \mathbf{v} \in \mathbb{C}^n\} = \lambda_n(A_I),$$

where $\lambda_n(A_I)$ is the largest (real) eigenvalue of A_I. In particular, the equality holds if and only if A is normal. Using this relationship, we will provide an upper bound on $\omega_{\max}$ using traces of powers of A_I. In what follows, we show a linear relationship between counts of certain subgraphs in G and $\mathrm{Tr}(A_I^\ell)$.

6.3.1 From Open-walks in G to Traces of Powers of A_I

Hereafter, we show that $\mathrm{Tr}(A_I^\ell)$ can be computed by a linear combination of the counts of specific subgraphs in G. To show this, we first provide a closed-form expression of the term $\mathrm{Tr}(A_I^\ell)$ using entries of A_I. On the one hand, since the spectrum of A_I is symmetric around the imaginary axis, we have that $\mathrm{Tr}(A_I^\ell) = 0$ for ℓ odd. On the other hand, when ℓ is an even number, we have that

$$\begin{aligned}
\mathrm{Tr}(A_I^\ell) &= \mathrm{Tr}(j^\ell (A - A^\top)^\ell), \\
&= (-1)^{\frac{\ell}{2}} \mathrm{Tr}\left((A - A^\top)^\ell\right), \\
&= (-1)^{\frac{\ell}{2}} \sum_{\substack{c_i, d_i \in \{0,1\} \\ c_i + d_i = 1}} (-1)^{\sum_{i=1}^{\ell} d_i} \mathrm{Tr}\left[A^{c_1}(A^\top)^{d_1} \cdots A^{c_\ell}(A^\top)^{d_\ell}\right].
\end{aligned} \tag{6.40}$$

Therefore, $\mathrm{Tr}(A_I^\ell)$ is equal to the sum of 2^ℓ terms. Using similar ideas than those used in the proof of Theorem 21, one can show that $\mathrm{Tr}\left[A^{c_1}(A^\top)^{d_1} \cdots A^{c_\ell}(A^\top)^{d_\ell}\right]$ is equal to

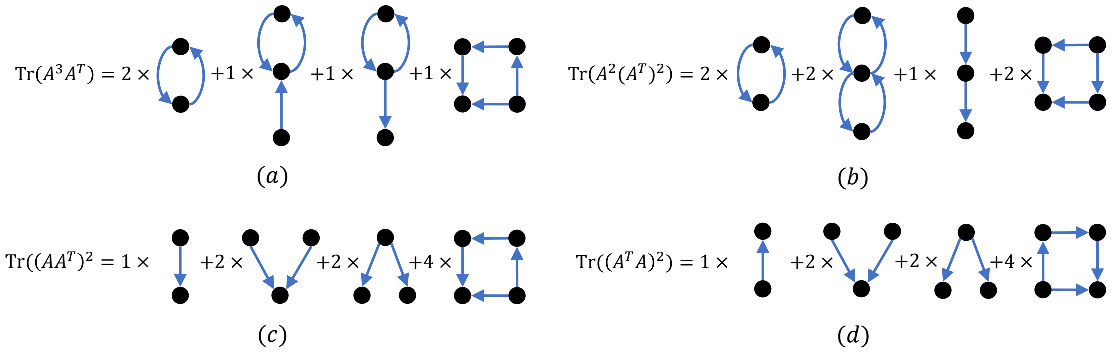


Figure 6-3: This figure shows the relationship between $\mathrm{Tr}(A_I^4)$ and the counts of certain subgraphs in G. More specifically, the traces in the figure are equal to sums of counts of certain subgraphs multiplied by the coefficients indicated in the figure. In (c), to calculate $\mathrm{Tr}((AA^\top)^2)$, we use the sum of in-degrees. In (d), to calculate $\mathrm{Tr}((A^\top A)^2)$, we use the sum of out-degrees. Since $\mathrm{Tr}((AA^\top)^2) = \mathrm{Tr}((A^\top A)^2)$, we can use either sum of in-degrees or out-degrees to obtain $\mathrm{Tr}((AA^\top)^2)$.

a linear combination of the counts of certain subgraphs in G. We illustrate this idea by considering the following examples.

Example 3. *When $\ell = 2$, we have that*

$$\begin{aligned}\mathrm{Tr}(A_I^2) &= -\mathrm{Tr}(A - A^\top)^2 \\ &= -\mathrm{Tr}(A^2 - AA^\top - A^\top A + (A^\top)^2) \\ &= -\mathrm{Tr}(A^2) + 2\mathrm{Tr}(AA^\top) - \mathrm{Tr}(A^\top)^2 \\ &= -2\mathrm{Tr}(A^2) + 2\mathrm{Tr}(AA^\top).\end{aligned} \tag{6.41}$$

In this particular case, we notice that Tr$(A^2) = \sum_{i,j:(i,j),(j,i)\in\mathcal{E}} 1$ (see (6.1)*) and Tr$(AA^\top) = \sum_{i,j:(j,i)\in\mathcal{E}} 1$. The latter term equals the sum of in-degrees of each vertex i in G. Consequently, Tr(A_I^2) equals twice the total number of edges minus twice the counts of bidirected-edge subgraphs in G.*

Let us consider an additional example when $\ell = 4$.

Example 4. *When $\ell = 4$, we have that*

$$\mathrm{Tr}(A_I^4) = \mathrm{Tr}((A - A^\top)^2(A - A^\top)^2). \tag{6.42}$$

Using the properties of matrix trace operations, the above term is simplified to

$$\mathrm{Tr}(A_I^4) = 2\mathrm{Tr}(A^4) - 8\mathrm{Tr}(A^3 A^\top) + 4\mathrm{Tr}(A^2(A^\top)^2) + 2\mathrm{Tr}((AA^\top)^2). \tag{6.43}$$

In what follows, we show that Tr$(A^3A^\top)$*, Tr*$(A^2(A^\top)^2)$ *and Tr*$((AA^\top)^2)$ *can all be calculated using the counts of certain subgraphs in* G*. We characterize those relationships in Figure 6-3. For demonstration purposes, we only show below the relationship between the term Tr*$((AA^\top)^2)$ *and subgraph counts. Others traces can be computed in a similar fashion.*

First, notice that

$$Tr((AA^\top)^2) = \sum_i \sum_{j,k,l} [A]_{ij}[A]_{kj}[A]_{kl}[A]_{il}, \tag{6.44}$$

where the inside term is non-zero if and only if $(j,i),(j,k),(l,k),(l,i)$ *are all edges of the digraph* G*. Next, we consider whether some of the indices* $i,j,k,l \in [n]$ *are equal. Since* G *is simple,* $A_{ii} = 0$ *for* $i \in [n]$*. Therefore, to ensure that* $[A]_{ij}[A]_{kj}[A]_{kl}[A]_{il}$ *is non-zero, it suffices to consider the following cases: (i)* i,j,k,l *are all distinct, (ii)* $i = k$ *while* $j \neq l$*, (iii)* $i = k$ *while* $j = l$*, and (iv)* $i \neq k$ *while* $j = l$*. In case (*i*), where* i,j,k,l *are all distinct, the ordered edges* $(j,i),(j,k),(l,k),(l,i)$ *together correspond to a subgraph of size four. However, notice that by permuting the indices, we obtain* $(l,i),(l,k),(k,j),(j,i)$*, which corresponds to the same subgraph (analogously for* $(l,k),(l,i),(j,i),(j,k)$ *and* $(j,k),(j,i),(l,i),(l,k)$*). As a result, the same subgraph is counted four times in the summation* (6.44)*. Thus, the coefficient corresponding to this subgraph is* 4 *when calculating Tr*$((AA^\top)^2)$*. In case (*ii*), the constraints given above induce a subgraph formed by the following ordered sequence of edges:* $\{(j,i),(l,i)\}$*, while* $\{(l,i),(j,i)\}$ *is another sequence representing the same subgraph. Thus, such a subgraph contributes twice in* (6.44) *and the associated coefficient is* 2*. In case (*iii*),* $i = k$ *and* $j = l$*, therefore* $[A]_{ij}[A]_{kj}[A]_{kl}[A]_{il} = [A]_{ij}$*, since* A *is unweighted. Hence, summing over all* i *in case (*iii*) corresponds to the sum of in-degrees of each vertex. Case (*iv*) is identical to case (*ii*).*

Remark 14. *Instead of* $\mathrm{Tr}((AA^\top)^2)$*, we may use* $\mathrm{Tr}((A^\top A)^2)$ *in* (6.43) *to obtain*

$\mathrm{Tr}(A_I^4)$. *In this case, the subgraphs under consideration are listed in Figure 6-3-(d). This observation can be generalized to all the terms involving traces of products of* A *and* $A^\top$ *in* (6.40).

Remark 15. *In general, finding a closed-form expression for the coefficients for the subgraphs using* (6.40) *is difficult. Moreover, to obtain* $\mathrm{Tr}(A_I^r)$, *one has to derive the counts for all subgraphs of size* r, *which can be computationally challenging when* r *is large. However, in most real networks, we obtain a tight approximation of the spectral radius by considering* $r \leq 6$, *as we will show empirically in Section 6.4.*

Next, we propose a method to upper bound the spectral radius of A_I using the K-moment problem.

6.3.2 Estimation of $\omega_{\max}(A)$

To upper bound the spectral radius of A_I, we follow a similar procedure as the one discussed in the previous section. Given A_I, we define the spectral measure of A_I as the following one-dimensional probability density:

$$\nu_{A_I}(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - \lambda_i(A_I)). \tag{6.45}$$

Since $\lambda_i(A_I) \in \mathbb{R}$, the measure ν_{A_I} is supported on $\mathbb{R}$. Without loss of generality, we can order the eigenvalues of A_I by $\lambda_1(A_I) \leq \cdots \leq \lambda_n(A_I)$. Since $A - A^\top$ is skew-symmetric, we have that $\lambda_1(A_I) = -\lambda_n(A_I)$. The support of ν_{A_I} must satisfy $\mathrm{Supp}(\nu(A_I)) \subseteq [-\lambda_n(A_I), \lambda_n(A_I)]$.

In addition to ν_{A_I}, we define the auxiliary spectral measure $\tilde{\nu}_{A_I}$ by

$$\tilde{\nu}_{A_I}(x) = \frac{1}{n-2}\sum_{i=2}^{n-1} \delta(x - \lambda_i(A_I)), \tag{6.46}$$

which is an $(n-2)$-atomic measure defined by removing both $\lambda_1(A_I)$ and $\lambda_n(A_I)$ from $\texttt{spec}(A_I)$. Different from $\tilde{\mu}_A$, we remove two atoms from $\texttt{spec}(A_I)$ to maintain the symmetry (with respect to the origin) of the auxiliary measure. Consequently, the supports of both ν_{A_I} and $\tilde{\nu}_{A_I}$ are contained in $[-\lambda_n(A_I), \lambda_n(A_I)]$.

Following an idea similar to the one presented in the previous section, we show that the trace of A_I^ℓ is related to the moments of both ν_{A_I} and $\tilde{\nu}_{A_I}$. More specifically, given a positive integer $r \in \mathbb{N}$, we compute the r-th moment of ν_{A_I}, denoted by $m_r(A_I)$, as follows:

$$m_r(A_I) = \int_{x\in\mathbb{R}} x^k d\nu_{A_I} = \frac{1}{n}\sum_{i=1}^{n} \lambda_i(A_I)^r = \frac{1}{n}\mathrm{Tr}(A_I^r). \tag{6.47}$$

Similarly, the r-th moment of $\tilde{\nu}_{A_I}$, denoted by $\tilde{m}_r(A_I)$, is equal to

$$\begin{aligned}\tilde{m}_r(A_I) &= \int_{x\in\mathbb{R}} x^k d\tilde{\nu}_{A_I} \\ &= \frac{1}{n-2}\sum_{j=2}^{n-1} \lambda_i(A_I)^r \\ &= \frac{1}{n-2}\left[\mathrm{Tr}(A_I^r) - \left((-1)^r + 1\right)\lambda_n(A_I)^r\right] \\ &= \frac{1}{n-2}\left[n m_r(A_I) - \left((-1)^r + 1\right)\lambda_n(A_I)^r\right].\end{aligned} \tag{6.48}$$

To obtain an upper bound on $\lambda_n(A_I)$, we first find necessary conditions that must be satisfied by all moment sequences of $\tilde{\nu}_{A_I}$, denoted by $\mathbf{w}_{2r+1} = \{w_\gamma\}_{\gamma\leq 2r+1}$. Since the spectrum of A_I is symmetric around 0, it follows that all odd moments of ν_{A_I} and $\tilde{\nu}_{A_I}$ are 0. As a result, in order for $\mathbf{w}_{2r+1}$ to be a moment sequence with respect to $\tilde{\nu}_{A_I}$, we must have:

$$w_\gamma = \begin{cases} 1, & \text{if } \gamma = 1, \\ 0, & \text{if } \gamma > 1 \text{ and } \gamma \text{ is an odd number,} \\ \frac{1}{n-2}\left(\mathrm{Tr}(A_I^\gamma) - 2\lambda_n(A_I)^\gamma\right), & \text{otherwise,} \end{cases} \tag{6.49}$$

for all $\gamma \leq 2r + 1$.

Moreover, the moment and localizing matrices of $\mathbf{w}_{2r+1}$ must be positive semidefinite, as required by Theorem 22. More specifically, we let the moment matrix of $\mathbf{w}_{2r+1}$ be defined entry-wise by:

$$[M_r(\mathbf{w})]_{\alpha,\beta} = w_{\alpha+\beta}, \tag{6.50}$$

where $\alpha, \beta \in \mathbb{N}_r$. Let $h_1(x) = x - \lambda_n(A_I)$ and $h_2(x) = x + \lambda_n(A_I)$; hence, we have that $[-\lambda_n(A_I), \lambda_n(A_I)] = \{x \in \mathbb{R} \colon h_1(x) \geq 0, h_2(x) \geq 0\}$. Next, we define the localizing matrices with respect to h_1 and h_2 by

$$[L_r(h_1, \mathbf{w})]_{\alpha,\beta} = \lambda_n(A_I) w_{\alpha+\beta} - w_{\alpha+\beta+1}, \tag{6.51}$$

and

$$[L_r(h_2, \mathbf{w})]_{\alpha,\beta} = \lambda_n(A_I) w_{\alpha+\beta} + w_{\alpha+\beta+1}. \tag{6.52}$$

Since the support of $\tilde{\nu}_{A_I}$ is contained in $[-\lambda_n(A_I), \lambda_n(A_I)]$, both $L_r(h_1, \mathbf{w})$ and $L_r(h_{,}\mathbf{w})$ must be positive semidefinite.

Subsequently, for $\rho = \lambda_n(A_I)$, the moment sequence $\mathbf{w}_{2r+1} = \{w_\gamma\}_{\gamma \leq 2r+1}$ of the auxiliary spectral measure $\tilde{\nu}_{A_I}$ must satisfy (6.49). Furthermore, the moment and localizing matrices defined in (6.50)–(6.52) must be positive semidefinite (by replacing $\lambda_n(A_I)$ with the parameter ρ). Next, we aim to find the maximum value of the parameter ρ such that all the constraints above are satisfied.

Theorem 25. *Let A be the adjacency matrix of a digraph G, and define $A_I = j(A - A^\top)$. Let r be an arbitrary positive integer and $d = 2r + 1$. Denote by $\omega_r^\star$ the solution to the following semidefinite program:*

$$\begin{aligned} &\underset{\rho, \mathbf{w}_{2r+1}}{\text{maximize}}\ \rho \\ &\text{subject to (6.49)}, \\ &M_r(\mathbf{w}) \succeq 0, L_r(g_1, \mathbf{w}) \succeq 0, L_r(g_2, \mathbf{w}) \succeq 0, \end{aligned} \tag{6.53}$$

where $M_r(\mathbf{w})$ and $L_r(g_i, \mathbf{w})$ are defined in (6.50)–(6.52). *Then, $\frac{\omega_r^\star}{2} \geq \omega_{\max}$ for all $r \in \mathbb{N}$. Furthermore, $\omega_r^\star$ is a non-increasing function of $r \in \mathbb{N}$.*

Proof. See Appendix A.5. □

Note that, as described in Subsection 6.3.1, the values of $\operatorname{Tr}(A_I^\ell)$ in (6.49) can be computed using counts of subgraphs of G. Hence, we have that $w_r^\star$ can be obtained using counts of subgraphs solely, providing an upper bound on the maximum imaginary part in the spectrum of A.

Corollary 4. *Let A be the adjacency matrix of a digraph G. Given a positive integer $r \in \mathbb{N}$, let $w_r^\star$ be the optimal solution to* (6.53)*. If $A = A^\top$, then $w_r^\star = 0$ for all positive integers $r \in \mathbb{N}$.*

Proof. When $A = A^\top$, $\mathrm{Tr}(A_I^\gamma) = 0$ for all $\gamma \in \mathbb{N}$. Therefore, from (6.60), given an even integer $\gamma \in \mathbb{N}$, we have that $(n-2)w_\gamma = -2\rho^\gamma$ (by replacing $\lambda_n(A_I)$ with the optimization parameter ρ). Since $M_r(\mathbf{w})$ is positive semidefinite, all its diagonal entries are non-negative. As a consequence, ρ must equal to zero. Therefore, $w_r^\star = 0$. □

The above corollary shows that the upper bound on $\omega_{\max}(A)$ is tight when A is a symmetric matrix. Consequently, the refined framework can also be used to obtain tight bounds for undirected graphs.

6.3.3 Refined Bounds on the Spectral Radius

In Section III, we have considered that the spectrum of A is contained in the square $S = [-\lambda_n, \lambda_n]^2$. However, more precisely, $\texttt{spec}(A)$ is contained in a rectangle $\hat{S} = [-\lambda_n, \lambda_n] \times [-\omega_{\max}, \omega_{\max}]$. Consequently, we define the polynomials $\hat{g}_3(\mathbf{x}) = \omega_{\max} - x_2$ and $\hat{g}_4(\mathbf{x}) = \omega_{\max} + x_2$. As required by Corollary 3, the localizing matrices of $\mathbf{y}_{2,d}$ with respect to $\hat{g}_3$ and $\hat{g}_4$ must be positive semidefinite. In other words, we impose additional constraints on the feasible sets in the optimization problems (6.24) and (6.39). This procedure is summarized in Algorithm 8.

Consequently, we have utilized counts of different subgraphs to provide upper and lower bounds on the spectral radius. In general, $\omega_{\max}(A)$ is much smaller than $\rho(A)$. Thus, the obtained solution from Algorithm 1 achieves better performance than the approach in Section 6.2. Notice that, not all subgraphs are needed to compute $\mathrm{Tr}(A_I^\ell)$ and $\mathrm{Tr}(A^\ell)$. For example, when we consider using subgraphs of order less or equal to 5, we only need the counts of those subgraphs depicted in Figure 6-4.

6.3.4 Upper Bound on λ_n

We have proposed a framework to provide a more accurate estimate on the spectral radius of A using the relation between the imaginary parts of A and $A - A^\top$. Similarly, we can also consider how the eigenvalues of $A_R = A + A^\top$ are related to the eigenvalues of A. More specifically, the relationship

$$\lambda_n \leq \frac{1}{2} \max\{\mathbf{v}^* A_R \mathbf{v} \colon \mathbf{v}^*\mathbf{v} = 1, \mathbf{v} \in \mathbb{C}^n\} = \lambda_n(A_R). \tag{6.56}$$

where $\lambda_n(A_R)$ is the largest eigenvalue of A_R. In particular, the equality holds if and only if A is a normal matrix. As shown previously, an upper bound on $\lambda_n(A_R)$ can be obtained by relating the counts of a subsets of subgraphs in A to the spectral moments of $A+A^\top$. Therefore, this bound is also an upper bound for λ_n due to (6.56). Subsequently, we can also provide an upper bound on λ_n by providing an upper bound on $\lambda_n(A_R)$ using counts of subgraphs in G.

To characterize the upper bound on the spectral radius of A_R, we follow the idea presented in the previous subsection. Given A_R, we define the spectral measure of A_R as the following one-dimensional probability density:

$$\nu_{A_R}(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - \lambda_i(A_R)). \tag{6.57}$$

We also define the $(n-1)$-atomic auxiliary spectral measure $\widetilde{\nu}_{A_R}$ by

$$\tilde{\nu}_{A_R}(x) = \frac{1}{n-1} \sum_{i=1}^{n-1} \delta(x - \lambda_i(A_R)). \tag{6.58}$$

From the definitions of ν_{A_R} and $\widetilde{\nu}_{A_R}$, we compute the r-th moment of $\widetilde{\nu}_{A_R}$, denoted by $\tilde{m}_r(A_R)$, as follows:

$$\begin{aligned} \tilde{m}_r(A_R) &= \int_{x \in \mathbb{R}} x^k d\tilde{\nu}_{A_R}, \\ &= \frac{1}{n-1} \left[\mathrm{Tr}(A_R^r) - \lambda_n(A_R)^r \right]. \end{aligned} \tag{6.59}$$

As illustrated in (6.44), as well as Theorem 2, $\mathrm{Tr}(A_R^r)$ can be computed using counts of certain subgraphs in G. As a consequence, $\tilde{m}_r(A_R)$ can also be computed as a linear combination of the counts of certain subgraphs in G. To find an upper bound on $\lambda_n(A_R)$, we provide below necessary conditions that must be satisfied by all moment sequences of $\tilde{\nu}_{A_R}$, denoted by $\mathbf{p}_{2r+1}$.

According to (6.59), in order for $\mathbf{p}_{2r+1}$ to be a potential moment sequence of the density $\tilde{\nu}_{A_R}$, we must have:

$$p_\gamma = \frac{1}{n-1}\left[\operatorname{Tr}(A_R^\gamma) - \lambda_n(A_R)^\gamma\right], \tag{6.60}$$

for all $\gamma \leq 2r+1$. Moreover, the moment matrix of $\mathbf{p}_{2r+1}$, defined entry-wise by

$$[M_r(\mathbf{p})]_{\alpha,\beta} = p_{\alpha+\beta}, \tag{6.61}$$

for $\alpha, \beta \in \mathbb{N}_r$, must be positive semidefinite. Since $A \in \{0,1\}^{n\times n}$, the matrix $A_R = A + A^\top$ is entry-wise non-negative. It further follows that the largest eigenvalue of A_R is non-negative, according to Perron-Frobenius theory. Subsequently, we have that $\texttt{spec}(A_R) \subseteq [-\lambda_n(A_R), \lambda_n(A_R)]$. Let us define the polynomials $\phi_1(x) = \lambda_n(A_R) - x$ and $\phi_2(x) = \lambda_n(A_R) + x$; hence, we have that $\texttt{spec}(A_R) \subseteq \{x \in \mathbb{R} \colon \phi_1(x) \geq 0, \phi_2(x) \geq 0\}$. Next, we define the localizing matrices with respect to ϕ_1 and ϕ_2 as

$$[L_r(\phi_1, \mathbf{p})]_{\alpha,\beta} = \lambda_n(A_R)p_{\alpha+\beta} - p_{\alpha+\beta+1}, \tag{6.62}$$

$$[L_r(\phi_2, \mathbf{p})]_{\alpha,\beta} = \lambda_n(A_R)p_{\alpha+\beta} + p_{\alpha+\beta+1}, \tag{6.63}$$

for $\alpha, \beta \in \mathbb{N}_r$. Then, Corollary 3 indicates that $L_r(\phi_1, \mathbf{p})$ and $L_r(\phi_2, \mathbf{p})$ must be positive semidefinite for the sequence $\mathbf{p}_{2r+1}$ to be a potential moment sequence of the density $\tilde{\nu}_{A_R}$.

Consequently, for $\rho = \lambda_n(A_R)$, the moment sequence $\mathbf{p}_{2r+1}$ of the auxiliary measure $\tilde{\nu}_{A_R}$ must satisfy the above constraints. The upper bound on $\lambda_n(A_R)$ can thus be found by maximizing the parameter ρ subjected to the above constraints, as shown in the following theorem.

Theorem 26. *Let A be the adjacency matrix of a digraph G, and define $A_R = A + A^\top$.*

Let r be an arbitrary positive integer and $d = 2r + 1$. Denote by $p_r^\star$ the optimal solution to the following semidefinite program:

$$\begin{aligned}
&\underset{\rho, \mathbf{p}_d}{\text{maximize}}\ \rho \\
&\text{subject to (6.60)}, \\
&M_r(\mathbf{p}) \succeq 0, \\
&L_r(\phi_1, \mathbf{p}) \succeq 0, L_r(\phi_2, \mathbf{p}) \succeq 0,
\end{aligned} \tag{6.64}$$

where $M_r(\mathbf{p})$, $L_r(\phi_1, \mathbf{p})$ and $L_r(\phi_2, \mathbf{p})$ are defined in (6.61)–(6.63)*. Then, $\frac{p_r^\star}{2} \geq \lambda_n$ for all $r \in \mathbb{N}$. Furthermore, $p_r^\star$ is a non-increasing function of $r \in \mathbb{N}$.*

Proof. See Appendix A.5. □

Since $\mathrm{Tr}(A_R^\ell)$ can be computed using counts of subgraphs of G, we have that $p_r^\star$ can be obtained via counts of subgraphs of A.

6.4 Empirical Results

In this section, we empirically demonstrate the validity of our bounds on random digraphs (Subsection 6.4.1) and on real networks (Subsection 6.4.2).

6.4.1 Random Directed Graphs

We generate random directed graphs according to the directed version of the Chung-Lu model [172]. More specifically, given a positive integer n, we consider two sequences $w_{in} = [w_1^{in}, w_2^{in}, \ldots, w_n^{in}]^\top$ and $w_{out} = [w_1^{out}, w_2^{out}, \ldots, w_n^{out}]^\top$, representing the in-degrees and out-degrees of each vertex. Furthermore, we let $\sum_{i=1}^n w_i^{in} = \sum_{i=1}^n w_i^{out} = m$. Then, according to [172], the entries in A are given by

$$A_{ij} = \begin{cases} 1, & \text{w.p. } \frac{w_i^{in} w_j^{out}}{m}, \\ 0, & \text{otherwise.} \end{cases} \tag{6.65}$$

Using this model, we can also generate Erdős-Renyi random digraphs by letting $w_{in} = w_{out} = [pn, \ldots, pn]^\top$ for a prescribed value $p \in (0,1)$. As an example, we consider the following parameters in our experiment: $n = 500$ and $p = \frac{\log n}{n} \approx 0.0124$. Using these parameters, we generate a numerical realization of the random digraph A. The spectral radius of A equals $\lambda_n \approx 6.3002$, whereas the $\omega_{\max} \approx 2.6373$. From Theorem 7, when $r = 2$, we have that $\rho_r^\star = 6.7806$.

In addition to Erdős-Renyi random digraph, we can specify w_{in} and w_{out} to generate random graphs power-law degree distributions. As shown in [173], given $c, \beta, i_0 \in \mathbb{R}$, we can define the sequence $w_i = c(i_0 + i)^{-\frac{1}{\beta-1}}$, to generate a random undirected graph whose degrees follow a power-law distribution with exponent β, i.e., the number of vertices with degree k is proportional to $k^{-\beta}$. In particular, it is possible to 'control' the maximum degree, denoted by Δ, and average degree, denoted by d, by using the following parameter selection:

$$c = \frac{\beta - 2}{\beta - 1} d n^{\frac{1}{\beta-1}}, \text{ and } i_0 = n \left(\frac{d(\beta - 2)}{\Delta(\beta - 2)} \right)^{\beta - 1}. \tag{6.66}$$

In our experiment, we generate a sequence w using the above method and let $w_{in} = w_{out} = w$. In addition, we consider the following parameters: $n = 1500$, $\beta = 5$, $d = 40$ and $\Delta = 120$. Subsequently, from a random digraph realization, we have that $\lambda_n \approx 42.8770$, while $\omega_{\max} \approx 6.3868$. In Figure 6-5-(a), we show the histogram of in-degrees for the particular random digraph realization under consideration, while in Figure 6-5-(b), we show the evolution of the upper and lower bounds proposed in this paper as the order of the subgraph counts used increases. For example, the outputs of Algorithm 8 using counts of subgraphs of order up to 6 are an upper bound of 42.8777 and a lower bound of 42.8763, which are very tight in this case. Next, we explore our framework on real artificial directed graphs.

6.4.2 Real-world Directed Graphs

We consider several real digraphs obtained from [171] and [174]. In our first example, we examine the directed graph representing flights between U.S. airports in 2010 containing 1,574 vertices and 28,236 edges. In this digraph, each directed edge represents a flight connection from one airport to another. In our experiment, we preserve connectivity of the digraph and remove the edge weights. The spectral radius of the resulting (unweighted) digraph equals $\lambda_n = 99.1175$, whereas $\omega_{\max} = 2.881$. We plot the eigenvalue spectrum of A in Figure 6-6. Using the moment framework described in Section 3, we obtain that our unrefined bounds are $\underline{\rho}_2^\star = 47.1184$ and $\overline{\rho}_2^\star = 172.2931$, when the counts of subgraphs of order up to 5 are considered. To improve these bounds, we first find an upper bound on $\omega_{\max}$. Using Algorithm 8, for $r = 2$, we obtained that $\omega_r^\star/2 \approx 8.2776$, which is an upper bound on $\omega_{\max}(A) = 2.881$. With the help on this additional information, we obtained that the refined lower bound and upper bound on the spectral radius equal 99.1167 and 102.9278, respectively.

In Tables 6.1 and 6.2, we illustrate the performance of our framework using other real-world directed graphs. In these experiments, we fix $r = 3$ and compare the performance of our bounds, with and without the refinement described in Subsection 6.3.3. As previously indicated, the refined bounds are guaranteed to be no worse than the bounds obtained without estimating the largest imaginary part. Moreover, as r increases, the difference between the estimates using the two proposed methods diminishes, as illustrated in Table 6.1 and Table 6.2. However, the convergence rate of our algorithm depends on the structure of the digraph. For example, we observe that using $r = 3$, the lower bound returned by Algorithm 8 equals $\underline{\rho}_3^\star$ computed using Theorem 24 when we are considering the social network with $n = 627$ vertices.

Type	**Size**	λ_n	$\omega_{\max}$	$\underline{\rho}_3^\star$	$\underline{\varrho}_3^\star$
Social	131	18.3488	1.2132	8.0349	8.4347
Social	168	21.8484	0.8023	9.6100	13.5492
Social	344	21.6719	1.26	10.7704	21.6712
Social	627	10.4766	1.2995	5.2389	5.2389
Airport	1574	99.1175	2.881	99.1167	99.1167
Wikipedia	8297	47.9430	8.4824	29.9651	29.9651

Table 6.1: This table shows lower bound on the spectral radius of various networks computed using Theorem 23 (fifth column) and Algorithm 8 (last column).

Type	**Size**	$\overline{\rho}_3^\star$	$p_3^\star$	$\overline{\varrho}_3^\star$
Social	131	22.2728	22.5450	20.7786
Social	168	35.9181	22.5630	24.9591
Social	344	24.7768	29.6324	24.7768
Social	627	12.8224	18.9572	12.3289
Airport	1574	99.1183	99.2906	99.1183
Wikipedia	8297	50.3321	49.0404	47.9438

Table 6.2: This table shows upper bounds computed using Theorem 24 (column 2, denoted by $\overline{\rho}_3^\star$), Theorem 26 (column 3, denoted by $p_3^\star$), and Algorithm 8 (last column), respectively.

```
Algorithm 8: Refined upper and lower bounds of $\lambda_n$
Input: Positive integer $r \in \mathbb{N}$, and $\{\mathrm{Tr}(A_I^\ell), \mathrm{Tr}(A^\ell)\}_{l=1}^{2r+1}$
Output: Lower bound and upper bound on the spectral radius of $G$, denoted
    by $\underline{\varrho}_r^\star$ and $\overline{\varrho}_r^\star$, respectively.
Let $d = 2r + 1$.
Solve (6.53) and obtain $w_r^\star$.
Define matrices $L_r(\hat{g}_3)$ and $L_r(\hat{g}_4)$ entry-wise by:
    $[L_r(\hat{g}_3)]_{\boldsymbol{\alpha}\boldsymbol{\beta}} = w_r^\star y_{\boldsymbol{\alpha}+\boldsymbol{\beta}} - y_{\boldsymbol{\alpha}+\boldsymbol{\beta}+[0,1]^\top}$, and
    $[L_r(\hat{g}_4)]_{\boldsymbol{\alpha}\boldsymbol{\beta}} = w_r^\star y_{\boldsymbol{\alpha}+\boldsymbol{\beta}} + y_{\boldsymbol{\alpha}+\boldsymbol{\beta}+[0,1]^\top}$.
Define matrices $\widetilde{L}_r(\hat{g}_3)$ and $\widetilde{L}_r(\hat{g}_4)$ entry-wise by
    $[\widetilde{L}_r(\hat{g}_3)]_{\boldsymbol{\alpha}\boldsymbol{\beta}} = w_r^\star z_{\boldsymbol{\alpha}+\boldsymbol{\beta}} - z_{\boldsymbol{\alpha}+\boldsymbol{\beta}+[0,1]^\top}$, and
    $[\widetilde{L}_r(\hat{g}_4)]_{\boldsymbol{\alpha}\boldsymbol{\beta}} = w_r^\star z_{\boldsymbol{\alpha}+\boldsymbol{\beta}} + z_{\boldsymbol{\alpha}+\boldsymbol{\beta}+[0,1]^\top}$.
Compute $\underline{\varrho}_r^\star$ via
    $\underline{\varrho}_r^\star = \arg \min_{\rho, \mathbf{y}_{2,d}} \rho$
        subject to (6.15)–(6.17),
            $M_r \succeq 0$,                                        (6.54)
            $L_r(g_i) \succeq 0$, for $i \in [4]$
            $L_r(\hat{g}_3) \succeq 0, L_r(\hat{g}_4) \succeq 0$.
Obtain $\overline{\varrho}_r^\star$ via:
    $\overline{\varrho}_r^\star = \arg \max_{\rho, \mathbf{z}_{2,d}} \rho$
        subject to (6.30)–(6.33),
            $\widetilde{M}_r \succeq 0$,                                        (6.55)
            $\widetilde{L}_r(g_i) \succeq 0$, for $i \in [4]$,
            $\widetilde{L}_r(\hat{g}_3) \succeq 0, \widetilde{L}_r(\hat{g}_4) \succeq 0$.
```

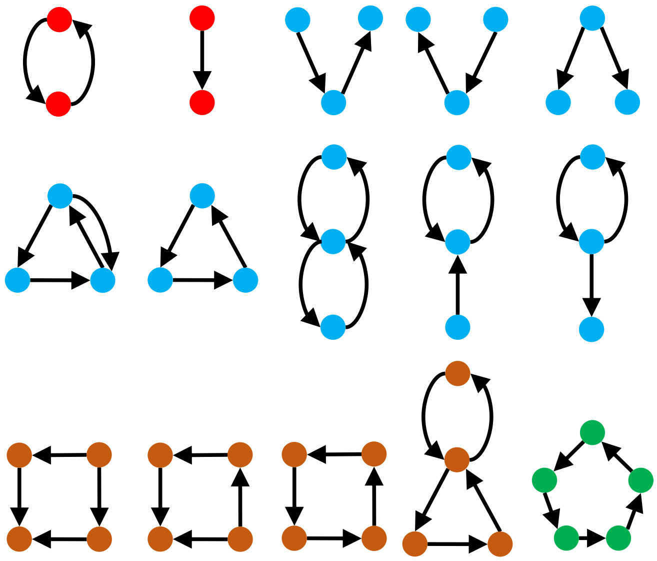
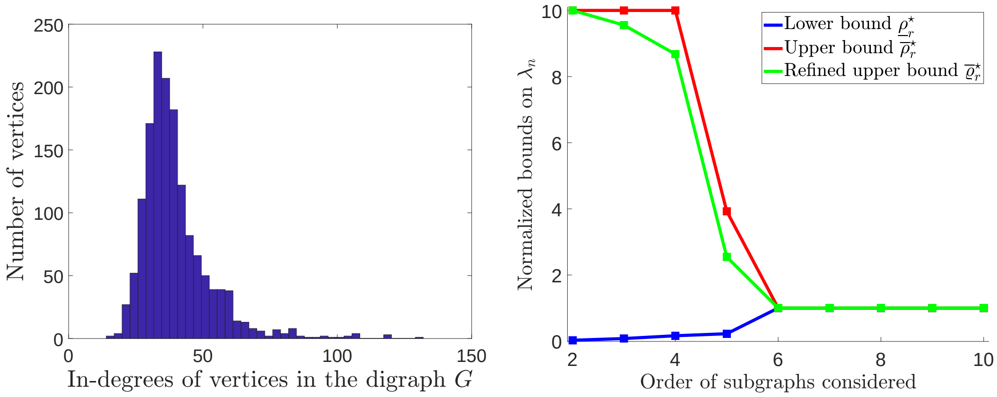


Figure 6-4: This figure shows those subgraphs whose counts are needed for estimating the spectral radius of A using Algorithm 1 with $r = 2$ (i.e., $d = 5$).



(a)

(b)

Figure 6-5: In (a), we show the histogram of in-degrees of one realization of the Chung-Lu random digraph. In (b), we show the normalized lower (in blue) and upper bounds, where the red and green lines show the upper bound obtained using Theorem 5 and Algorithm 1, respectively.

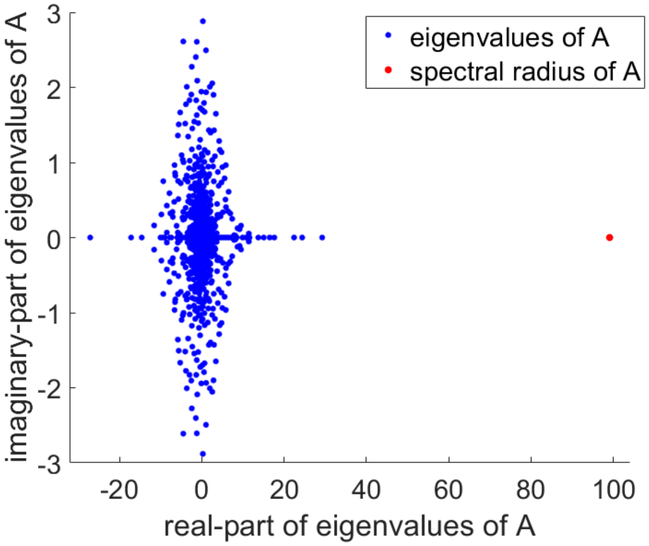


Figure 6-6: This figure shows the eigenvalue spectrum of the digraph representing flights between airports in the U.S. The x-axis and y-axis are the real and imaginary parts of the eigenvalues of A, respectively.

Chapter 7

Moment-closure Analysis and Control of Spreading Processes

As illustrated in the previous chapter, we can the leverage connection between the multidimensional moment problem and semidefinite programming to characterize global property of a directed graph. In this chapter, we further exploit this connection to analyze networked stochastic spreading processes.

The rest of the paper is organized as follows. In Section 7.1, we provide preliminaries and a description of the nonhomogeneous SIS spreading process, as well as additional background on the multidimensional moment problem. The proposed moment-closure framework is introduced in Section 7.2, where we focus our attention on the networked SIS model. In Section 7.3, we discuss how to apply this moment-closure technique to both the SI and the SIR epidemic models. In Section 7.4, we illustrate the performance of our framework by numerically analyzing several spreading processes taking place over a real-world social network. In Section 7.5, we analyze the spread of multiple diseases in a multilayer contact network and design efficient methodologies to contain all diseases simultaneously.

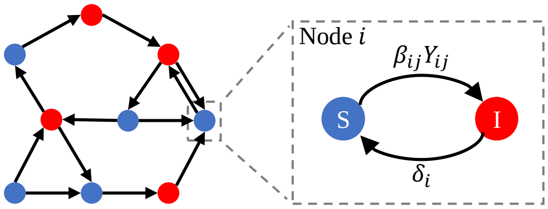


Figure 7-1: Illustration of the SIS spreading model on a directed graph. In the left subfigure, the black arrows represent the edges in the digraph, whereas the red and blue circles represent the infected and susceptible nodes, respectively. In the right subfigure, we show the possible transitions between states of a node i. The variable Y_{ij} represents the event that an in-neighbor j of i is infected.

7.1 Networked SIS Spreading Model and the K-moment Problem

In this section, we introduce notions related to networked epidemic models and multidimensional moment problem. Throughout this chapter, we adopt standard notions from graph theory (see Section 2.1.2 in Chapter 2).

7.1.1 Heterogeneous Networked SIS Spreading Model

We first describe the *Susceptible-Infected-Susceptible* (SIS) model, which is commonly used to characterize epidemics over networked populations. In the coming sections, we introduce a novel technique to analyze the stochastic dynamics of this, and other, epidemic models. For clarity in our exposition, we first illustrate the proposed technique using the SIS epidemic model; we then extend our analysis to other models, such as the SI and SIR models, in Section 7.3.

Next, we describe the continuous-time heterogeneous SIS spreading model on the graph G, [127]. In this model, at a given time $t \geq 0$, each node can be in one of the following two states: (i) '*Susceptible*', representing the case of a healthy node, and (ii) '*Infected*', in which the node is infected by a disease propagating through the network. On one hand, whenever node i is in the Susceptible state, i can be infected by one of its infected in-neighbour $j \in \mathcal{N}_i^-$ according to a Poisson process with parameter $\beta_{ij} > 0$, called the *infection rate* of edge (j, i). On the other hand, if node i is in the Infected state at a given time t, it cures itself according to a Poisson process with parameter δ_i, called the *recovery rate* of node i. We use a binary variable $x_i(t) \in \{0, 1\}$ to represent the state of node $i \in \mathcal{V}$ at time $t \geq 0$. More specifically, $x_i(t) = 0$ if node i is Susceptible, and $x_i(t) = 1$ if it is Infected at time $t \geq 0$. We illustrate the SIS spreading model in Figure 7-1.

The exact evolution of the random variables $x_i(t)$ can be characterized by a continuous-time Markov process with the following transition probabilities:

$$\begin{aligned} \mathbb{P}\left(x_i(t+h) = 1 \mid x_i(t) = 0\right) &= \sum_{j \in \mathcal{N}_i^-} \beta_{ij} x_j(t) h + o(h), \\ \mathbb{P}\left(x_i(t+h) = 0 \mid x_i(t) = 1\right) &= \delta_i h + o(h). \end{aligned} \tag{7.1}$$

Notice that the dimension of the state space of the Markov process in (7.1) is 2^n; hence, an exact analysis of the stochastic process is computationally challenging when the size of the underlying network is large. In what follows, we are interested in analyzing the dynamics of the probability of a node $i \in \mathcal{V}$ being infected at time t, i.e., $\mathbb{P}(x_i(t) = 1) = \mathbb{E}[x_i(t)]$. As illustrated in [84], the governing equations for the evolution of the expectation of $x_i(t)$ is given by[1]

$$\frac{d}{dt}\mathbb{E}\left[x_i\right] = -\delta_i \mathbb{E}\left[x_i\right] + \sum_{j \in \mathcal{N}_i^-} \beta_{ij} \mathbb{E}\left[x_j\right] - \sum_{j \in \mathcal{N}_i^-} \beta_{ij} \mathbb{E}\left[x_i x_j\right]. \tag{7.2}$$

We refer to (7.2) as the *mean SIS dynamics* of node i. In order to solve (7.2), it is necessary to characterize the second-order moment $\mathbb{E}[x_i x_j]$ for all $j \in \mathcal{N}_i^-$. However, as shown in [135], the evolution of $\mathbb{E}[x_i x_j]$ depends on third-order moments of the form $\mathbb{E}[x_i x_j x_k]$, which in turn, forces us to characterize $\mathbb{E}[x_i x_j x_k]$. More generally, one can prove that, in order to characterize the evolution of a k-th order moment, one needs to characterize the time derivatives of moments of order $k+1$. As a result of this recursive dependency, the evolution of the mean SIS dynamics is fully characterized by 2^n ordinary

[1] Whenever clear from the context, we shall remove the time-dependent notation from the random variable $x_i(t)$.

differential equations.

In order to obtain a computationally tractable approximation of the mean SIS dynamics, it is common to use moment-closure techniques in which one approximates k-th order moments using lower-order moments (see for example [135]). In particular, the mean-field approximation (MFA) is a widely adopted moment-closure technique in which one assumes that $\mathbb{E}[x_i x_j] = \mathbb{E}[x_i]\mathbb{E}[x_j]$. Hence, defining the moment variable $\mu_i = \mathbb{E}[x_i]$, (7.2) turns into the following system of n non-linear differential equations:

$$\dot{\mu}_i = -\delta_i \mu_i + \sum_{j \in \mathcal{N}_i^-} \beta_{ij} \mu_j - \sum_{j \in \mathcal{N}_i^-} \beta_{ij} \mu_i \mu_j. \tag{7.3}$$

Nonetheless, using this approximation, we do not have any quality guarantee about whether $\mu_i(t)$ is an upper or lower bound of the expectation $\mathbb{E}[x_i(t)]$.

In this paper, we develop a systematic framework to perform moment-closure with quality guarantees by using recent results on the K-moment problem [108]. The proposed framework is capable of providing both upper and lower bounds on the mean SIS spreading process. In the next subsection, we provide necessary background on the K-moment problem. We use the SIS model as a running example to illustrate the proposed technique. In Section 7.3, we will extend this technique to analyze other epidemic models, such as the networked SI and SIR models.

7.1.2 The K-moment Problem

To explain our approach, we recall the following concepts from the theory behind the K-moment problem (see Section 6.2.3 in Chapter 6 for more details). As noted by Theorem 22, a necessary and sufficient condition for the feasibility of the K-moment problem, restricted to the case when K is semi-algebraic and compact, can be stated in terms of linear matrix inequalities involving *moment matrices* and *localizing matrices* (see Definition 21 and Definition 22). We recall these two important notions here to ease the readability. In order to define these matrices, we introduce the following notions.

Given an integer $r \in \mathbb{N}$, we define the vector

$$\mathbf{v}_r(\mathbf{x}) := \left[1, x_1, \dots, x_n, x_1^2, x_1x_2, \dots, x_1^r, \dots, x_n^r\right]^\top, \tag{7.4}$$

i.e., the vector containing the monomials of the canonical basis of real-valued polynomials of degree at most r. Furthermore, given an integer vector $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_n]^\top \in \mathbb{N}_r^n$, we define $[\mathbf{v}_r]_{\boldsymbol{\alpha}} = \mathbf{x}^{\boldsymbol{\alpha}}$.

Definition 21. *Given an $\mathbb{R}^n$-valued random variable $\mathbf{x}$, the moment matrix of $\mathbf{x}$ of order $2r$ is defined as $M_r = \mathbb{E}[\mathbf{v}_r(\mathbf{x})\mathbf{v}_r(\mathbf{x})^\top]$.*

Let $\mathbf{y} = \{y_{\boldsymbol{\alpha}}\}_{|\boldsymbol{\alpha}| \leq 2r}$ be a finite multi-sequence such that $y_{\boldsymbol{\alpha}} = \mathbb{E}\left[\mathbf{x}^{\boldsymbol{\alpha}}\right]$ for all $|\boldsymbol{\alpha}| \leq 2r$. Then, we can index the entries of the moment matrix M_r using two elements of $\mathbb{N}_r^n$ as follows. Given two elements $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{N}_r^n$, the $(\boldsymbol{\alpha}, \boldsymbol{\beta})$-th entry of M_r, denoted by $[M_r]_{\boldsymbol{\alpha},\boldsymbol{\beta}}$, is equal to $\mathbb{E}[[\mathbf{v}_r]_{\boldsymbol{\alpha}}[\mathbf{v}_r]_{\boldsymbol{\beta}}] = \mathbb{E}[\mathbf{x}^{\boldsymbol{\alpha}}\mathbf{x}^{\boldsymbol{\beta}}] = y_{\boldsymbol{\alpha}+\boldsymbol{\beta}}$.

Similarly, we define the *localizing matrices* as follows.

Definition 22. *Given an $\mathbb{R}^n$-valued random variable $\mathbf{x}$, and a polynomial $g : \mathbb{R}^n \to \mathbb{R}$, we define the localizing matrix of $\mathbf{x}$ with respect to g as $L_r(g) = \mathbb{E}\left[g(\mathbf{x})\mathbf{v}_r(\mathbf{x})\mathbf{v}_r(\mathbf{x})^\top\right]$.*

Let $deg(g)$ be the degree of the polynomial g. Then, g can be written as

$$g(\mathbf{x}) = \sum_{\boldsymbol{\gamma} \in \mathbb{N}^n_{deg(g)}} c_{\boldsymbol{\gamma}} \mathbf{x}^{\boldsymbol{\gamma}},$$

where $\mathbf{x}^{\boldsymbol{\gamma}}$ is a monomial (i.e., an entry in $\mathbf{v}_{deg(g)}(\mathbf{x})$) and $c_{\boldsymbol{\gamma}}$ is its corresponding coefficient. Let $\mathbf{y} = \{y_{\boldsymbol{\alpha}}\}_{|\boldsymbol{\alpha}| \leq 2r+deg(g)}$ be the multi-sequence of moments such that $y_{\boldsymbol{\alpha}} = \mathbb{E}\left[\mathbf{x}^{\alpha}\right]$ for all $\boldsymbol{\alpha} \in \mathbb{N}^n_{2r+deg(g)}$. Hence, the entries of the localizing matrix can be indexed using two entries of $\mathbb{N}^n_{r+deg(g)}$, as follows. Given two elements $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{N}_r^n$, the $(\boldsymbol{\alpha}, \boldsymbol{\beta})$-th entry of $L_r(g)$, denoted by $[L_r(g)]_{\boldsymbol{\alpha},\boldsymbol{\beta}}$, is equal to $\mathbb{E}[g(\mathbf{x})[\mathbf{v}_r]_{\boldsymbol{\alpha}}[\mathbf{v}_r]_{\boldsymbol{\beta}}] = \mathbb{E}\left[\sum_{\boldsymbol{\gamma} \in \mathbb{N}^n_{deg(g)}} c_{\boldsymbol{\gamma}} \mathbf{x}^{\boldsymbol{\gamma}} \mathbf{x}^{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\beta}}\right] = \sum_{\boldsymbol{\gamma} \in \mathbb{N}^n_{deg(g)}} c_{\boldsymbol{\gamma}} y_{\boldsymbol{\alpha}+\boldsymbol{\beta}+\boldsymbol{\gamma}}$. In Chapter 6, we have stated a condition for a finite multi-sequence to be K-feasible for a compact and semi-algebraic set K (see Theorem 22). In this section, we state an anlogous necessary and sufficient condition for the K-feasibility of an *infinite* multi-sequence. Before stating this theorem, we first need to define the following notion [108].

Definition 23. *A polynomial $p : \mathbb{R}^n \to \mathbb{R}$ is a sum-of-squares (SOS) if p can be written as*

$$p(\mathbf{x}) = \sum_{j=1}^{J} p_j(\mathbf{x})^2, \tag{7.5}$$

for some finite set of polynomials $\{p_j \colon j \in [J]\}$.

A necessary and sufficient condition for an *infinite* multi-sequence $\mathbf{y} = \{y_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathbb{N}^n}$ to be K-feasible, restricted to the case when K is both compact and semi-algebraic, is as follows.

Theorem 27. (Putinar's Positivstellensatz, [175]) *Consider an infinite multi-sequence $\mathbf{y} = \{y_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathbb{N}^n}$, and a collection of polynomials $g_i : \mathbb{R}^n \to \mathbb{R}$, for all $i \in [m]$. Define a compact semi-algebraic set $K = \{\mathbf{x} \in \mathbb{R}^n \colon g_i(\mathbf{x}) \geq 0,\ i \in [m]\}$. Assume that there exists a polynomial $u = u_0 + \sum_{i=1}^m u_i g_i$, where u_i are SOS polynomials for all $i \in [m]$, such that the set $\{\mathbf{x} \colon u(\mathbf{x}) \geq 0\}$ is compact. Then, the multi-sequence $\mathbf{y}$ has a K-representing measure, if and only if,*

$$\begin{aligned} &M_r(\mathbf{y}) \succeq 0, \ \textit{and} \\ &L_r(g_j \mathbf{y}) \succeq 0, \ \textit{for all } j \in [m], \ \textit{and } r \in \mathbb{N}. \end{aligned} \tag{7.6}$$

Remark 16. *Theorem 22 in Chapter 6 and Theorem 27 stated necessary and sufficient conditions for a finite and an infinite multi-sequence to be K-feasible, respectively.*

Remark 17. *Using* (7.6) *to verify the K-feasibility of a given multi-sequence requires checking the positive semi-definiteness of $m + 1$ matrices for each $r \in \mathbb{N}$. Moreover, the dimension of these matrices grows with r.*

Based on Theorem 27, one can verify whether a given multi-sequence is a K-feasible moment sequence by solving an infinite sequence of semi-definite programs. On the other hand, given a finite moment sequence, up to a certain order, one can use Theorem 27 to derive conditions on higher-order moments for the multi-sequence of moments to be feasible. In the next section, we use this idea to provide upper and lower bounds on the evolution of $\mathbb{E}[x_i]$, described in (7.2).

7.2 SDP-based Moment-closure

In this section, we first characterize the dynamics of the $\boldsymbol{\alpha}$-moment of the random vector describing the state of the SIS model, for an arbitrary $\boldsymbol{\alpha} \in \mathbb{N}^k$ (Subsection 7.2.1). Then, we show that the problem of obtaining upper and lower bounds on the evolution of the $\boldsymbol{\alpha}$-moment is closely related to the K-moment problem (Subsection 7.2.2). Finally, we obtain a closed-form expression for the mean dynamics of the SIS spreading process (Subsection 7.2.3). In Section 7.3, we will extend our results to other networked epidemic models, such as the SI and SIR models.

7.2.1 Dynamics of the $\boldsymbol{\alpha}$-moment in the SIS Spreading Process

As discussed in Subsection 7.1.1, in order to close the system of differential equations describing the mean SIS dynamics (7.2), it is necessary to characterize the dynamics of second-order moments $\mathbb{E}[x_i x_j]$ for all $(i,j) \in \mathcal{E}$. More generally, in order to characterize the mean dynamics of any k-th order moment of the form $\mathbb{E}[x_{i_1} \cdots x_{i_k}]$, we need to obtain an expression for the $(k+1)$-th order differential $dx_{i_1} \cdots x_{i_{k+1}}$. To undertake the problem of finding a closed system of differential equations to describe the mean SIS spreading process, we propose the following three-step approach: First, we describe the stochastic dynamics of the networked SIS process using jump processes [176]. Second, we use Ito's formula for jump processes to obtain a governing equation for high-order differentials, i.e., an expression for $dx_{i_1} \cdots x_{i_k}$ for arbitrary k. Finally, we derive explicit differential equations allowing us to upper and lower bound the dynamics of any k-th order moment of the SIS spreading process. To achieve our goals, we first introduce related notions on Poisson jump processes.

Definition 24. [176] *Given $\gamma > 0$, a stochastic process P_t^γ is called a Poisson jump process with rate γ if: (i) for every $s, t > 0$, the random variable $P_{s+t}^\gamma - P_s^\gamma$ is independent of $\{P_{t'}^\gamma : t' \leq s\}$ and follows the same distribution as $P_t^\gamma - P_0^\gamma$, and (ii) the random variable $P_t^\gamma - P_0^\gamma$ follows a the Poisson distribution with mean γt, i.e., $\mathbb{P}(P_t^\gamma - P_0^\gamma = k) = e^{\gamma t} \frac{(\gamma t)^k}{k!}$.*

In what follows, we abbreviate P_t^γ as P_γ for convenience. Using Poisson jump processes, the evolution of the states $x_i(t)$ in the SIS spreading process described in (7.1) can be characterized by the following set of stochastic differential equations:

$$dx_i = -x_i dP_{\delta_i} + (1 - x_i) \sum_{j \in \mathcal{N}_i^-} x_j dP_{\beta_{ij}}, \tag{7.7}$$

with $x_i(0) \in \{0, 1\}$ for all $i \in \mathcal{V}$. Notice that, we can recover the first-order mean dynamics of the SIS spreading process in (7.2) by taking expectation of (7.7). In order to obtain the dynamics of the second-order differential $dx_i x_j$, we use Ito's formula for jump processes, as stated below:

Theorem 28. [176] *Let* $\mathbf{x}(t)$ *be an* $\mathbb{R}^n$*-valued random variable for all* $t > 0$, *and* $\phi : \mathbb{R}^n \to \mathbb{R}$ *be a twice continuously-differentiable function. If*

$$d\mathbf{x}(t) = \sum_{k=1}^{n_p} \mathbf{h}_k(\mathbf{x}) dP_{\gamma_k}, \tag{7.8}$$

where $\mathbf{h}_k : \mathbb{R}^n \to \mathbb{R}^n$, *for all* $k \in [n_p]$, *then*

$$d\phi(\mathbf{x}) = \sum_{k=1}^{n_p} \left[\phi(\mathbf{x} + \mathbf{h}_k(\mathbf{x})) - \phi(\mathbf{x})\right] dP_{\gamma_k}. \tag{7.9}$$

As an example, we let $\phi(\mathbf{x}) = x_i x_j$, and apply Theorem 28 on (7.7). Subsequently, after tedious (but simple) algebraic manipulations, we obtain

$$\begin{aligned} dx_i x_j = {} & -x_i x_j (dP_{\delta_i} + dP_{\delta_j}) + (1 - x_i) x_j \sum_{k \in \mathcal{N}_i^-} x_k dP_{\beta_{ik}} \\ & + (1 - x_j) x_i \sum_{k \in \mathcal{N}_j^-} x_k dP_{\beta_{jk}}. \end{aligned} \tag{7.10}$$

If the SIS spreading process is homogeneous, i.e., $\delta_i = \delta$ for all $i \in [n]$ and $\beta_{ij} = \beta$ for all $(i,j) \in \mathcal{E}$, then taking the expectation of (7.10) results in:

$$\begin{aligned}\frac{d\mathbb{E}[x_i x_j]}{dt} &= -2\delta\mathbb{E}[x_i x_j] - \beta\sum_{k=1}^{n}(a_{jk}+a_{ik})\mathbb{E}[x_i x_j x_k] \\ &+ \beta\left[\sum_{k\in\mathcal{N}_i^-}\mathbb{E}[x_i x_k] + \sum_{k\in\mathcal{N}_j^-}\mathbb{E}[x_j x_k]\right],\end{aligned}$$

which reduces to the result in [135]. More generally, we can use (7.9) to derive explicit expressions of higher-order differentials, i.e., $dx_1^{\alpha_1}x_2^{\alpha_2}\cdots x_n^{\alpha_n}$ for arbitrary $\alpha_1,\ldots,\alpha_n \in \mathbb{N}$, for the SIS spreading model, as stated in the following theorem.

Theorem 29. *Given a collection of k integers $i_1,\ldots,i_k \in [n]$ and a vector of positive integers $\boldsymbol{\alpha} \in \mathbb{N}^k$, we define the following monomials $\phi_{\boldsymbol{\alpha}}(\mathbf{x}) = x_{i_1}^{\alpha_1}\cdots x_{i_k}^{\alpha_k}$, $\phi_{\boldsymbol{\alpha}}^{-s}(\mathbf{x}) = x_{i_1}^{\alpha_1}\cdots x_{i_{s-1}}^{\alpha_{s-1}}x_{i_{s+1}}^{\alpha_{s+1}}\cdots x_{i_k}^{\alpha_k}$, and $\phi_{\mathbf{1}}(\mathbf{x}) = x_{i_1}\cdots x_{i_k}$. Consider a directed graph $G = (\mathcal{V},\mathcal{E})$, and the set of stochastic differential equations described in (7.7). Then, the evolution of the $\boldsymbol{\alpha}$-moment satisfies*

$$\begin{aligned}\frac{d\mathbb{E}[\phi_{\boldsymbol{\alpha}}(\mathbf{x})]}{dt} &= -\sum_{s=1}^{k}\delta_{i_s}\mathbb{E}[\phi_{\mathbf{1}}(\mathbf{x})] \\ &+\sum_{s=1}^{k}\sum_{\ell\in\mathcal{N}_{i_s}^-}\beta_{i_s\ell}\left(\mathbb{E}[\phi_{\mathbf{1}}^{-s}(\mathbf{x})x_\ell] - \mathbb{E}[\phi_{\mathbf{1}}(\mathbf{x})x_\ell]\right).\end{aligned} \tag{7.11}$$

Proof. See Appendix A.6. □

This theorem shows that the time derivative of the $\boldsymbol{\alpha}$-moment $\mathbb{E}[\phi_{\boldsymbol{\alpha}}(\mathbf{x})]$ depends on $\mathbb{E}[\phi_{\mathbf{1}}(\mathbf{x})x_\ell]$, which is a moment of higher order. In order to close the differential equation in (7.11), we propose to approximate $\mathbb{E}[\phi_{\mathbf{1}}(\mathbf{x})x_\ell]$ using $\mathbb{E}[\phi_{\boldsymbol{\alpha}}(\mathbf{x})]$ for $|\boldsymbol{\alpha}| \le k$. In the next section, we achieve this goal by upper and lower bound the term $\mathbb{E}[\phi_{\mathbf{1}}(\mathbf{x})x_\ell]$.

7.2.2 SDP-based Moment Closure

In this subsection, we will develop a framework to obtain both upper and lower bounds on the dynamics of the $\boldsymbol{\alpha}$-moment, $\mathbb{E}[\phi_{\boldsymbol{\alpha}}]$. In this direction, we will bound the higher-order term $\mathbb{E}[\phi_{\mathbf{1}}(\mathbf{x})x_\ell]$ using lower-order moments $\mathbb{E}[\phi_{\boldsymbol{\beta}}(\mathbf{x})]$ for $|\boldsymbol{\beta}| \le k$. For example, the widely used mean-field approximation [135] is an approach to close the first-order mean-dynamics $\mathbb{E}[x_i]$ by approximating the second-order term $\mathbb{E}[x_i x_j]$ using the following product of two first-order terms $\mathbb{E}[x_i]\mathbb{E}[x_j]$.

In what follows, we develop a framework to find two systems of differential equations whose solutions are guaranteed to upper and lower bound the dynamics of any $\boldsymbol{\alpha}$-moment. Our approach utilizes Putinar's Positivstellensatz to derive bounds on an $\boldsymbol{\alpha}$-moment in terms of lower-order moments. Before we present this approach, we first introduce several definitions. Given a set $\mathcal{I} \subseteq [n]$, we define $\mu_{\mathcal{I}} = \mathbb{E}[\Pi_{i\in\mathcal{I}} x_i]$. Furthermore, given a set of k *distinct* indices $\mathcal{I}_k = \{i_1, \ldots, i_k\} \subseteq [n]$, we define the (finite) multi-sequence of moments $\mathbf{y}(\mathcal{I}_k) = \{\mathbb{E}\left[\Pi_{s\in S} x_s\right]\}_{S\subseteq\mathcal{I}_k, |S|<k}$.

In what follows, we bound the moment $\mathbb{E}[\phi_{\boldsymbol{\alpha}}] = \mu_{\mathcal{I}_k}$ using lower-order moments contained in the set $\mathbf{y}(\mathcal{I}_k)$. To achieve this goal, we notice that at each time $t > 0$, $\mathbf{x}(t)$ is a $\{0,1\}^n$-valued random variable. Subsequently, for a given time t, $[x_{i_1}(t), \ldots, x_{i_k}(t)]^\top$ follows a distribution supported on $\{0,1\}^k$. In particular, the $\boldsymbol{\alpha}$-moment of the random vector $[x_{i_1}(t), \ldots, x_{i_k}(t)]^\top$ for $\boldsymbol{\alpha} = \mathbf{1}_k$ is equal to $\mu_{\mathcal{I}_k}$. Therefore, the sequence of moments $\hat{\mathbf{y}}(\mathcal{I}_k) = \mathbf{y}(\mathcal{I}_k) \cup \{\mu_{\mathcal{I}_k}\}$ must be $\{0,1\}^k$-feasible. Consequently, an upper bound (respectively, lower bound) on $\mu_{\mathcal{I}_k}$ can be obtained by finding the largest (respectively, smallest) value of $\mu_{\mathcal{I}_k}$ such that $\hat{\mathbf{y}}(\mathcal{I}_k)$ is $\{0,1\}^k$-feasible.

To achieve the above objective, we propose to exploit the semidefinite inequalities in Theorem 27, regarding the moment and localizing matrices of $\mathbf{y}(\mathcal{I}_k)$. However, (7.6) provides conditions for an *infinite* sequence to be K-feasible whereas the sequence $\mathbf{y}(\mathcal{I}_k)$ is finite. To circumvent this issue, we will extend the finite multi-sequence of moments $\mathbf{y}(\mathcal{I}_k)$ into an infinite sequence such that the results in Theorem 27 are applicable. As we discuss below, this extension is possible due to the binary nature of the random variables x_i. More specifically, although $\mathbf{y}(\mathcal{I}_k)$ contains a finite sequence of moments, we can extend this sequence using the following observation: Given a set of q disjoint indices $\{i_1, \ldots, i_q\} \subseteq [n]$, since x_i are binary random variables, we have that:

$$\mathbb{E}[\Pi_{s=1}^q x_{i_s}] = \mathbb{E}[\Pi_{s=1}^q x_{i_s}^{\alpha_s}], \tag{7.12}$$

for all $q \leq k$, where $\alpha_s > 0$ for all $s \in [q]$. Subsequently, $\mathbf{y}(\mathcal{I}_k)$ can be extended uniquely into an infinite sequence, as follows: Given $\hat{\mathbf{y}}(\mathcal{I}_k)$, we construct its associated *infinite extension* as $\mathbf{y}_\infty(\mathcal{I}_k) = \{y_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathbb{N}^k}$, with $y_{\boldsymbol{\alpha}}$ satisfying (7.12). Consequently, given a compact semi-algebraic set K, the K-feasibility of $\mathbf{y}(\mathcal{I}_k)$ is equivalent to the K-feasibility of $\mathbf{y}_\infty(\mathcal{I}_k) = \{y_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathbb{N}^k}$.

Secondly, in order to apply Theorem 27, we show below (in Lemma 15) that the infinite-dimensional matrices in (7.6) are positive semidefinite, if and only if, certain finite-dimensional matrices are positive semidefinite. Before rigorously stating this claim, we need to introduce several additional notions. Given $k \leq n$, we let $\kappa = \lceil k/2 \rceil$, and $N_k = \binom{k+\kappa}{k}$. Finally, given $s \in [k]$, we let e_s denote the s-th standard basis vector of $\mathbb{R}^k$. With the help of these notions, we define the following finite-dimensional matrices. Let $M_\kappa(\hat{\mathbf{y}}(\mathcal{I}_k)) \in \mathbb{R}^{N_k \times N_k}$ be defined entry-wise by

$$[M_\kappa(\hat{\mathbf{y}}(\mathcal{I}_k))]_{\boldsymbol{\alpha},\boldsymbol{\beta}} = y_{\boldsymbol{\alpha}+\boldsymbol{\beta}}, \tag{7.13}$$

for all $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{N}^k_\kappa$. Essentially, if $y_{\boldsymbol{\gamma}} = \mathbb{E}[\mathbf{x}^{\boldsymbol{\gamma}}]$ for all $\boldsymbol{\gamma} \in \mathbb{N}^k$, then M_κ is the principal sub-matrix of size N_k of the infinite moment matrix in Theorem 27. In addition to M_κ, we now construct a collection of finite-dimensional matrices to replace the infinite-dimensional localizing matrices in Theorem 27. To achieve this goal, we first notice that the measure of the random vector $[x_{i_1}, \ldots, x_{i_k}]^\top$ is supported on $\tilde{S}_k = [0,1]^k \supset \{0,1\}^k$, which is both compact and semi-algebraic. By defining $g_s^1(\mathbf{x}) = 1 - x_{i_s}$ and $g_s^0(\mathbf{x}) = x_{i_s}$ for all $s \in [k]$, the hypercube $\tilde{S}_k$ can be represented as

$$\tilde{S}_k = \{\mathbf{x} \in \mathbb{R}^k : g_s^1(\mathbf{x}) \geq 0, g_s^0(\mathbf{x}) \geq 0, \forall s \in [k]\}. \tag{7.14}$$

Next, for each $s \in [k]$, we define two matrices $[L^1_\kappa(\hat{\mathbf{y}}(\mathcal{I}_k), s)]$ and $[L^0_\kappa(\hat{\mathbf{y}}(\mathcal{I}_k), s)]$, as follows:

$$[L^1_\kappa(\hat{\mathbf{y}}(\mathcal{I}_k), s)]_{\boldsymbol{\alpha},\boldsymbol{\beta}} = y_{\boldsymbol{\alpha}+\boldsymbol{\beta}} - y_{\boldsymbol{\alpha}+\boldsymbol{\beta}+e_s}, \tag{7.15}$$

and

$$[L_\kappa^0(\hat{\mathbf{y}}(\mathcal{I}_k), s)]_{\boldsymbol{\alpha},\boldsymbol{\beta}} = y_{\boldsymbol{\alpha}+\boldsymbol{\beta}+e_s}. \tag{7.16}$$

As we will prove in Lemma 15, the matrices in (7.15) and (7.16) can be used as finite-dimensional localizing matrices for $g_s^1(\mathbf{x})$ and $g_s^0(\mathbf{x})$, respectively.

Remark 18. *From* (7.12)*, we see that whenever $y_{\boldsymbol{\alpha}+\boldsymbol{\beta}} = y_{\boldsymbol{\alpha}+\boldsymbol{\beta}+e_s}$, the corresponding entry in* (7.15) *is zero. More specifically, given an integer vector $\boldsymbol{\alpha} \in \mathbb{N}^k$, we define $\mathcal{A} = \{i : [\boldsymbol{\alpha}]_i \neq 0\}$ and $\mathcal{B} = \{i : [\boldsymbol{\beta}]_i \neq 0\}$. Consequently, $[L_\kappa^1(\hat{\mathbf{y}}(\mathcal{I}_k), s)]_{\boldsymbol{\alpha},\boldsymbol{\beta}} = 0$, if and only if,*

$$\mathcal{A} \cup \mathcal{B} = \mathcal{A} \cup \mathcal{B} \cup \{i_s\}. \tag{7.17}$$

Next, we show that the positive semidefiniteness of the infinite-dimensional matrices in Theorem 27 is equivalent to the positive semidefiniteness of the finite-dimensional matrices in (7.13), (7.15), and (7.16).

Lemma 15. *Let $\{x_{i_s}\}_{s\in[k]}$ be a collection of binary random variables such that $\hat{\mathbf{y}}(\mathcal{I}_k) = \{\mathbb{E}\left[\Pi_{s\in S} x_s\right]\}_{S\subseteq \mathcal{I}_k}$, and denote its associated infinite extension by $\mathbf{y}_\infty(\mathcal{I}_k)$. Then, the sequence $\mathbf{y}_\infty(\mathcal{I}_k)$ is $\tilde{S}_k$-feasible, if and only if,*

$$\begin{aligned} M_\kappa(\hat{\mathbf{y}}(\mathcal{I}_k)) \succeq 0, \ \textit{and} \\ L_\kappa^1(\hat{\mathbf{y}}(\mathcal{I}_k), s) \succeq 0, L_\kappa^0(\hat{\mathbf{y}}(\mathcal{I}_k), s) \succeq 0, \forall s \in [k]. \end{aligned} \tag{7.18}$$

Proof. See Appendix A.6. □

Remark 19. *From* (7.13), (7.15), *and* (7.16)*, we have that $M_\kappa(\hat{\mathbf{y}}(\mathcal{I}_k)) = L_\kappa^1(\hat{\mathbf{y}}(\mathcal{I}_k), s) + L_\kappa^0(\hat{\mathbf{y}}(\mathcal{I}_k), s)$. Subsequently, positive semidefiniteness of $L_\kappa^1(\hat{\mathbf{y}}(\mathcal{I}_k), s)$ and $L_\kappa^0(\hat{\mathbf{y}}(\mathcal{I}_k), s)$ implies that $M_\kappa(\hat{\mathbf{y}}(\mathcal{I}_k)) \succeq 0$.*

Based on the above lemma, we can derive upper and lower bounds on the moment $\mathbb{E}[\phi_{\mathbf{1}}(\mathbf{x})x_\ell] = \mu_{\mathcal{I}_k\cup\{\ell\}}$ in (7.11) by solving, respectively, the following semidefinite programs:

$$\overline{\mu}_{\mathcal{I}_k\cup\{\ell\}} = \max_{\mu_{\mathcal{I}_k\cup\{\ell\}}} \mu_{\mathcal{I}_k\cup\{\ell\}} \text{ s.t. (7.18) holds.} \tag{7.19}$$

$$\underline{\mu}_{\mathcal{I}_k\cup\{\ell\}} = \min_{\mu_{\mathcal{I}_k\cup\{\ell\}}} \mu_{\mathcal{I}_k\cup\{\ell\}} \text{ s.t. (7.18) holds.} \tag{7.20}$$

Hence, given a set $\mathcal{I}_k$, we have that $\mu_{\mathcal{I}_k\cup\{\ell\}} \in [\underline{\mu}_{\mathcal{I}_k\cup\{\ell\}}, \overline{\mu}_{\mathcal{I}_k\cup\{\ell\}}]$. Based on this, one could be tempted to obtain an upper (respectively, a lower) bound on the evolution of $\mathbb{E}[\phi_{\boldsymbol{\alpha}}]$ by solving the ODE in (7.11) after replacing the higher-order term $\mathbb{E}[\phi_{\mathbf{1}}(\mathbf{x})x_\ell]$ by $\underline{\mu}_{\mathcal{I}_k\cup\{\ell\}}$ (respectively, $\overline{\mu}_{\mathcal{I}_k\cup\{\ell\}}$). However, this is not true, since a monotone relationship between derivatives does not preserve the monotonicity between the solutions of the ODEs, as discussed in [177].

To address this issue, we propose to make slight modifications on the entries of the localizing matrices to invoke a multidimensional version of Grönwall's comparison lemma [177]. More specifically, for a given set $\mathcal{J} \subseteq \mathcal{I}_k$, let $\hat{\mu}_{\mathcal{J}}$ and $\check{\mu}_{\mathcal{J}}$ be upper and lower bounds on the moment $\mu_{\mathcal{J}}$, i.e., $\mu_{\mathcal{J}} \in [\check{\mu}_{\mathcal{J}}, \hat{\mu}_{\mathcal{J}}]$. For a given $\boldsymbol{\gamma} \in \mathbb{N}^k_\kappa$, let us define $\mathcal{J} = \{i \in [n] \colon [\boldsymbol{\gamma}]_i \neq 0\}$, as well as $\hat{y}_{\boldsymbol{\gamma}} = \hat{\mu}_{\mathcal{J}}$ and $\check{y}_{\boldsymbol{\gamma}} = \check{\mu}_{\mathcal{J}}$. Let us also define the following modifications on the localizing matrices described in (7.15) and (7.16):

$$[\tilde{L}^1_\kappa(\hat{\mathbf{y}}(\mathcal{I}_k), s)]_{\boldsymbol{\alpha},\boldsymbol{\beta}} = \begin{cases} 0, \text{ if (7.17) holds,} \\ \hat{y}_{\boldsymbol{\alpha}+\boldsymbol{\beta}} - \check{y}_{\boldsymbol{\alpha}+\boldsymbol{\beta}+e_s}, \text{ otherwise.} \end{cases}, \tag{7.21}$$

and

$$[\tilde{L}^0_\kappa(\mathbf{y}(\mathcal{I}_k), s)]_{\boldsymbol{\alpha},\boldsymbol{\beta}} = \hat{y}_{\boldsymbol{\alpha}+\boldsymbol{\beta}+e_s}. \tag{7.22}$$

In the next theorem, we formally show how to obtain upper and lower bounds on the evolution of $\mathbb{E}[\phi_{\boldsymbol{\alpha}}]$ using a modification of the ODE in (7.11) involving (7.21) and (7.22).

Theorem 30. *Given a directed graph $G = (\mathcal{V}, \mathcal{E})$, let us define a sequence of functions $\{\hat{\mu}_{\mathcal{I}}(t), \check{\mu}_{\mathcal{I}}(t)\}_{\mathcal{I}\subseteq[n],|\mathcal{I}|\leq k}$ satisfying the following ODEs:*

$$\begin{aligned}\frac{d\hat{\mu}_{\mathcal{I}}(t)}{dt} = &- \sum_{s=1}^{|\mathcal{I}|} \delta_{i_s}\hat{\mu}_{\mathcal{I}}(t) \\ &+ \sum_{s=1}^{|\mathcal{I}|} \sum_{\ell\in\mathcal{N}^-_{i_s}} \beta_{i_s\ell}\left(\hat{\mu}_{\mathcal{I}\cup\{\ell\}\setminus\{i_s\}}(t) - \underline{\mu}_{\mathcal{I}\cup\{\ell\}}(t)\right)\end{aligned}$$

and

$$\begin{aligned}\frac{d\check{\mu}_{\mathcal{I}}(t)}{dt} = &-\sum_{s=1}^{|\mathcal{I}|} \delta_{i_s} \check{\mu}_{\mathcal{I}}(t) \\ &+ \sum_{s=1}^{|\mathcal{I}|} \sum_{\ell \in \mathcal{N}_{i_s}^-} \beta_{i_s \ell} \left(\check{\mu}_{\mathcal{I}\cup\{\ell\}\setminus\{i_s\}}(t) - \overline{\mu}_{\mathcal{I}\cup\{\ell\}}(t) \right),\end{aligned}$$

for all $\mathcal{I} \subseteq [n], |\mathcal{I}| \leq k$, where

$$\underline{\mu}_{\mathcal{I}\cup\{\ell\}} = \begin{cases} \check{\mu}_{\mathcal{I}\cup\{\ell\}}, & \textit{if } |\mathcal{I}\cup\{\ell\}| \leq k, \\ \check{\mu}^{\star}_{\mathcal{I}\cup\{\ell\}}, & \textit{otherwise,} \end{cases} \tag{7.23}$$

and

$$\overline{\mu}_{\mathcal{I}\cup\{\ell\}} = \begin{cases} \hat{\mu}_{\mathcal{I}\cup\{\ell\}}, & \textit{if } |\mathcal{I}\cup\{\ell\}| \leq k, \\ \hat{\mu}^{\star}_{\mathcal{I}\cup\{\ell\}}, & \textit{otherwise.} \end{cases} \tag{7.24}$$

In particular, $\check{\mu}^{\star}_{\mathcal{I}\cup\{\ell\}}$ and $\hat{\mu}^{\star}_{\mathcal{I}\cup\{\ell\}}$ are, respectively, the solutions that minimize/maximize the following SDPs:

$$\begin{aligned} &\min_{\mu_{\mathcal{I}\cup\{\ell\}}} / \max_{\mu_{\mathcal{I}\cup\{\ell\}}} \mu_{\mathcal{I}\cup\{\ell\}} \\ &\quad \textit{subject to } \tilde{L}^1_{\kappa}(\mathbf{y}(\mathcal{I}\cup\{\ell\}), s) \succeq 0, \forall s \in [k], \\ &\quad\quad\quad\quad\quad\;\; \tilde{L}^0_{\kappa}(\mathbf{y}(\mathcal{I}\cup\{\ell\}), s) \succeq 0, \forall s \in [k]. \end{aligned} \tag{7.25}$$

Let $\mu_{\mathcal{I}}(t) = \mathbb{E}[\phi_{\boldsymbol{\alpha}}]$ be the solution of the ODE in (7.11)*. Then, if $\hat{\mu}_{\mathcal{I}}(0) \geq \mu_{\mathcal{I}}(0) \geq \check{\mu}_{\mathcal{I}}(0)$, we have that $\hat{\mu}_{\mathcal{I}}(t) \geq \mu_{\mathcal{I}}(t) \geq \check{\mu}_{\mathcal{I}}(t)$, for all $t \geq 0$ and $\mathcal{I} \in [n], |\mathcal{I}| \leq k$.* ◇

Proof. See Appendix A.6. □

In the above theorem, we have provided an SDP-based moment-closure procedure for SIS spreading process. More specifically, when $|\mathcal{I}| < k$, the ODEs in the statement of the above theorem resemble the ODE in (7.11). Nonetheless, when $|\mathcal{I}| = k$, the term $\mathbb{E}[\phi_{\mathbf{1}}(\mathbf{x})x_\ell]$ in (7.11) may be of order $k+1$; hence, the resulting system of ODEs cannot be solved. In the above theorem, we derive bounds for moments that are of order $k+1$ by solving the finite-dimensional SDPs in (7.25). Notice that these SDPs involve, solely, moments of order up to k. Consequently, all the moments in the ODEs in Theorem 30 are of order less or equal to k, resulting in a closed system of differential equations. The theorem also states that, when the ODEs in Theorem 30 share the same initial conditions as the ODE in (7.11), the solutions are upper and lower bounds on the exact dynamics of $\mathbb{E}[\phi_{\boldsymbol{\alpha}}(\mathbf{x})]$. In the next section, we illustrate the proposed approach to perform a first-order moment-closure of the mean dynamics of the SIS spreading process.

7.2.3 First Order Moment-closure

In theory, we can upper and lower bound the dynamics of the mean spreading process in (7.2) by solving the ODEs in Theorem 30. In practice, these ODEs are solved via numerical methods using a discretized time interval. Notice that, according to Theorem 30, we need to solve the SDPs in (7.23) and (7.24) in each time step, which can be computationally very challenging in large-scale applications. To undertake this issue, we will develop a simplified procedure by finding a closed-form solution of the SDPs for the first-order mean dynamics. Our approach towards deriving a closed-form expression consists of two steps: First, we explicitly write the moment and localizing matrices in (7.18) for the first-order mean dynamics; then, we use a generalized version of Sylvester's criterion [168] to find a closed-form solution of the resulting SDPs.

When considering the first-order mean-dynamics in (7.2), we aim to derive upper and lower bounds on the second-order moments $\mu_{ij} = \mathbb{E}[x_i x_j]$ for $i \neq j$, in terms of first-order moments $\mu_i = \mathbb{E}[x_i]$. In this case, since $k = 2$, we have that $\mathcal{I}_2 = \{i, j\}$ (i.e., $i_1 = i$ and $i_2 = j$). Subsequently, the multi-sequence of interest $\mathbf{y}(\mathcal{I}_2)$ is given by $\mathbf{y}(\mathcal{I}_2) = \{1, \mu_i, \mu_j\}$, and its associated infinite extension is equal to $\mathbf{y}_\infty(\mathcal{I}_2) = \{1, \mu_i, \mu_j, \mu_{ij}, \ldots\}$. More specifically, in the multisequence $\mathbf{y}_\infty(\mathcal{I}_2)$, the entries are indexed as follows: $\mu_{\boldsymbol{\alpha}} = \mu_i$ if $\alpha_2 = 0$, $\mu_{\boldsymbol{\alpha}} = \mu_j$ if $\alpha_1 = 0$, and $\mu_{\boldsymbol{\alpha}} = \mu_{ij}$ otherwise. Subsequently, according to

(7.13), we have that

$$M_1(\hat{\mathbf{y}}(\mathcal{I}_2)) = \begin{bmatrix} 1 & \mu_i & \mu_j \\ \mu_i & \mu_i & \mu_{ij} \\ \mu_j & \mu_{ij} & \mu_j \end{bmatrix}. \tag{7.26}$$

Moreover, from (7.15) and (7.16), we construct the following four matrices:

$$\begin{aligned} L_1^0(\hat{\mathbf{y}}(\mathcal{I}_2), i) &= \begin{bmatrix} \mu_i & \mu_i & \mu_{ij} \\ \mu_i & \mu_i & \mu_{ij} \\ \mu_{ij} & \mu_{ij} & \mu_{ij} \end{bmatrix}, \\ L_1^0(\hat{\mathbf{y}}(\mathcal{I}_2), j) &= \begin{bmatrix} \mu_j & \mu_{ij} & \mu_j \\ \mu_{ij} & \mu_{ij} & \mu_{ij} \\ \mu_j & \mu_{ij} & \mu_j \end{bmatrix}, \end{aligned} \tag{7.27}$$

$$L_1^1(\hat{\mathbf{y}}(\mathcal{I}_2), i) = \begin{bmatrix} 1-\mu_i & 0 & \mu_j-\mu_{ij} \\ 0 & 0 & 0 \\ \mu_j-\mu_{ij} & 0 & \mu_j-\mu_{ij} \end{bmatrix}, \tag{7.28}$$

and

$$L_1^1(\hat{\mathbf{y}}(\mathcal{I}_2), j) = \begin{bmatrix} 1-\mu_j & \mu_i-\mu_{ij} & 0 \\ \mu_i-\mu_{ij} & \mu_i-\mu_{ij} & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{7.29}$$

Hence, according to Lemma 15, the sequence $\mathbf{y}_\infty(\mathcal{I}_2)$ has an $\tilde{S}_2$-representing measure with $\tilde{S}_2 = \left\{\mathbf{x} \in \mathbb{R}^2 \colon x_i, x_j \in [0,1]\right\}$, if and only if, (7.26)–(7.29) are all positive semidefinite. The main idea of our approach is to use a generalized version of Sylvester's criterion to replace the linear matrix inequalities in (7.25) by polynomial inequalities, as shown in the theorem below.

Theorem 31. *Consider a directed graph $G = (\mathcal{V}, \mathcal{E})$ and a set of n initial values $\{\mu_i(0)\}_{i=1}^n$. Let us define two sequences of functions $\{\hat{\mu}_i(t)\}_{i=1}^n$ and $\{\check{\mu}_i(t)\}_{i=1}^n$ satis-*

fying the following ODEs:

$$\frac{d\hat{\mu}_i}{dt} = -\delta_i \hat{\mu}_i + \sum_j \beta_{ij} \hat{\mu}_j - \sum_j \beta_{ij} \overline{\mu}_{ij},$$

$$\frac{d\check{\mu}_i}{dt} = -\delta_i \check{\mu}_i + \sum_j \beta_{ij} \check{\mu}_j - \sum_j \beta_{ij} \underline{\mu}_{ij},$$

with $\hat{\mu}_i(0) = \check{\mu}_i(0) = \mu_i(0)$, *where*

$$\overline{\mu}_{ij} = \max\{\hat{\mu}_i + \hat{\mu}_j - 1, 0\}, \tag{7.30}$$

$$\underline{\mu}_{ij} = \min\{\check{\mu}_i, \check{\mu}_j\}. \tag{7.31}$$

Then $\hat{\mu}_i(t) \geq \mu_i(t) \geq \check{\mu}_i(t)$, *for all* $t \geq 0$ *and* $i \in [n]$. ◇

Proof. See Appendix A.6. □

Several remarks are in order. First, note that we do not need to solve a semidefinite program to numerically find the upper and lower bounds stated in the above theorem. Instead, we need to solve a system of $2n$ piece-wise affine differential equations, where the piece-wise nonlinearities are described in (7.30) and (7.31). Furthermore, it is, in principle, possible to use the proposed approach to obtain a whole hierarchy of moment closures by considering higher-order moments. For example, we could derive a system of $n + m$ differential equations, where m is the number of edges in the graph, using both n first-order and m second-order moments. Finally, it is worth noting that the proposed technique can be generalized to analyze the mean dynamics of other spreading processes, as we illustrate in the next section.

7.3 Moment-closure of Other Popular Epidemic Model

In this section, we will apply the SDP-based moment closure framework herein proposed to find upper and lower bounds on the stochastic dynamics of two other networked

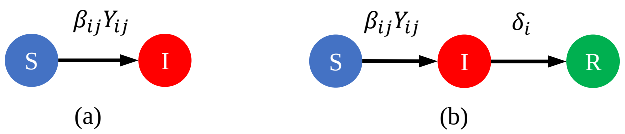


Figure 7-2: In (a) and (b), we show the transition between states in the SI-spreading process and SIR-spreading process, respectively.

epidemic models, namely, the SI and the SIR models.

7.3.1 Susceptible-Infected (SI) Epidemic Model

In the SI networked epidemic model [19], a susceptible node can be infected by its infected in-neighbors; however, once the node is infected, it remains infectious forever (see Figure 7-2-(a) for a detailed transition diagram). Let x_i be a binary random variable representing the state of node i, where $x_i(t) = 0$ if node i is susceptible at time t and $x_i(t) = 1$ if it is infected. The stochastic dynamics of the networked SI process can be modeled using the following jump-process:

$$dx_i = (1 - x_i) \sum_{j \in \mathcal{N}_i^-} x_j dP_{\beta_{ij}}. \tag{7.32}$$

Notice that this SDE is similar to (7.7), after removing the term describing the recovery process. Consequently, using the techniques used to prove Theorem 29, we can readily obtain the following ODE describing the evolution of any moment $\mu_{\mathcal{I}}(t)$, for any choice of $\mathcal{I} \subseteq [n]$:

$$\frac{d\mu_{\mathcal{I}}(t)}{dt} = \sum_{s=1}^{|\mathcal{I}|} \sum_{\ell \in \mathcal{N}_{i_s}^-} \beta_{i_s \ell} \left(\mu_{\mathcal{I} \cup \{\ell\} \backslash \{i_s\}} - \mu_{\mathcal{I} \cup \{\ell\}} \right). \tag{7.33}$$

Since the random variables defining the states of nodes in the network are binary, we can use the techniques used in the analysis of the SIS model to find upper and lower bounds in the moment dynamics. In particular, the finite-dimensional moment and localizing matrices proposed in Subsection 7.2.2 can be directly used in here. Thus, we can obtain the following corollary from Theorem 30.

Corollary 5. *Given a directed graph $G = (\mathcal{V}, \mathcal{E})$, let us define two sequences of functions $\{\hat{\mu}_{\mathcal{I}}(t)\}_{\mathcal{I}\subseteq[n],|\mathcal{I}|\leq k}$ and $\{\check{\mu}_{\mathcal{I}}(t)\}_{\mathcal{I}\subseteq[n],|\mathcal{I}|\leq k}$ satisfying the following ODE's:*

$$\frac{d\hat{\mu}_{\mathcal{I}}(t)}{dt} = \sum_{s=1}^{|\mathcal{I}|} \sum_{\ell\in\mathcal{N}_{i_s}^-} \beta_{i_s\ell} \left(\hat{\mu}_{\mathcal{I}\cup\{\ell\}\setminus\{i_s\}}(t) - \underline{\mu}_{\mathcal{I}\cup\{\ell\}}(t)\right)$$

$$\frac{d\check{\mu}_{\mathcal{I}}(t)}{dt} = \sum_{s=1}^{|\mathcal{I}|} \sum_{\ell\in\mathcal{N}_{i_s}^-} \beta_{i_s\ell} \left(\check{\mu}_{\mathcal{I}\cup\{\ell\}\setminus\{i_s\}}(t) - \overline{\mu}_{\mathcal{I}\cup\{\ell\}}(t)\right),$$

for all $\mathcal{I} \subseteq [n], |\mathcal{I}| \leq k$, where $\overline{\mu}_{\mathcal{I}\cup\{\ell\}}$ and $\underline{\mu}_{\mathcal{I}\cup\{\ell\}}$ are defined as in (7.23) *and* (7.24)*, respectively. If $\hat{\mu}_{\mathcal{I}}(0) \geq \mu_{\mathcal{I}}(0) \geq \check{\mu}_{\mathcal{I}}(0)$, then $\hat{\mu}_{\mathcal{I}}(t) \geq \mu_{\mathcal{I}}(t) \geq \check{\mu}_{\mathcal{I}}(t)$, for all $t \geq 0$ and $\mathcal{I} \in [n], |\mathcal{I}| \leq k$.* ◇

7.3.2 Susceptible-Infected-Removed (SIR) spreading process

In the case of SIR spreading process, nodes in G can be in one out of three states: *susceptible*, *infected*, or *removed*, at any time instance. A node is in the *removed* state when it has been infected in the past, it has recovered from the infection, and has developed permanent immunity to the disease (see Figure 7-2-(b) for a detailed transition diagram); hence, it cannot be infected again in the future. In what follows, we use $\{0, 1\}$-valued random variables $x_{i,S}(t), x_{i,I}(t)$, and $x_{i,R}(t)$ to indicate whether node i is susceptible, infected, or removed at time t, respectively. Since node i can only be in exactly one compartment at every time instance, we have that $x_{i,S}(t)+x_{i,I}(t)+x_{i,R}(t) = 1$ for all $t \geq 0$. With these definitions, the two transition probabilities among states are characterized by:

$$\begin{aligned} \mathbb{P}(x_{i,I}(t+h) = 1 \mid x_{i,S}(t) = 1) &= h \sum_{j\in\mathcal{N}_i^-} \beta_{ij} x_{j,I}(t) + o(h), \\ \mathbb{P}(x_{i,R}(t+h) = 1 \mid x_{i,I}(t) = 1) &= \delta_i h + o(h). \end{aligned} \tag{7.34}$$

We assume that a node is either susceptible or infected at time $t = 0$. The evolution of the network states are characterized by the following set of SDEs:

$$d\begin{bmatrix} x_{i,S} \\ x_{i,I} \\ x_{i,R} \end{bmatrix} = \begin{bmatrix} 0 \\ -x_{i,I} \\ x_{i,I} \end{bmatrix} dP_{\delta_i} + \sum_{j\in\mathcal{N}_i^-} \begin{bmatrix} -x_{i,S}x_{j,I} \\ x_{i,S}x_{j,I} \\ 0 \end{bmatrix} dP_{\beta_{ij}}, \tag{7.35}$$

for all $i \in [n]$, where all the Poisson jump-processes are independent. From (7.35), the expectation of the random variable $x_{i,S}$ satisfies

$$\frac{d\mathbb{E}[x_{i,S}(t)]}{dt} = \sum_{j\in\mathcal{N}_i^-} \beta_{ij}\mathbb{E}[x_{i,S}(t)x_{j,I}(t)]. \tag{7.36}$$

Therefore, in order to solve for $\mathbb{E}[x_{i,S}(t)]$, it is necessary to characterize the evolution of $\mathbb{E}[x_{i,S}x_{j,I}]$ over time.

In what follows, we apply the framework herein proposed to derive a closed system of ODEs bounding the mean SIR spreading process. We start by computing the mean dynamics of the SIR spreading process via Ito's formula, as follows.

Theorem 32. *Consider the networked SIR process described in* (7.35)*. Given the vectors $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma} \in \mathbb{N}^n$, define the monomial $\phi_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathbf{x}) = \Pi_{i=1}^n x_{i,S}^{\alpha_i} x_{i,I}^{\beta_i} x_{i,R}^{\gamma_i}$. Then,*

$$\begin{aligned} \frac{d\mathbb{E}[\phi_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathbf{x})]}{dt} &= -\sum_{s=1}^n \delta_s \mathbf{1}_{\beta_s\neq 0}\mathbb{E}[\phi_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathbf{x})] \\ &+ \sum_{s=1}^n \delta_s \mathbf{1}_{\beta_s=0\cap\gamma_s\neq 0}\mathbb{E}[\Pi_{k\in[n],k\neq\ell} x_{k,S}^{\alpha_k} x_{k,I}^{\beta_k} x_{k,R}^{\gamma_k} x_{\ell,S}^{\alpha_\ell} x_{\ell,I}] \\ &- \sum_{s=1}^n \sum_{\ell\in\mathcal{N}_s^-} \beta_{s\ell}\mathbf{1}_{\alpha_s\neq 0}\mathbb{E}[\phi_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathbf{x})x_{\ell,I}] \\ &+ \sum_{s=1}^n \sum_{\ell\in\mathcal{N}_s^-} \beta_{s\ell}\mathbf{1}_{\alpha_s=0\cap\beta_s\neq 0}\mathbb{E}[\phi_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathbf{x})x_{\ell,I}] \end{aligned} \tag{7.37}$$

Proof. See Appendix A.6. □

Hereafter, given two sets of indices $\mathcal{I}, \mathcal{J} \subseteq [n]$, we define

$$\mu_{\mathcal{I},\mathcal{J}} = \mathbb{E}[\Pi_{i\in\mathcal{I}} x_{i,S} \Pi_{j\in\mathcal{J}} x_{j,I}]. \tag{7.38}$$

In particular, when $\mathcal{I}$ (resp. $\mathcal{J}$) is a singleton, i.e., $\mathcal{I} = \{i\}$ (resp., $\mathcal{J} = \{j\}$), we also write $\mu_{\mathcal{I},\mathcal{J}} = \mu_{i,\mathcal{J}}$ (resp., $\mu_{\mathcal{I},\mathcal{J}} = \mu_{\mathcal{I},j}$). Letting $\boldsymbol{\gamma} = \mathbf{0}_n$ in (7.37), we obtain

$$\begin{aligned} \frac{d\mu_{\mathcal{I},\mathcal{J}}(t)}{dt} &= -\sum_{s\in\mathcal{J}} \delta_s \mu_{\mathcal{I},\mathcal{J}}(t) - \sum_{s\in\mathcal{I}} \sum_{\ell\in\mathcal{N}_s^-} \beta_{s\ell}\mu_{\mathcal{I},\mathcal{J}\cup\{\ell\}} \\ &+ \sum_{s\in\mathcal{J}\setminus\mathcal{I}} \sum_{\ell\in\mathcal{N}_s^-} \beta_{s\ell}\mu_{\mathcal{I},\mathcal{J}\cup\{\ell\}}, \end{aligned} \tag{7.39}$$

which depends only on moments of $x_{i,S}$ and $x_{i,I}$ for $i \in [n]$. In order to solve the above system of ODEs, we need to provide bounds on $\mu_{\mathcal{I},\mathcal{J}\cup\{\ell\}}$ using lower-order moments. To achieve this goal, similar to the case of SIS spreading process, we aim to construct moment and localizing matrices, as listed in (7.18). We consider a slight abuse of notations by replacing k in $\mathbf{y}(\mathcal{I}_k)$ with $|\mathcal{I}| + |\mathcal{J}|$, and $\mathbf{y}(\mathcal{I}_k)$ with

$$\mathbf{y}(\mathcal{I}, \mathcal{J}) = \{\mu_{\mathcal{I}',\mathcal{J}'}\}_{|\mathcal{I}'|+|\mathcal{J}'|<k}.$$

With this definition, we construct finite-dimensional matrices analogous to the ones in (7.13), (7.15), and (7.16) using elements in $\mathbf{y}_{\mathcal{I},\mathcal{J}}$ accordingly. For example, to close the first-order mean dynamics of the SIR spreading process, the moment matrix defined in (7.13) becomes:

$$M_1(\mathbf{y}(\{i\},\{j\})) = \begin{bmatrix} 1 & \mu_{i,\emptyset} & \mu_{\emptyset,j} \\ \mu_{i,\emptyset} & \mu_{i,\emptyset} & \mu_{\emptyset,j} \\ \mu_{\emptyset,j} & \mu_{i,j} & \mu_{\emptyset,j} \end{bmatrix}. \tag{7.40}$$

Since $x_{i,S}$ and $x_{i,I}$ are binary random variables for all $i \in \mathcal{V}$, Lemma 15 can be applied without loss of generality.

To provide upper and lower bounds for the moment $\mu_{\mathcal{I},\mathcal{J}\cup\{\ell\}}$, we build $2k+1$ matrices using elements in $\mathbf{y}(\mathcal{I}, \mathcal{J}\cup\{\ell\})$ and solve for the maximum and minimum value $\mu_{\mathcal{I},\mathcal{J}\cup\{\ell\}}$ such that those matrices are positive semidefinite. Denoting those extreme values by

$\underline{\mu}_{\mathcal{I},\mathcal{J}\cup\{\ell\}}$ and $\overline{\mu}_{\mathcal{I},\mathcal{J}\cup\{\ell\}}$, we have that $\mu_{\mathcal{I},\mathcal{J}\cup\{\ell\}} \in [\underline{\mu}_{\mathcal{I},\mathcal{J}\cup\{\ell\}}, \overline{\mu}_{\mathcal{I},\mathcal{J}\cup\{\ell\}]}$. Finally, we adopt a similar treatment to the localizing matrices as in (7.21) and (7.22), i.e., replacing the entries within localizing matrices by upper and lower estimates $\hat{\mu}_{\mathcal{I},\mathcal{J}}$ and $\check{\mu}_{\mathcal{I},\mathcal{J}}$. We use $\check{\mu}^{\star}_{\mathcal{I},\mathcal{J}\cup\{\ell\}}$ and $\hat{\mu}^{\star}_{\mathcal{I},\mathcal{J}\cup\{\ell\}}$ to denote the lower and upper estimates of $\mu_{\mathcal{I},\mathcal{J}}$ obtained by solving SDPs using modified localizing matrices. As a result, we obtain the following theorem for the networked SIR epidemic model:

Theorem 33. *Consider the networked SIR process described in* (7.35)*. Let us define a sequence of functions* $\{\hat{\mu}_{\mathcal{I},\mathcal{J}}(t), \check{\mu}_{\mathcal{I},\mathcal{J}}(t)\}_{\mathcal{I},\mathcal{J}\subseteq[n],|\mathcal{I}|\leq k}$ *satisfying the following ODEs:*

$$\begin{aligned}\frac{d\hat{\mu}_{\mathcal{I},\mathcal{J}}(t)}{dt} =& -\sum_{s\in\mathcal{J}} \delta_s \hat{\mu}_{\mathcal{I},\mathcal{J}}(t) - \sum_{s\in\mathcal{I}} \sum_{\ell\in\mathcal{N}_s^-} \beta_{s\ell}\underline{\mu}_{\mathcal{I},\mathcal{J}\cup\{\ell\}},\\ &+ \sum_{s\in\mathcal{J}\setminus\mathcal{I}} \sum_{\ell\in\mathcal{N}_s^-} \beta_{s\ell}\overline{\mu}_{\mathcal{I},\mathcal{J}\cup\{\ell\}},\end{aligned}$$

$$\begin{aligned}\frac{d\check{\mu}_{\mathcal{I},\mathcal{J}}(t)}{dt} =& -\sum_{s\in\mathcal{J}} \delta_s \check{\mu}_{\mathcal{I},\mathcal{J}}(t) - \sum_{s\in\mathcal{I}} \sum_{\ell\in\mathcal{N}_s^-} \beta_{s\ell}\overline{\mu}_{\mathcal{I},\mathcal{J}\cup\{\ell\}},\\ &+ \sum_{s\in\mathcal{J}\setminus\mathcal{I}} \sum_{\ell\in\mathcal{N}_s^-} \beta_{s\ell}\underline{\mu}_{\mathcal{I},\mathcal{J}\cup\{\ell\}},\end{aligned}$$

where

$$\overline{\mu}_{\mathcal{I},\mathcal{J}\cup\{\ell\}} = \begin{cases} \hat{\mu}_{\mathcal{I}\cup\{\ell\}}, & \text{if } |\mathcal{I}| + |\mathcal{J}\cup\{\ell\}| \leq k, \\ \hat{\mu}^{\star}_{\mathcal{I},\mathcal{J}\cup\{\ell\}}, & \text{otherwise,} \end{cases} \tag{7.41}$$

and

$$\underline{\mu}_{\mathcal{I},\mathcal{J}\cup\{\ell\}} = \begin{cases} \check{\mu}_{\mathcal{I},\mathcal{J}\cup\{\ell\}}, & \text{if } |\mathcal{I}| + |\mathcal{J}\cup\{\ell\}| \leq k, \\ \check{\mu}^{\star}_{\mathcal{I},\mathcal{J}\cup\{\ell\}}, & \text{otherwise,} \end{cases} \tag{7.42}$$

for all $\mathcal{I},\mathcal{J} \subseteq [n]$*. If* $\hat{\mu}_{\mathcal{I},\mathcal{J}}(0) \geq \mu_{\mathcal{I},\mathcal{J}}(0) \geq \check{\mu}_{\mathcal{I},\mathcal{J}}(0)$*, then* $\hat{\mu}_{\mathcal{I},\mathcal{J}}(t) \geq \mu_{\mathcal{I},\mathcal{J}}(t) \geq \check{\mu}_{\mathcal{I},\mathcal{J}}(t)$*, for all* $\mathcal{I},\mathcal{J} \subseteq [n]$ *and* $t \geq 0$.

Proof. See Appendix A.6. □

In the next section, we demonstrate the performance of the moment-closure framework herein proposed on both the SIS and SIR epidemic processes taking place in a real social network.

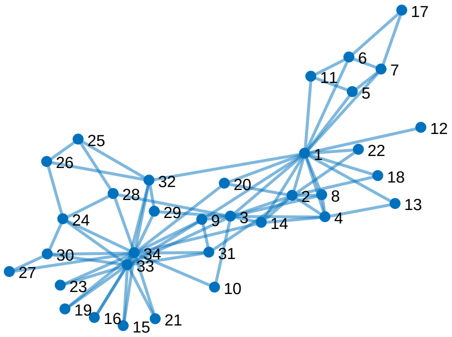


Figure 7-3: Topology of the Zachary's Karate Club, representing friendships among 34 individuals.

7.4 Simulation

In this section, we demonstrate the SDP-based moment-closure framework by finding upper and lower bounds on the probabilities of infection of all nodes in a real social network. In our first set of simulations (Subsection 7.4.1), we implement the exact stochastic SIS spreading process, as described in [127]. We simulate 10,000 realizations of the stochastic process using the same initial conditions and compute the evolutions of the empirical average of the probabilities of infection, which is an approximation of the mean SIS dynamics. We then execute our SDP-based moment-closure technique, using Theorem 31, in order to obtain the upper and lower bounds on the mean SIS dynamics, $\hat{\mu}_i(t)$ and $\check{\mu}_i(t)$. Furthermore, we compare the time evolution of these bounds with the widely used mean-field approximation (7.3). In our second set of simulations (Subsection 7.4.2), we apply similar analysis to the SIR spreading process.

7.4.1 Moment-closure of the SIS Epidemic Process

In this subsection, we run the stochastic SIS dynamics over the Zachary's Karate Club [178], plotted in Figure 7-3. In our experiments, we choose the individuals with labels $\mathcal{S} = \{3, 5, 6, 14, 16, 17, 20, 23\}$ to be initially infected. The infection rates satisfy $\beta_{ij} = \beta = 1$ for all $(i, j) \in \mathcal{E}$ and the recovery rates are $\delta_i = \delta = 7.4$ for all nodes.

According to [87], the expected number of infected individuals converges towards zero exponentially fast if $\tau = \frac{\beta}{\delta} < \frac{1}{\lambda_1(A)}$, where $\lambda_1(A)$ is the largest absolute eigenvalue of the adjacency matrix. In our case, the largest eigenvalue of the Zachary's network equals to $\lambda_1(A) = 6.7257$; hence, the condition $\tau < \frac{1}{\lambda_1(A)}$ is satisfied. As illustrated in Figure 7-4, the empirical average of number of infected nodes decreases exponentially over time. In Figure 7-5, we plot the evolution of the mean SIS dynamics of each node in the Zachary's network.Our simulations show the validity of the bounds obtained by our moment-closure framework.

7.4.2 Moment-closure of the SIR Epidemic Process

We proceed to demonstrate our SDP-based moment-closure scheme on the SIR model. In these experiments, we use again Zachary's network. In our simulations, we have selected the following set of initially infected nodes: $\mathcal{D} = \{5, 22, 28, 31, 32\}$; all remaining nodes are initially in the susceptible state. We set the infection rates to be $\beta_{ij} = \beta = 10$, whereas the recovery rates are $\delta_i = \delta = 6.7257$ for all nodes. Due to space limitations, we show in Figure 7-6 the evolution of $\{\hat{\mu}_{i,S}(t), \hat{\mu}_{i,I}(t), \hat{\mu}_{i,R}(t)\}$ and $\{\check{\mu}_{i,S}(t), \check{\mu}_{i,I}(t), \check{\mu}_{i,R}(t)\}$ for the nodes in the subset $\{2, 7, 22, 29\}$. Notice that the proposed moment-closure technique does indeed upper and lower bounds the true mean dynamics of the SIR model. Nonetheless, the performance of these bounds varies. For example, in Figure 7-6-(c), both bounds remain close to the true mean dynamics. However, in Figure 7-6-(a), the upper estimate $\hat{\mu}_{2,I}$ fails to keep track of the true evolution of $\mu_{2,I}(t)$. There are several possible reasons for this to happen. For example, as shown in (7.35), at every time instance, we have $\mu_{i,S}(t) + \mu_{i,I}(t) + \mu_{i,R}(t) = 1$; however, the proposed upper and lower estimates fail to preserve this property.

7.5 Coinfection Control in Multi-layer Networks

The preceding sections of this chapter mainly concerns about the dynamic behavior of single-disease processes in single-layer networks. In the following sections, we analyze

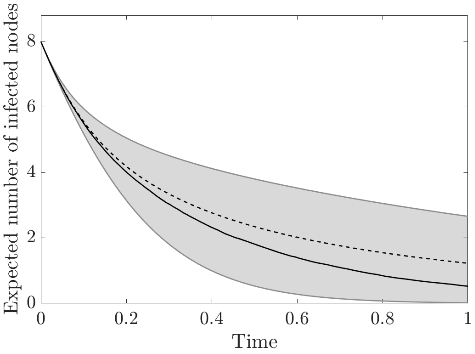


Figure 7-4: This figure depicts upper and lower bounds on the expected number of infected nodes. The solid black line represent the empirical average over 10000 realizations of the number of infected nodes over time. The dashed line and the shaded region represent the expected number of infected nodes calculated via the mean-field approximation and the SDP-based moment-closure technique, respectively.

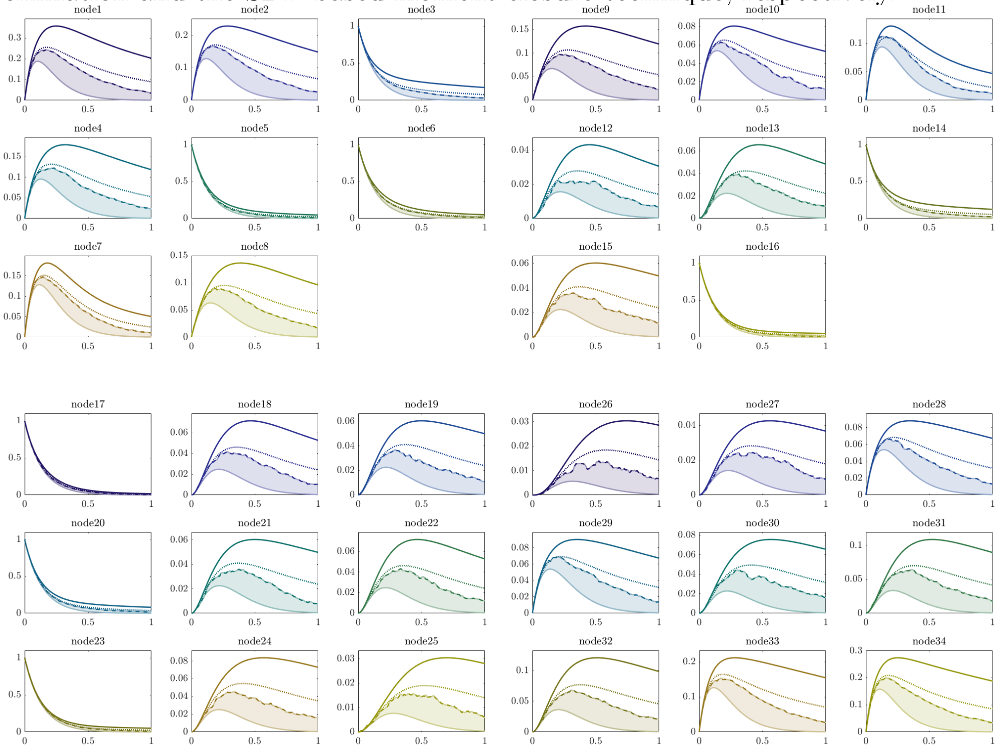


Figure 7-5: Dashed lines represent the empirical averages of 10,000 realizations of the stochastic SIS dynamics for each node i. The dotted lines represent the trajectories obtained from the mean-field approximation for each node. The solid lines represent $\hat{\mu}_i(t)$ and $\check{\mu}_i(t)$ for each node i. Finally, the shaded areas are filling the gap between the empirical average and $\check{\mu}_i(t)$ for each node i.

the problem of simultaneously controlling the spread of several diseases by distributing different types of vaccines throughout the nodes of a multilayer contact network (see

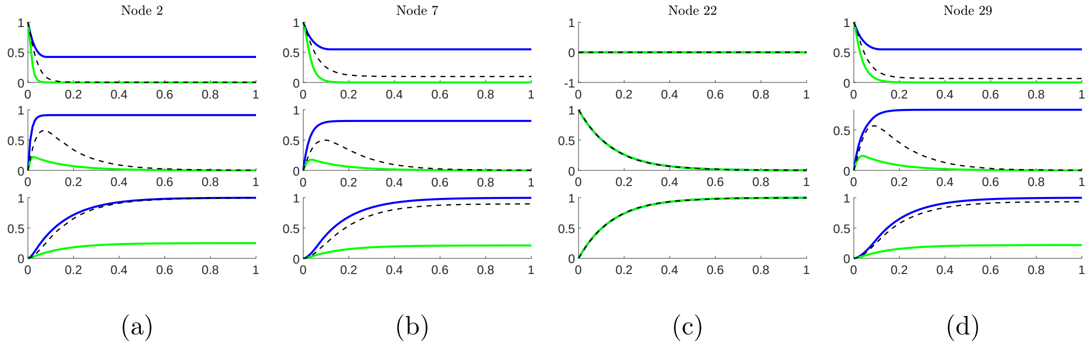


Figure 7-6: Dashed lines represent the average of 10,000 realizations of stochastic SIR dynamics for each node i. In subfigures (a)–(d), we show the evolution of $\hat{\mu}_{i,C}(t)$ and $\check{\mu}_{i,C}(t)$, where $C \in \{S, I, R\}$, for the nodes $i = 2, 7, 22, 29$, respectively. For instance, in (a), blue and green lines in each of the subplots (from up to down) show the evolution of $\{\hat{\mu}_{2,S}(t), \check{\mu}_{2,S}(t)\}$, $\{\hat{\mu}_{2,I}(t), \check{\mu}_{2,I}(t)\}$, and $\{\hat{\mu}_{2,R}(t), \check{\mu}_{2,R}(t)\}$, respectively.

Subsection 7.5.1). We based our work on the so-called SI_1SI_2S, analyzed by Sahneh and Scoglio in [143], which extends the popular SIS epidemic model to the case of competitive viruses in a two-layer network of identical agents. We further propose an extension of this model to the case of non-identical agents and an arbitrary number of layers and viruses, which we call $(SIS)^L$ spreading model (see Subsection 7.5.2). In our setting, we assume that we can modify the susceptibility of an individual to a particular disease by inoculating an antidote specifically design to fight that disease. In real applications, fabricating and distributing antidotes throughout a population has an associated cost. Hence, while facing simultaneous diseases, the agency responsible for disease control must decide how to invest its budget on the fabrication and allocation of different vaccines throughout the nodes and layers of the network (see Subsection 7.5.4). We obtain a global vaccination strategy using Geometric Programming (see Subsection 7.5.5). Additionally, we provide an alternative vaccination strategy based, solely, on local structural information of the network (Subsection 7.5.6). Finally, we illustrate the performance of global vaccination strategy and local vaccination strategy on the $(SIS)^L$ spreading process taking place in a synthesized network.

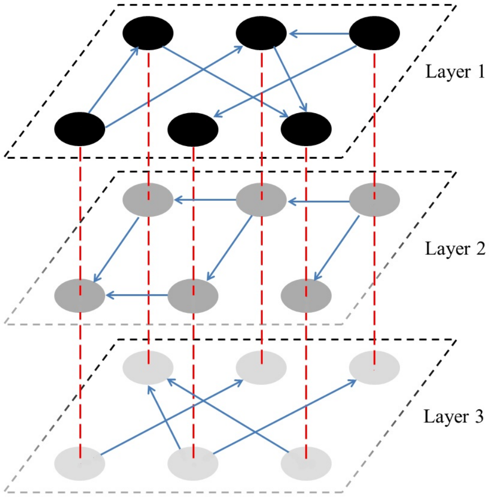


Figure 7-7: A three-layer network, $\mathcal{G} = \{\mathcal{V}, \mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3\}$.

7.5.1 Multi-layered Network

To formulate the problem of interest, we first adopt the notations in Subsection 2.1.2. Furthermore, in a directed graph $G(\mathcal{V}, \mathcal{E})$, we define the *out-neighborhood* of node i by $N_i^+ = \{j | j \in \mathcal{V}, (i,j) \in \mathcal{E}\}$. The *in-neighborhood* is similarly defined as $N_i^- = \{j | j \in \mathcal{V}, (j,i) \in \mathcal{E}\}$. The *out-degree* and *in-degree* of node i are defined as $d_i^+ = |N_i^+|$ and $d_i^- = |N_i^-|$, respectively.

We study spreading processes in multilayer networks. A multilayer network is represented by the tuple $\mathcal{G} = (\mathcal{V}, \mathcal{E}_1, \cdots, \mathcal{E}_L)$, where $\mathcal{V} = [n]$ is the set of nodes, and $\mathcal{E}_l$ is the set of edges corresponding to layer l. The edge-set $\mathcal{E}$ in a multilayer network is the union of edge sets $\mathcal{E}_l$ for $l \in [L]$. As in the case of simple graphs, the structure of layer $l \in [L]$ can be algebraically represented using the adjacency matrix of that layer, which we denote by $A_l = (a_{ij,l})_{ij}$. Furthermore, we define the *out-neighbor* and *in-neighbor* of node i in layer l as $N_{i,l}^+ = \{j | j \in \mathcal{V}, (i,j) \in \mathcal{E}_l\}$ and $N_{i,l}^- = \{j | j \in \mathcal{V}, (j,i) \in \mathcal{E}_l\}$, respectively.

7.5.2 Non-Homogeneous $(SIS)^L$ Spreading Model

In this section, we describe the non-homogeneous $(SIS)^L$ model. This model is an extension of the continuous-time competitive spreading model, denoted by SI_1SI_2S,

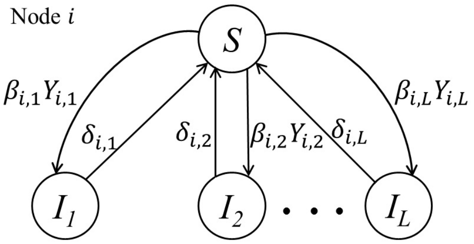


Figure 7-8: $(SIS)^L$ model for a particular node $i \in \mathcal{V}$.

was studied by Sahneh and Scoglio in [143]. In the SI_1SI_2S, two competing diseases propagate through a two-layer network. In the $(SIS)^L$ model, we consider L diseases propagating through different layers of a multilayer network. Fig. 1 shows the structure of a network with three spreading diseases. The state of each node in this model can fall into one of two cases: (i) '*Susceptible*' or healthy state, in which the node is not infected by any diseases, (ii) '*Infection*' state, in which the node is infected by one (and only one) of the L diseases propagating in the network. Fig. 2 represents the transition diagram, with $L+1$ states, for each node $i \in \mathcal{V}$. Whenever node i is in the susceptible state, it can be infected by the l-th diseases with a rate proportional to the infection rate $\beta_{i,l} > 0$, which is both node- and disease-dependent. In contrast, if node i is infected by disease l, it cures itself at rate $\delta_{i,l}$. Notice that, in this model, once a person is infected by one of the diseases, he/she is immune to other diseases until recovery.

The dynamics of such network can be modeled by a continuous-time Markov process. Following the modeling procedure in [143] for the case of a network with two layers, and extending it to multiple layers we can derive a mean-field approximation of the $(SIS)^L$ model, as follows. Let us denote by $p_{i,l}(t)$ the probability of node i being infected by disease l at time t. Hence, the mean-field approximation provides us the following set of ordinary differential equations:

$$\frac{dp_{i,l}\left(t\right)}{dt} = \beta_{i,l}(1 - \sum_{k=1}^{L} p_{i,k}(t)) \sum_{j=1}^{n} a_{ij,l} p_{j,l}\left(t\right) - \delta_{i,l} p_{i,l}\left(t\right), \tag{7.43}$$

for $i \in [n]$ and $l \in [L]$. This set of equations can be converted into a compact matrix form:

$$\frac{d\mathbf{p_l}\left(t\right)}{dt} = \left(B_l A_l - D_l\right)\mathbf{p_l}\left(t\right) - \mathbf{P}\left(t\right) B_l A_l \mathbf{p_l}\left(t\right), \tag{7.44}$$

for $l = 1, 2, \cdots, L$, where $\mathbf{p_l}\left(t\right) := \left(p_{1,l}\left(t\right), \ldots, p_{n,l}\left(t\right)\right)^T$, $B_l := diag(\beta_{i,l})$, $D_l := diag\left(\delta_{i,l}\right)$, and $\mathbf{P}\left(t\right) := diag(\sum_{l=1}^{L} p_{i,l})$.

7.5.3 Stability Analysis of Competitive Spreading Model

The nonlinear system of ODEs in (7.44) can be linearized around the disease-free equilibrium, $\mathbf{p_l^*} = \mathbf{0}, \forall l = 1, 2, \cdots, L$. We can therefore linearize (7.44) around this equilibrium, resulting in the following system of linear ODEs:

$$\frac{d\widetilde{\mathbf{p}}_l\left(t\right)}{dt} = \left(B_l A_l - D_l\right)\widetilde{\mathbf{p}}_l\left(t\right), \ \forall l \in [L]. \tag{7.45}$$

If all the eigenvalues of $B_l A_l - D_l$ lie in the open left half-plane, i.e., $\rho_l := \max\{\Re(\lambda_i(B_l A_l - D_l))\} < 0$ for all $l = 1, 2, \cdots, L$, the nonlinear dynamics is locally asymptotically stable. Furthermore, since $\mathbf{P}\left(t\right), B_l, A_l \geq 0$, then $\frac{d\mathbf{p_l}(t)}{dt} \leq \left(B_l A_l - D_l\right)\mathbf{p}_l\left(t\right)$. In other words, the linearized dynamics in (7.45) upper-bounds the nonlinear dynamics in (7.44) for all layers and identical initial conditions. Therefore, if $\rho_l < 0$ the nonlinear dynamics in (7.44) is not only locally, but also globally exponentially stable. Moreover, for $\mathbf{p_l}(0)$ close to the origin, $\mathbf{p_l} \to \mathbf{0}$ exponentially fast and the exponential decay rate at layer l is given by ρ_l.

7.5.4 Budget-constrained Vaccine Allocation Problem

In this subsection, we describe the problem of simultaneously controlling several diseases propagating in a multilayer network. We assume that we are able to distribute vaccines able to reduce the infection rate of disease l in node i, $\beta_{i,l}$. To make the modelling more realistic, we assume the feasible infection rates satisfy: $\beta_{i,l} \in \left[\underline{\beta}_{i,l}, \overline{\beta}_{i,l}\right]$. In addition, we also assign costs to these resources. We define $f_{i,l} : \mathbb{R}_+ \to \mathbb{R}_+$ as a node-dependent (indicated by subscript i) and disease-dependent (indicated by subscript l) *vaccination* *cost function*, such that $f_{i,l}(\beta)$ represents the cost of achieving an infection rate β in node i for disease l. We assume that the cost functions $f_{i,l}$ is monotonically decreasing with respect to β (i.e., the lower the β, the higher the cost). In what follows, we put ourselves in the position of the agency responsible for controlling all the diseases propagating in the network. Hence, given a fixed budget C, how should we invest on different vaccines to control all these different diseases? In this paper, we propose a convex optimization framework to find the optimal budget allocation to maximize the exponential decay rate of the slowest decaying disease. This problem can be formulated as follows:

Problem 12. *(*Budget-constrained allocation*) Given the following elements: (i) A multilayer graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}_1, \cdots, \mathcal{E}_L)$ (directed or undirected), where the adjacency matrix of layer l is A_l; (ii) a set of cost functions $\{f_{i,l}\}_{i\in[n],l\in[L]}$; (iii) limits on the infection rates $\underline{\beta}_{i,l}, \overline{\beta}_{i,l}$; and (iv) a total budget C. Find the optimal budget allocation to maximize the minimum exponential decay rate among all diseases.*

Next, we propose a convex formulation to efficiently solve this problem. As shown above, the linearized dynamics upper-bound the nonlinear dynamics and is a good approximation of the nonlinear dynamics for small densities of infection. Hence, the decaying rate of disease l is given by the eigenvalue with the largest real part of $B_l A_l - D_l$, which we denote by $\lambda_{\max}(B_l A_l - D_l)$. Hence, we reformulate Problem 12 as the following optimization:

$$
\begin{aligned}
&\underset{\{\varepsilon_l, \beta_{i,l}\}_{i\in[n]}^{l\in[L]}}{\text{minimize}} \max\{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_L\}\\
&\text{subject to } \lambda_{\max}(B_l A_l - D_l) \le \varepsilon_l,\ l \in [L]\\
&\sum_{l=1}^{L}\left(\sum_{i=1}^{n} f_{i,l}\left(\beta_{i,l}\right)\right) \le C,\\
&\underline{\beta}_{i,l} \le \beta_{i,l} \le \overline{\beta}_{i,l},\\
&\forall i = 1, \ldots, n,\ l = 1, \cdots, L.
\end{aligned}
$$

Notice that the optimal values for $\beta_{i,l}$ represent the achieved infection rates after the budget is allocated optimally. In the next three subsections, we reformulate the above as a convex optimization using Geometric Programming. Moreover, we also propose a relaxed formulation when global knowledge about the structure of the network is not available.

7.5.5 GP for Coinfection Control in Directed Networks

We now introduce some basic concepts about geometric programming that are useful in our formulation, which works for both directed and undirect networks. Let $x_1, \ldots, x_n > 0$ denote n decision variables and define $\mathbf{x} \triangleq (x_1, \ldots, x_n) \in \mathbb{R}^n_{++}$. A *monomial* $h(\mathbf{x})$ is defined as a function of the form $h(\mathbf{x}) \triangleq dx_1^{a_1} x_2^{a_2} \ldots x_n^{a_n}$, where $d > 0$ and $a_i \in \mathbb{R}$. A *posynomial* function $q(\mathbf{x})$ is defined as a non-negative sum of monomials: $q(\mathbf{x}) \triangleq \sum_{k=1}^{K} c_k x_1^{a_{1k}} x_2^{a_{2k}} \ldots x_n^{a_{nk}}$, where $c_k > 0$. There are a few algebraic manipulation properties related to posynomials. They are closed under addition, multiplication, and nonnegative scaling. In addition, a posynomial divided by a monomial is also a posynomial. A geometric program (GP) is an optimization problem of the form:

$$\begin{aligned} &\text{minimize } f(\mathbf{x}) \\ &\text{subject to } q_i(\mathbf{x}) \leq 1,\ i = 1, ..., m, \\ &\qquad\qquad h_i(\mathbf{x}) = 1,\ i = 1, ..., p, \end{aligned} \tag{7.46}$$

where q_i are posynomial functions, h_i are monomials, and f is a log-convex function. A GP is a quasi-convex optimization problem [109], by a change of variables, GP can be converted it into a convex optimization problem.

We make some assumptions on the structure of the network and the cost functions. First, we assume is that the adjacency matrices A_ls are strongly connected. Therefore, by *Perron-Frobenius* lemma [168], we have that ρ_l is a real, simple, and positive eigenvalue. Second, we also assume that the cost functions $f_{i,l}$ can be approximated using posynomials. By defining $\widehat{\delta}_{i,l} = 1 - \delta_{i,l}$ for all $i \in [n], l \in [L]$, hence ensuring entrywise nonnegativity of matrix $B_l A_l + I - D_l$, we can solve Problem 1, for directed networks, using the following formulation:

$$\underset{\{\rho_l, u_{i,l}, \beta_{i,l}\}_{i\in[n]}^{l\in[L]}}{min.} \epsilon \tag{7.47}$$

$$subject\ to\ \rho_l \leq \epsilon,\ l \in [L] \tag{7.48}$$

$$\frac{\beta_{i,l} \sum_{j=1}^{n} A_{ij,l} u_{j,l} + \widehat{\delta}_{i,l} u_{i,l}}{\rho_l u_{i,l}} \leq 1,\ l \in [L] \tag{7.49}$$

$$\sum_{l=1}^{L} \left(\sum_{i=1}^{n} f_{i,l}\left(\beta_{i,l}\right) \right) \leq C, \tag{7.50}$$

$$\underline{\beta}_{i,l} \leq \beta_{i,l} \leq \overline{\beta}_{i,l}, \tag{7.51}$$

$$\forall i \in [n], l \in [L] \tag{7.52}$$

The above formulation is a geometric program that can be solved using standard off-the-shelf software in polynomial time [132].

7.5.6 Local Policy for Coinfection Control

The previous approach is based on the assumption that a central agency have full information about the entire structure of the contact network. Nonetheless, in many practical situations, it is impossible to retrieve the complete network structure. In contrast, it is likely that each agent has access to a local, myopic view of its neighborhood, N_i. In this case, it is convenient to develop alternative disease-control strategies that make use of local information, solely. In this subsection, we propose such an approach, in which each agent in the network is able to decide about its protection level based solely on its local view of the network. The following definitions and results are relevant in our derivations:

Definition 25 (*Gershgorin disk*)**.** *Let $M = [m_{ij}]$ be a $n \times n$ matrix, and define $R_i := \sum_{j\neq i} |m_{ij}|$. Then, the* Gershogrin disks *are defined as the closed disks $D(m_{ii}, R_i) := \{z \in \mathbb{C} : |z - m_{ii}| \leq R_i\}$, for $i \in [n]$.*

Theorem 34 (*Gershgorin circle theorem,* [168])**.** *Every eigenvalue of M lies within at* *least one Gershgorin disk $D(m_{ii}, R_i)$, $i \in [n]$.*

Since M and M^T has a same set of eigenvalues. We have that every eigenvalue of M lies within at least one Gershogrin disk $D(m_{ii}, \hat{R}_i)$, with $\hat{R}_i = \sum_{i \neq j} |m_{ij}|$. Based on Gershgorin's theorem, we have the following result.

Theorem 35. *The linearized dynamics in* (7.45) *is stable if either (i)* $\sum_{j=1}^{n} \beta_{j,l} a_{ji,l} < \delta_{i,l}$*, or (ii)* $\beta_{i,l} d_{i,l}^{out} < \delta_{i,l}$ *are satisfied for all* $i = 1, 2, \ldots, n$ *and* $l = 1, 2, \cdots, L$*.*

Proof. The proof is a direct application of Theorem 34 to $M_l := B_l A_l - D_l$ for $B_l = diag(\beta_{i,l})$ and $D_l = diag\,(\delta_{i,l})$. □

Corollary 6. *Assume homogeneous infection rates, i.e.* $\beta_l = \beta_{i,l}$ *for all* $i = 1, 2, \ldots, n$*. Then, the linearized system (3) is stable if* $\beta_l d_{i,l}^{in} < \delta_{i,l}$ *holds for all* $i = 1, \cdots, n$ *and* $l = 1, \cdots, L$*, where* $d_{i,l}^{in}$ *is the* in-degree *of node* $i \in \mathcal{V}$ *at layer* l*.*

Proof. As shown in the above proof in Theorem 35, we have that $m_{ii,l} = -\delta_{i,l}$ and $R_{i,l} = \sum_{j, a_{ji,l} \neq 0} \beta_{j,l}$. Since $\beta_l = \beta_{j,l}$, then $R_{i,l} = \beta_{i,l} \sum_j a_{ji,l} = \beta_l d_{i,l}^{in}$. The rest follows from Theorem 35. □

From Theorem 35, whether the linearized dynamics in (7.45) is stable or not is directly linked to the number of *out-neighbors* of each node. This allow us to use only local information to distribute the resources (i.e. changing $\beta_{i,l}$ and $\delta_{i,l}$). Notice that in Theorem 35, inequalities are strict. To cope with strict inequalities, we can introduce slack variables $\epsilon > 0$ to convert into $\beta_{i,l} d_{i,l}^{out} + \epsilon \leq \delta_{i,l}$. Therefore, the problem proposed in section II-D can be relaxed into the following linear program.

$$\underset{\epsilon,\{\beta_{i,l}\}_{i\in[n]}^{l\in[L]}}{min.} \epsilon \tag{7.53}$$

$$subject\ to\ \delta_{i,l}-\epsilon \geq \sum_{j=1}^{n}\beta_{j,l}a_{ji,l} \tag{7.54}$$

$$\delta_{i,l}-\epsilon \geq \beta_{i,l}d_{i,l}^{out},\ \forall i,l \tag{7.55}$$

$$\sum_{l=1}^{L}\left(\sum_{i=1}^{n}f_{i,l}\left(\beta_{i,l}\right)\right)\leq C, \tag{7.56}$$

$$\underline{\beta}_{i,l}\leq\beta_{i,l}\leq\overline{\beta}_{i,l}, \tag{7.57}$$

$$\forall i=1,\ldots,n,\ l=1,\cdots,L \tag{7.58}$$

Notice that, when B_l is homogeneous, we can simplify equation (7.55) into $\delta_l-\epsilon\mathbf{1}\succeq\beta_l\mathbf{d}_l^{out}$ where $\mathbf{d}_l^{out}$ is the degree vector at layer l of the form $\left[d_{1,l}^{out},d_{2,l}^{out},\cdots,d_{n,l}^{out}\right]^T$ and $\succeq$ denotes component-wise inequality. Futhermore, if the underlying network is undirected, the *out-degrees* are the same as the *in-degrees* and hence (7.54) and (7.55) are equivalent. Notice that the feasibility of such formulation ensures the stability of the linearized system described in (7.45). However, we have no information on the location of the ρ_ls. Therefore, it is natural to see that increasing the budget limitation may or may not lead to a lower decay rate. Using such local distribution strategy, we have no guarantee on the performance of the decay rate of spreading process.

7.5.7 Simulations on Vaccination Strategies

In this subsection, through simulations, we will study performance of the convex formulations. In addition, we will compare the difference between global vaccination and local vaccination strategies.

Global vaccination Strategies: We illustrate the approaches in the previous subsection using a two-layer network with $n=50$ nodes. The first layer of the network is a circular network with each node connecting to its first and second neighbors, i.e,

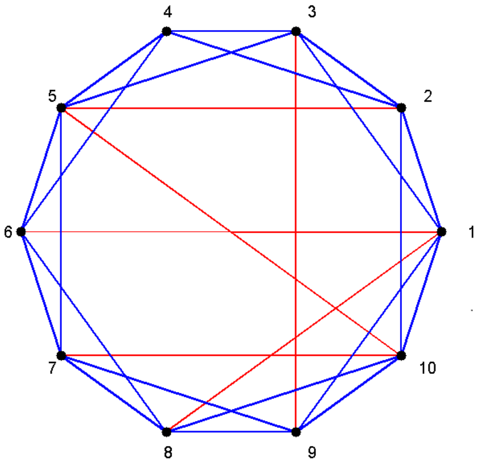


Figure 7-9: Graph visualization of A_1(shown in blue lines) and A_2(show in red lines) with 10 nodes.

$v_{i+1} \sim v_{\mod (i+1,n)+1}$ and $v_{i+1} \sim v_{\mod (i+2,n)+1}$ for $i = 0, 2, \cdots, n-1$. The second layer of the network is an *Erdös-Rényi* graph with edge-probability $p = 0.3$. A visualization of both layers with 10 nodes and $p = 0.1$ is depicted in Figure 7-9.

We assume that the infection rates are homogenous for both layers, thence $\beta_{i,1} = \beta_1$ and $\beta_{i,2} = \beta_2$, with $\beta_1 = 0.0625$ and $\beta_2 = 0.0181$. The curing rates $\delta_{i,l}$'s are non-homogenous for all $i \in [n]$ and $l = 1, 2$ and are drawn randomly in the range $[0.1, 0.3]$. In our example, after we sample $\delta_{i,l}$, the maximum eigenvalue of $\beta_l A_l - diag\{\delta_{i,l}\}$ are $\rho_1 \approx 0.072$ and $\rho_2 \approx 0.061$. In this case, since the maximum eigenvalues are positive, from Section II, we know that the uncontrolled $(SIS)^L$ system is unstable and an epidemic outbreak may happen.

We then use our formulation in the previous subsection to allocate vaccines. Figure 7-10 and Figure 7-11 show the optimal resource allocation using a total budget of $C = 5000$, and a vaccination cost function $f_{i,l} = \frac{1}{\beta_{i,l}}$ for all $i \in [n]$ and $l \in [L]$. We can expect that if one node is connected to more neighbors, it should have higher priority in getting antidotes to lower its infection rate. As shown in Figure 7-10, there is a negative correlation between degrees and optimal spreading rate with a linear correlation coefficient 0.2180. Nonetheless, the relation is non-trivial, as shown by the allocation result in layer 1 (the ring graph). Each of the node in layer 1 has exactly 4 neighbors, but

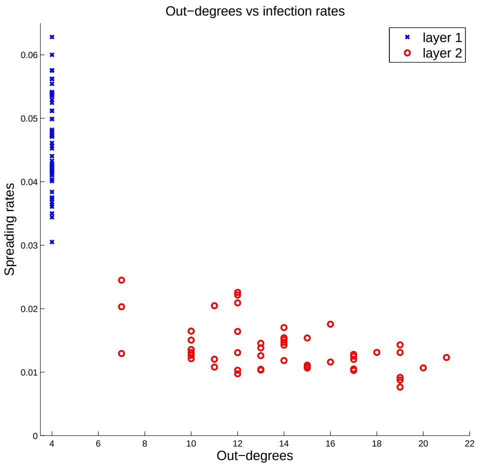


Figure 7-10: Optimal infection rates versus out-degrees for each one of the layers in a two-layer network.

the assigned infection rates varies drastically. In Figure 7-11, we plot the relationship between the optimal β's and eigen-centrality of each node. Similarly, we can observe that the higher the centrality of a node, the lower its infection rate. This relation is far from linear after we derive its linear correlation coefficient to be 0.2197 using linear regression. Therefore, to control multiple diseases within such framework, we cannot simply allocate vaccines to those who has large amount of neighbors (measured using degrees) or who has certain impact in the network (measured by eigen-centrality).

Local Vaccination Strategies: Hereafter, we compare the performance of the LP formulation based on local information to the global GP formulation. We use the same network and parameters described in the simulation for global vaccine strategies. In Figure 7-12, we plot the maximum decay rate obtained using both local and global strategies as we increase the available budget from $C = 1800$ to 8900. As expected, the higher the budget, the more the faster the diseases die out. Though no theoretical conclusion can be drawn on how suboptimal the local allocation is compared to the global one, we can see from Figure 7-12 that the local resource allocation approximates the global allocation for $C < 4000$. Nonetheless, using only degree information is not enough to described the topological structure of the network. Therefore, the rate achieved by the local strategy upper bounds the rate achieved by the global one. The tradeoff

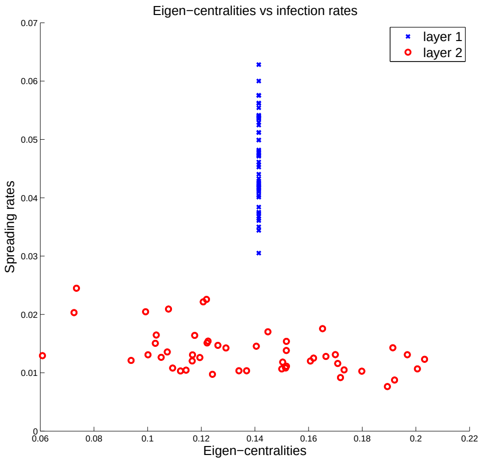


Figure 7-11: Optimal infection rates versus eigen-centralities for each one of the layers in a two-layer network.

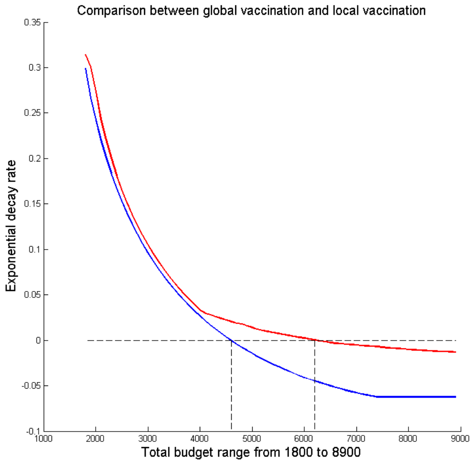


Figure 7-12: Comparison between global resource allocation and local resource allocation with varying budget limitations. The blue dots shows the maximum simultaneous decay rate of global allocation while red shows the maximum decay rate using local allocation

between global and local strategy is that, global method provides a more detailed plan for vaccination than the local one, but it requires more information about the network topology. In contrast, the local method is less efficient in controlling the outbreak, but requires only myopic information about the network structure.

Chapter 8

Applications of the K-moment Problem

In Chapter 6 and 7, we have leveraged the connection between semidefinite programming and multidimensional moment problem to bound the spectral radius of a simple directed graph as well as providing moment-closure of several popular networked spreading processes. In this section, we further leverage this connection to lower-bound the algebraic connectivity of an *undirected* graph (see Section 8.1) and characterize safety in nonlinear dynamical systems (see Section 8.2).

8.1 Lower Bound on the Graph Algebraic Connectivity

In this section, we will provide an lower bound on the algebraic connectivity of an undirected graph by leveraging the results related to the K-moment problem. Before explaining our method, we first introduce notions in graph theory that are useful in the development of our framework. Consider an undirected graph G, we define a diagonal matrix D by $D_{ii} = \sum_{j=1}^{n} A_{ij}$ and the *Laplacian matrix* of G by $L = D - A$. When G is undirected, L is a positive semidefinite matrix [168]. In what follows, we use $\lambda_1, \ldots, \lambda_n$ to denote the eigenvalues of L, and the *eigenvalue spectrum* of L is denoted by $\texttt{spec}(L) =$

$\{\lambda_i\}_{i=1}^n$. Since L is positive semidefinite, we assume that $0 = \lambda_1 \leq \lambda_2 \cdots \leq \lambda_n$ without loss of generality. In particular, λ_2 is called the *algebraic connectivity* of the graph G. Given a positive number α, we define the *perturbed Laplacian* by $L_\alpha = \alpha I - L$. Consequently, the spectrum of L_α is equal to $\texttt{spec}(L_\alpha) = \{\alpha - \lambda_n, \ldots, \alpha - \lambda_2, \alpha\}$. To ease notations, we also define $\tilde{\lambda}_i = \alpha - \lambda_{n-i+1}$ for $i \in [n]$, and define $I_\alpha = \alpha I - D$.

8.1.1 Moment-based lower bound on λ_2

In this subsection, we derive a relationship between closed walks in G and the power-sums of the eigenvalues in L_α. As a first step in our approach, we consider an undirected graph G with Laplacian matrix L, and adjacency matrix A. Given a positive value α and a positive integer k, we have that

$$\begin{aligned}
\mathrm{Tr}(L_\alpha^k) &= \mathrm{Tr}\left[(\alpha I - (D-A))^k\right] \\
&= \mathrm{Tr}\left[(I_\alpha + A)^k\right] \\
&= \mathrm{Tr}\left[\sum_{\ell=0}^{k} \binom{k}{\ell} I_\alpha^\ell A^{k-\ell}\right] \\
&= \sum_{\ell=0}^{k} \binom{k}{\ell} \mathrm{Tr}\left(I_\alpha^\ell A^{k-\ell}\right) \\
&= \sum_{\ell=0}^{k} \binom{k}{\ell} \sum_{i=1}^{n} \left((\alpha - d_i)^\ell [A^{k-\ell}]_{ii}\right).
\end{aligned} \tag{8.1}$$

Since $[A^k]_{ii}$ is equal to the total number of closed-walks of length k starting from vertex $i \in \mathcal{V}$, the above derivation shows that the closed walks in G are related to the power-sums of of eigenvalues in L_α.

We next provide a relationship between the closed walks in G and the moments of a probability measure defined on the eigenvalue spectrum of L_α. To achieve this goal, we first introduce some notions from probability theory that are crucial in the development of our framework. Consider an $\mathbb{R}$-valued random variable $x \sim \mu$, the k-th moment of the random variable x is defined as $m_k = \int_{\mathbb{R}} x^k d\mu$. Given an undirected graph G, we define the *auxiliary spectral distribution* as

$$\mu_{L_\alpha}(x) = \frac{1}{n-2}\sum_{i=1}^{n-2}\delta(x-\tilde{\lambda}_i), \tag{8.2}$$

where $\delta()$ is the Dirac's delta measure, i.e., the probability measure on $\mathbb{R}$ assigning unit mass to the origin, and zero elsewhere. In other words, the measure μ_{L_α} is a discrete probability measure assigning a mass $1/(n-2)$ to each one of the $n-2$ points in the set $\mathtt{spec}(L)\setminus\{\tilde{\lambda}_{n-1},\tilde{\lambda}_n\}$. Moreover, the k-th moment this measure is equal to

$$\begin{aligned} m_k &= \int_{\mathbb{R}} x^k d\mu_{L_\alpha} \\ &= \frac{1}{n-2}\sum_{i=1}^{n-2}\tilde{\lambda}_i^k \\ &= \frac{1}{n-2}\left[\operatorname{Tr}(L_\alpha^k) - \alpha^k - (\alpha-\lambda_2)^k\right]. \end{aligned} \tag{8.3}$$

Since L is positive semidefinite, if $\alpha \geq \lambda_n$, defining $\gamma = \tilde{\lambda}_{n-2}$, we obtain that L_α is also positive semidefinite. More precisely, $\mu_{L_\alpha}(x)$ is supported on $[0,\gamma]$. Consequently, the sequence of moments of $\mu_{L_\alpha}(x)$ must be $[0,\gamma]$-feasible. Therefore, finding the maximum value of γ for which the sequence of moments of μ_{L_α} is $[0,\gamma]$-feasible will give us an upper bound on $\tilde{\lambda}_{n-1}$. In what follows, we use the theory behind the K-moment problem to derive necessary conditions that must be satisfied for all K-feasible sequences $y_r = \{y_k\}_{k\leq r}$. In particular, the set K under consideration is the closed interval $[0,\gamma]$.

If y_r is the moment sequence for the auxiliary spectral distribution μ_{L_α}, then $y_k = \mathbb{E}_{\mu_{L_\alpha}}[x^k]$, for all $k \in \mathbb{N}$. Furthermore, as shown in (8.3), the following relationship has to be satisfied

$$y_k = \frac{1}{n-2}\left[\operatorname{Tr}(L_\alpha^k) - \alpha^k - \gamma^k\right], \forall k \in \mathbb{N}. \tag{8.4}$$

The moment matrix associated with y_r is equal to

$$[M_r]_{ij} = m_{i+j}, \text{ for } i,j \in \mathbb{N}_r. \tag{8.5}$$

Moreover, the measure μ_{L_α} is supported on interval $\mathcal{S} = [0,\gamma]$. By defining $g_1(x) = x$

and $g_2(x) = \gamma - x$, the set $\mathcal{S}$ can be characterized, equivalently by, $\mathcal{S} = \{x \in \mathbb{R} \colon g_1(x) \geq 0, g_2(x) \geq 0\}$. Subsequently, the localizing matrices of y_r with respect to g_1 and g_2, are equal to,

$$[L_r(g_1)]_{ij} = m_{i+j+1}, \tag{8.6}$$

and

$$[L_r(g_2)]_{ij} = \gamma m_{i+j} - m_{i+j+1}, \tag{8.7}$$

for $i, j \in \mathbb{N}_r$, respectively. According to Corollary 3 in Chapter 6, if y_r is a $[0, \gamma]$-feasible moment sequence for the auxiliary spectral distribution, the matrices (8.5)–(8.7) must be positive semidefinite. Therefore, we find an upper bound on $\tilde{\lambda}_{n-1}$ by finding the largest value of γ subject to these matrices being positive semidefinite, as described in the following theorem.

Theorem 36. *Let r be an arbitrary positive integer and $d = 2r + 1$. Denote by $\overline{\gamma}_r^\star$ the solution of the following semidefinite program:*

$$\begin{aligned} &\underset{\gamma, \mathbf{y}_d}{\text{maximize}}\ \gamma \\ &\text{subject to } M_r \succeq 0 \\ &L_r(g_i) \succeq 0, \text{ for all } i \in [2], \end{aligned} \tag{8.8}$$

where M_r and $L_r(g_i)$ are defined in (8.5)–(8.7)*. Let $\lambda_r^\star = \alpha - \overline{\gamma}_r^\star$, then, $\lambda_r^\star \leq \lambda_2$ for all $r \in \mathbb{N}$. Furthermore, $\lambda_r^\star$ is a non-decreasing function of $r \in \mathbb{N}$.*

Proof. The proof follows similar idea as in the proof of Theorem 23 from Chapter 6. □

8.1.2 Simulations

In this section, we empirically demonstrate the validity of our bounds on random undirected graphs and on real-world networks.

We first examine our bounds on Erdős-Rényi random graphs with $n = 1000$ nodes.

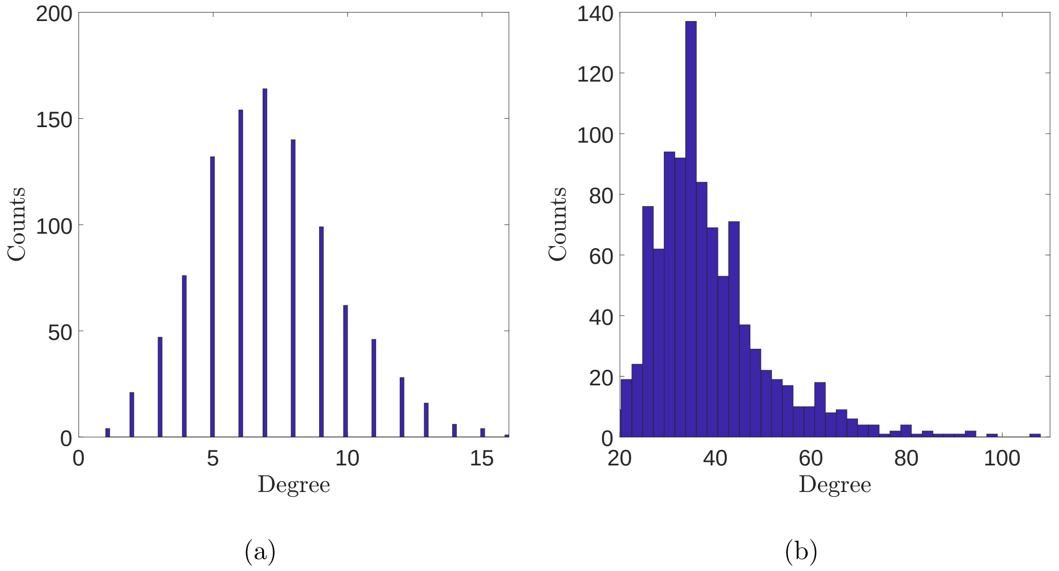
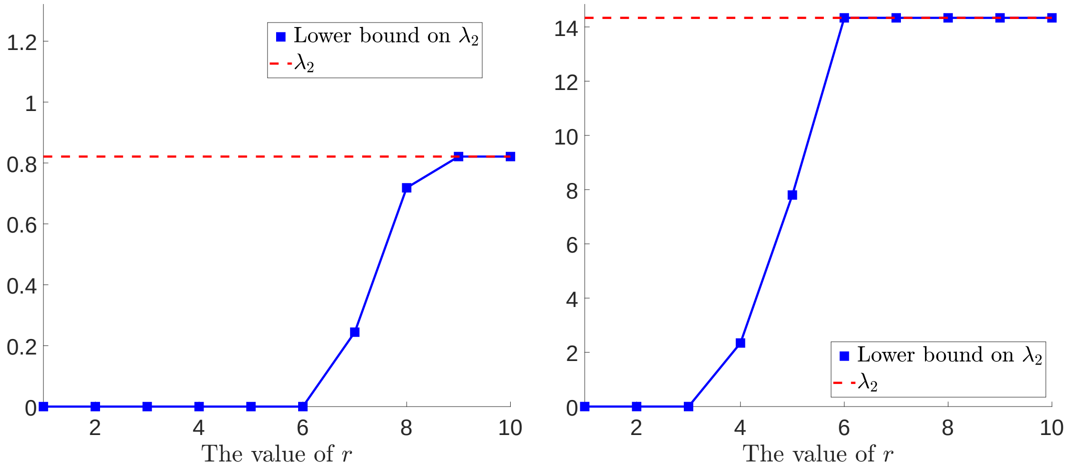


Figure 8-1: The degree distribution of a realization of the (a) Erdős-Rényi random graph with $n = 100$ and $p = 0.007$, and (b) Chung-Lu random graph with $n = 1000$.

We set the probability of edge to be $p = 0.007$. With these parameter settings, we generate a numerical realization of the random graph G with Laplacian L and the degree distribution of this graph in Figure 8-1-(a). In this case, the second smallest eigenvalue of L, i.e., the algebraic connectivity, is equal to 0.8211, whereas the edge-cut is equal to 0.8. In Figure 8-2, we show the evolution of lower bounds computed using Theorem 36 with different values of r. As shown in the figure, $\lambda_r^\star$ is equal to zero when $r \leq 6$. The intuition behind this is that, it is possible that there exists a disconnected graph G' with the same amount of closed-walks of length up to 12 as in G. In this case, the algebraic connectivity of G' is equal to 0. As r increases, we obtain information about long closed-walks in the graph, hence the quality of our lower bounds increase.

We next generate random graphs according to the Chung-Lu model [179]. As described in Section 6.4 of Chapter 6, it is possible to generate random graphs whose degree distribution obeys a power-law distribution. In this experiment, we consider the following parameters: $n = 1000$, $\beta = 5$, $d = 40$ and $\Delta = 100$. We obtain a sample from this ensemble of random graphs, whose algebraic connectivity is equal to $\lambda_2 \approx 14.3211$. We depict the degree-distribution of this particular sample in Figure 8-1-(b). Using Theorem 36, we obtain a sequence of lower bounds on λ_2 with varying r, as shown in Figure 8-2-(b).



(a)

(b)

Figure 8-2: The lower bounds on algebraic connectivity of a sample of (a) Erdős-Rényi, and (b) Chung-Lu random graph computed using Theorem 36 with varying r.

Different from the case of Erdős-Rényi random graph, the lower bound becomes tight when $r = 6$. Next, we explore our framework on a selected set of real-world networks obtained from [180]. In order to provide reasonable lower bounds, we assume that all networks have only one connected component, i.e., $\lambda_2 > 0$. In particular, we concentrate on finding the algebraic connectivity of the largest connected component in an undirected graph when it contains multiple connected components.

Type	**Size**	λ_2	r	$\lambda_r^\star$
Crime	439	0.00682	5	0.00680
Wikipedia	889	0.0822	8	0.0822
Social	1446	0.3219	5	0.3218
Social	2235	0.3726	10	0.3725
Protein	2224	0.0599	4	0.0150

Table 8.1: This table shows the true algebraic connectivity (column 2, denoted by λ_2), the first value of r at which $\lambda_r^\star > 0$ (column 3), and lower bounds on λ_2 computed using Theorem 36 (column 4, denoted by $\lambda_r^\star$), respectively.

8.2 Safety Verification of Non-linear Systems

The ability to provide safety certificates about the behavior of complex systems is critical in many engineering applications [181–185]. Although safety verification is a mature area with many success stories [186, 187], the verification of nonlinear dynamical systems over nonconvex unsafe regions remains a challenging problem [188, 189].

In the past decades, various solutions have been proposed to verify the safety of dynamical systems. The solution approaches often fall into the following two categories: (i) reachable set methods [190–192], and (ii) Lyapunov function methods [193–196]. Essentially, reachable set methods aim to find a set containing all possible states at a given time, for a given set of initial conditions. Subsequently, if the reachable set does not intersect with the pre-specified unsafe regions, the system is considered to be safe. For example, in [190] the reachable set is found for continuous-time linear systems, whereas in [191] and [192] the reachable sets are computed via approximations for nonlinear dynamical systems. In [197], the authors applied a reachable set method to plan safe trajectories for autonomous vehicles.

While reachable set methods can be used to obtain quantitative guarantees for safety, the reliability of the result largely depends on the assumptions made about the system, as well as the form of the unsafe regions. For instance, calculating the volume of the intersection of two sets, such as the reachable set and the unsafe regions, can become computationally challenging [189], jeopardizing the practical application of reachable set methods. An alternative approach to safety verification is based on using Lyapunov-like functions. In [194], the authors proposed the use of barrier certificates for safety verification of nonlinear systems. In contrast with the reachable set method, this line of work does not require to solve differential equations and is computationally more tractable. Furthermore, it also allows to provide safety certificates for various types of hybrid [193] and stochastic systems [195].

Despite a tremendous amount of solutions proposed to solve the safety verification problem, the majority of existing methods only provide binary safety certificates. More specifically, these certificates concern only *whether the system is safe* rather than *how safe the system is.* Lacking a detailed analysis of how unsafe a system is may result in a restricted and conservative design space. To illustrate this point, let us consider the operation of a solar-powered autonomous vehicle. Naturally, regions without solar exposure are considered to be unsafe, since the battery of the vehicle could be drained after a period of time. However, it would be inefficient to plan a path for the vehicle completely avoiding all these shaded regions. Instead, a more suitable requirement would be that the amount of time the vehicle spends in the shaded regions is bounded. More generally, this framework can be useful in those situations where the system is able to tolerate the exposure to a deteriorating agent, such as excessive heat or radiation, for a limited amount of time.

In this section, we consider this alternative, more flexible notion of safety. More precisely, we aim to compute the time that a (nonlinear) system spends in the unsafe regions. In particular, we focus our analysis on the case of systems described by a polynomial dynamics and unsafe regions described by a collection of polynomial inequalities. To calculate the amount of time spent in the unsafe regions, we use *occupation measures* to quantify how much time the system trajectory spends in a particular set [198]. Using this alternative viewpoint of the system dynamics, the safety quantity of interest can be calculated by finding the volume of the unsafe region with respect to the occupation measure [199]. The usage of occupation measures allows us to leverage powerful numerical procedures developed in the context of control of polynomial systems [200–202]. More specifically, we will show that the safety notion under consideration is the solution of an infinite-dimensional LP. Furthermore, we provide a hierarchy of relaxations that can be efficiently solved using semidefinite programming by leveraging the results in the K-moment problem.

In the following subsections, we use $\delta_{\mathbf{x}}$ to denote the Dirac measure centered on a fixed point $\mathbf{x} \in \mathbb{R}^n$ and we use $\otimes$ to denote the product between two measures. The ring of polynomials in $\mathbf{x}$ with real coefficients is denoted by $\mathbb{R}[\mathbf{x}]$, and $\mathbb{R}[\mathbf{x}]_r$ denotes the subset of polynomials of degree up to r.

8.2.1 Problem Statement

We consider a continuous-time autonomous dynamical system whose dynamics is captured by the following equation:

$$\begin{aligned} \dot{\mathbf{x}}(t) &= f(t, \mathbf{x}), \quad t \in [0, T] \\ \mathbf{x}(0) &= \mathbf{x}_0 \end{aligned} \tag{8.9}$$

where $\mathbf{x}(t) \in \mathbb{R}^n$ is the state vector, $\mathbf{x}_0$ is the initial condition, and $T > 0$ is the terminal time. We consider that the states of (7.2) are constrained to live within the set $\mathcal{X} \subseteq \mathbb{R}^n$ for all $t \in [0, T]$. Furthermore, we consider that the system evolves from an initial condition $\mathbf{x}_0$, with $\mathbf{x}_0 \in \mathcal{X}_0 \subseteq \mathcal{X}$. In this paper, we are interested in the case that the set $\mathcal{X}$ is semi-algebraic (see Definition 6.10 in Chapter 6). According to the definition, the set $\mathcal{X}$ can be defined using polynomials $g_i^{\mathcal{X}}(\mathbf{x}) \in \mathbb{R}[\mathbf{x}]$, as follows:

$$\mathbf{x}(t) \in \mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n \mid g_i^{\mathcal{X}}(\mathbf{x}) \geq 0, \forall i \in [n_{\mathcal{X}}]\} \tag{8.10}$$

for all $t \in [0, T]$. In this subsection, we consider the following problem:

Problem 13. *Consider a* compact *and* semi-algebraic *set $\mathcal{X}$, defined by* (8.10)*, and $\mathcal{X}_u \subseteq \mathcal{X}$, defined by:*

$$\mathcal{X}_u = \{\mathbf{x} \in \mathbb{R}^n \mid g_i^{\mathcal{X}_u}(\mathbf{x}) \geq 0, \forall i \in [n_{\mathcal{X}_u}]\}. \tag{8.11}$$

Given the autonomous system described in (8.9)*, with $x_0 \sim \mu_0(\mathcal{X}_0)$, where μ_0 is a probability distribution supported on $\mathcal{X}_0$, compute the expected amount of time that the system trajectory spends in the unsafe region $\mathcal{X}_u$.*

Notice that this expected time can be computed as:

$$\mathbb{E}\left[\int_0^T \mathbf{1}_{\mathcal{X}_u}(\mathbf{x}(t))dt\right], \tag{8.12}$$

where the expectation in (8.12) is taken with respect to the distribution of the initial condition $\mathbf{x}_0$. We remark that the above formulation is also capable of providing safety certificate for the system when the initial state is known exactly, i.e., $\mu_0 = \delta_{\mathbf{x}_0}$.

8.2.2 Occupation Measure-based Reformulation

In this section, we introduce a measure-theoretic approach to characterize the trajectories of the autonomous system described in (8.9) presented in Subsection 8.2.2.2. Using this method, we show that the expectation in (8.12) can be computed via an infinite-dimensional linear program – see Subsection 8.2.2.3 and Subsection 8.2.2.4. To explain our approach, we first introduce some notions of measure theory.

8.2.2.1 Notations and Preliminaries

Given a topological space $\mathcal{S}$, we denote by $\mathcal{M}(\mathcal{S})$ the space of finite signed Borel measures on $\mathcal{S}$, and $\mathcal{M}_+(\mathcal{S})$ its positive cone. Let $\mathcal{C}(\mathcal{S})$ and $\mathcal{C}^1(\mathcal{S})$ be the space of continuous functions and continuously differentiable functions on $\mathcal{S}$, respectively. The topological dual of $\mathcal{M}(\mathcal{S})$ and $\mathcal{C}(\mathcal{S})$ are denoted by $\mathcal{M}(\mathcal{S})^*$ and $\mathcal{C}(\mathcal{S})^*$.

Given a function $h \in \mathcal{C}(\mathcal{S})$ and a measure $\mu \in \mathcal{M}(\mathcal{S})$, we define the duality bracket between h and μ by

$$\langle h, \mu \rangle = \int_{\mathcal{S}} h d\mu. \tag{8.13}$$

By Riesz-Markov-Kakutani representation theorem [203], when $\mathcal{S}$ is locally compact Hausdorff, the dual space of $\mathcal{C}(\mathcal{S})$ is $\mathcal{M}(\mathcal{S})$, in which the norm of $\mathcal{C}(\mathcal{S})$ is the sup-norm of functions and the norm of $\mathcal{M}(\mathcal{S})$ is the total variation norm of measures. In the rest of the paper, we consider compact topological spaces $\mathcal{S} \subseteq \mathbb{R}^n$. As a consequence, both local compactness and separability conditions required to form the duality between $\mathcal{M}(\mathcal{S})$ and $\mathcal{C}(\mathcal{S})$ are satisfied. Given a measure $\mu \in \mathcal{M}(\mathcal{S})$, the support of μ, denoted by $\texttt{supp}(\mu)$, is the smallest closed set $C \subseteq \mathcal{S}$ such that $\mu(\mathcal{S} \setminus C) = 0$ where smallest is understood in the set-inclusion sense.

8.2.2.2 Occupation Measure and Liouville's Equation

Given an initial condition $\mathbf{x}_0$, let $\mathbf{x}(t \mid \mathbf{x}_0)$ be the solution to (8.9). Given a trajectory $\mathbf{x}(t \mid \mathbf{x}_0)$, we define the *occupation measure* $\mu(\cdot \mid \mathbf{x}_0)$ of $\mathbf{x}(t \mid \mathbf{x}_0)$ as

$$\mu(A \times B \mid \mathbf{x}_0) = \int_{[0,T] \cap A} \mathbf{1}_B(\mathbf{x}(t \mid \mathbf{x}_0))dt \tag{8.14}$$

for all $A \times B \subseteq [0,T] \times \mathcal{X}$. Therefore, given sets A and B, the value $\mu(A \times B)$ equals the total amount of time out of A that the state trajectory $\mathbf{x}(t \mid \mathbf{x}_0)$ spends in the set B. Similarly, we define the *final measure* $\mu_T(\cdot \mid \mathbf{x}_0)$ as

$$\mu_T(B \mid \mathbf{x}_0) = \mathbf{1}_B(\mathbf{x}(T \mid \mathbf{x}_0)) \tag{8.15}$$

for $B \subseteq \mathcal{X}$. Notice that the occupation measure $\mu(\cdot \mid \mathbf{x}_0)$ is supported on $[0,T] \times \mathcal{X}$ whereas the final measure $\mu_T(\cdot \mid \mathbf{x}_0)$ is supported on $\mathcal{X}$.

Given a test function $v \in \mathcal{C}^1([0,T] \times \mathcal{X})$, we define the operator $\mathcal{L}$ as:

$$v \mapsto \mathcal{L}v = \frac{\partial v}{\partial t} + \nabla v \cdot f(t, \mathbf{x}). \tag{8.16}$$

The *adjoint operator* $\mathcal{L}^* : \mathcal{M}([0,T] \times \mathcal{X}) \to \mathcal{C}^1([0,T] \times \mathcal{X})^*$ is given by

$$\langle v, \mathcal{L}^* \nu \rangle = \langle \mathcal{L}v, \nu \rangle. \tag{8.17}$$

From (8.16), we have that

$$\begin{aligned} v(T, \mathbf{x}(T \mid \mathbf{x}_0)) &= v(0, \mathbf{x}_0) + \int_0^T \frac{d}{dt} v(t, \mathbf{x}(t \mid \mathbf{x}_0))dt \\ &= v(0, \mathbf{x}_0) + \int_{[0,T] \times \mathcal{X}} \mathcal{L}v(t, \mathbf{x}) d\mu(t, \mathbf{x} \mid \mathbf{x}_0) \\ &= v(0, \mathbf{x}_0) + \langle \mathcal{L}v, \mu(\cdot \mid \mathbf{x}_0) \rangle. \end{aligned} \tag{8.18}$$

Hence, we can further rewrite (8.18) as

$$\langle v, \delta_T \otimes \mu_T(\cdot \mid \mathbf{x}_0)\rangle = \langle v, \delta_0 \otimes \delta_{\mathbf{x}_0}\rangle + \langle \mathcal{L}v, \mu(\cdot \mid \mathbf{x}_0)\rangle. \tag{8.19}$$

In the view of (8.17), since the above equation holds for all $v \in \mathcal{C}^1([0,T] \times \mathcal{X})$, we obtain the following equality:

$$\delta_T \otimes \mu_T(\cdot \mid \mathbf{x}_0) = \delta_0 \otimes \delta_{\mathbf{x}_0} + \mathcal{L}^* \mu(\cdot \mid \mathbf{x}_0). \tag{8.20}$$

Essentially, (8.20) describes the evolution of the distribution of states, given an initial distribution, under the flow of the dynamics (8.9) – see [204] for a more detailed discussions.

The measures defined in (8.14) and (8.15) depend on a given initial condition $\mathbf{x}_0$. In what follows, we extend these definitions to handle the case when the system is evolving from a set of possible initial conditions. Given an initial distribution μ_0 with $\mathtt{supp}(\mu_0) \subseteq \mathcal{X}_0$, we define the *average occupation measure* $\mu \in \mathcal{M}([0,T] \times \mathcal{X})$ as

$$\mu(A \times B) = \int_{\mathcal{X}_0} \mu(A \times B \mid \mathbf{x}_0) d\mu_0 \tag{8.21}$$

and the *average final measure* $\mu_T \in \mathcal{M}(\mathcal{X})$ as

$$\mu_T(B) = \int_{\mathcal{X}_0} \mu_T(B \mid \mathbf{x}_0) d\mu_0. \tag{8.22}$$

By integrating the left- and right-hand side of (8.18) with respect to μ_0, we have that

$$\delta_T \otimes \mu_T = \delta_0 \otimes \mu_0 + \mathcal{L}^* \mu. \tag{8.23}$$

Note that any family of solutions $\mathbf{x}(t)$ of (8.9) with an initial distribution μ_0 induces an occupation measure (8.21) and a final measure (8.22) satisfying (8.23). Conversely, for any tuple of measures (μ_0, μ, μ_T) satisfying (8.23), one can identify a distribution on the admissible trajectories starting from μ_0 whose average occupation measure and average final measure coincide with μ and μ_T, respectively (see Lemma 3 in [201] and Lemma 6 in [205] for more details).

8.2.2.3 Infinite-dimensional Linear Program Reformulation

Hereafter, we will show that the value in (8.12) can be obtained by solving a linear program on the occupation measure and the final measure, defined in (8.21) and (8.22). According to the definition of average occupation measure, we have that

$$\begin{aligned}\mathbb{E}\left[\int_0^T \mathbf{1}_{\mathcal{X}_u}(\mathbf{x}(t))dt\right] &= \int_{\mathcal{X}_0}\int_0^T \mathbf{1}_{\mathcal{X}_u}(\mathbf{x}(t))dtd\mu_0 \\ &= \int_{\mathcal{X}_0} \mu([0,T]\times\mathcal{X}_u \mid \mathbf{x}_0)d\mu_0 \\ &= \mu([0,T]\times\mathcal{X}_u).\end{aligned} \tag{8.24}$$

Leveraging the above measure-theoretical formulation, the value in (8.12) is equal to

$$\mu([0,T]\times\mathcal{X}_u). \tag{8.25}$$

Subsequently, finding the solution to Problem 13 is equivalent to finding the *volume* of the set $[0,T]\times\mathcal{X}_u$, where this volume is measured using the average occupation measure, instead of the Lebesgue measure. Next, we show that the value of (8.25) can be obtained by solving the following optimization problem: Given a polynomial $g : [0,T]\times\mathcal{X}\to\mathbb{R}$, such that $g(t,\mathbf{x}) > 0, \forall(t,\mathbf{x}) \in [0,T]\times\mathcal{X}_u$, consider the following optimization problem

$$\begin{aligned}\mathtt{P}:\ &\sup \int g d\widetilde{\mu} \\ &\text{subject to } \widetilde{\mu}+\widehat{\mu}=\mu \\ &\delta_T\otimes\mu_T = \delta_0\otimes\mu_0+\mathcal{L}^*\mu \\ &\mu,\widehat{\mu}\in\mathcal{M}_+([0,T]\times\mathcal{X}) \\ &\widetilde{\mu}\in\mathcal{M}_+([0,T]\times\mathcal{X}_u) \\ &\mu_T\in\mathcal{M}_+(\mathcal{X})\end{aligned} \tag{8.26}$$

where the supremum is taken over a tuple of measures $(\widetilde{\mu}, \widehat{\mu}, \mu, \mu_T) \in \mathcal{M}_+([0,T] \times \mathcal{X}_u) \times \mathcal{M}_+([0,T] \times \mathcal{X}) \times \mathcal{M}_+([0,T] \times \mathcal{X}) \times \mathcal{M}_+(\mathcal{X})$. The constraint $\widetilde{\mu} + \widehat{\mu} = \mu$ is equivalent to $\widetilde{\mu} \leq \mu$, i.e., the measure $\widetilde{\mu}$ is dominated by μ. Using duality brackets, we can write the objective in (8.26) as $\langle g, \widetilde{\mu} \rangle$. It follows that (8.26) is a linear program in the decision variable $(\widetilde{\mu}, \widehat{\mu}, \mu, \mu_T)$. Denote by $\sup \mathtt{P}$ the optimal value of $\mathtt{P}$ and by $\max \mathtt{P}$ the supremum attained. When $g \equiv 1$, we show below that the optimal value to the above program, if it exists, is equal to (8.12).

Theorem 37. *Let $\mathcal{X}_u$ be a compact and semi-algebraic subset of $\mathcal{X}$ and $\mathcal{B}$ be the Borel σ-algebra of Borel subsets of $[0,T] \times \mathcal{X}$. Let $\widetilde{\mu}^* \in \mathcal{M}([0,T] \times \mathcal{X}_u)$ be defined by*

$$\widetilde{\mu}^*(S) = \mu(S \cap [0,T] \times \mathcal{X}_u), \forall S \in \mathcal{B}. \tag{8.27}$$

Given a polynomial $g : [0,T] \times \mathcal{X} \to \mathbb{R}$, if $g(t, \mathbf{x}) > 0, \forall (t, \mathbf{x}) \in [0,T] \times \mathcal{X}_u$, then $\widetilde{\mu}^*$ is the $\widetilde{\mu}$-component of an optimal solution to $\mathtt{P}$. Furthermore, $\sup \mathtt{P} = \max \mathtt{P} = \int g d\widetilde{\mu}^*$. In particular, if $g \equiv 1$, then $\max \mathtt{P} = \mu([0,T] \times \mathcal{X}_u)$.

Proof. See Appendix A.7. □

As a result of Theorem 37, the solution of $\mathtt{P}$ is equal to the expected time in (8.12). In the next subsection, we consider the Lagrangian dual of $\mathtt{P}$.

8.2.2.4 Dual Infinite-dimensional Program

As mentioned in Section 8.2.2.1, the dual space of $\mathcal{M}(\mathcal{S})$ is the Banach space of continuous functions on $\mathcal{S}$ with the sup-norm. Let $\mathcal{C}_+(\mathcal{S}) \subseteq \mathcal{C}(\mathcal{S})$ be the set of continuous functions that are nonnegative on $\mathcal{S}$. Using duality theory, the dual program of (8.26) is equal to

$$
\begin{aligned}
\mathtt{D} : \inf_{v,w} & \int v(0,\mathbf{x}) d\mu_0 \\
& \text{s.t. } w(t,\mathbf{x}) - g(t,\mathbf{x}) \geq 0, \forall (t,\mathbf{x}) \in [0,T] \times \mathcal{X}_u \\
& -\mathcal{L}v(t,\mathbf{x}) - w(t,\mathbf{x}) \geq 0, \forall (t,\mathbf{x}) \in [0,T] \times \mathcal{X} \\
& v(T,\mathbf{x}) \geq 0, \forall \mathbf{x} \in \mathcal{X} \\
& w(t,\mathbf{x}) \geq 0, \forall (t,\mathbf{x}) \in [0,T] \times \mathcal{X}
\end{aligned}
\tag{8.28}
$$

where the decision variables in the above program are the continuously differentiable function $v(t,\mathbf{x}) \in \mathcal{C}^1([0,T] \times \mathcal{X})$ and the continuous function $w(t,\mathbf{x}) \in \mathcal{C}([0,T] \times \mathcal{X})$. The dual problem D always provides an upper bound on the optimal value of the primal P. In the sequel, we show that the optimal values of (8.26) and (8.28) are actually equal. Thus, *strong duality* holds in this infinite-dimensional linear program.

Theorem 38. *Let $p^\star$ and $d^\star$ be the optimal values of* P *and* D*, respectively. Then, $p^\star = d^\star$, i.e., there is no duality gap between* P *and* D*.*

Proof. See Appendix A.7. □

Consequently, the value of (8.12) can be obtained by solving (8.26) or (8.28). However, these two optimization problems are taking arguments from a tuple of measures or a tuple of continuous functions; hence both programs are hard infinite-dimensional optimization problems. In the next section, we leverage recent results from the multi-dimensional moment problem [108] to approximate the solution to (8.26). Furthermore, we show that it is possible to obtain increasingly tighter bounds on (8.25) by solving a sequence of semidefinite programs.

8.2.3 Semidefinite and Sum-of-Squares Relaxation

In the previous section, we have shown that (8.12) can be computed by solving an infinite-dimensional linear program. Although the optimal solutions to P or D provide exact solutions to Problem 1, it is computationally intractable to solve them. To address this issue, in Subsection 8.2.3.2, we will provide a method to approximate the optimal solutions to P and D using sequences of semidefinite programs (SDPs) and sum-of-squares (SOS) programs, respectively. We utilize tools developed in the context of the multi-dimensional moment problem allowing us to replace the tuple of measures in P by sequences of moments.

The following observation plays a key role in our approximation scheme. Notice that the equality constraint in (8.26) is equivalent to

$$\langle v, \delta_T \otimes \mu_T \rangle = \langle v, \delta_0 \otimes \mu_0 \rangle + \langle \mathcal{L}v, \mu \rangle \tag{8.29}$$

for all $v \in \mathcal{C}([0,T] \times \mathcal{X})$. Since the set of polynomials are dense in $\mathcal{C}([0,T] \times \mathcal{X})$ and the ring $\mathbb{R}[t, \mathbf{x}]$ is closed under addition and multiplication, (8.29) is equivalent to

$$\begin{gathered} \int_{\mathcal{X}} v(T, \mathbf{x}) d\mu_T = \int_{\mathcal{X}} v(0, \mathbf{x}) d\mu_0 + \int_{[0,T]\times\mathcal{X}} \mathcal{L}v d\mu \\ \text{for all } v(t, \mathbf{x}) = t^a \mathbf{x}^{\boldsymbol{\alpha}}, (a, \boldsymbol{\alpha}) \in \mathbb{N} \times \mathbb{N}^n, \end{gathered} \tag{8.30}$$

where $a \in \mathbb{N}$, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n) \in \mathbb{N}^n$ and $\mathbf{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}$. Using the above procedure, the linear constraints in P hold provided that (8.29) holds for all monomial functions $v(t, \mathbf{x})$. A standard relaxation is then to require that (8.29) holds for all monomials up to a given fixed degree r, i.e., $a + |\boldsymbol{\alpha}| = a + \sum_{i=1}^n \alpha_i \leq r$.

Since $v(t, \mathbf{x})$ is a monomial, the integration of v with respect to a measure μ results in a moment of μ. Therefore, (8.30) is a linear constraint on the moments of μ_0, μ and μ_T. In this case, instead of finding a tuple of measures satisfying the constraints in (8.26), we aim to find (finite) sequences of numbers that satisfy the constraint (8.30). Moreover, the sequences of numbers are moments of measures $\widetilde{\mu}, \widehat{\mu}, \mu, \mu_T$. As required by (8.26), these measures must be supported on certain specified sets. To formalize this idea, in order to obtain an approximated solution to (8.12), we want to find sequences of numbers that are moments of the tuple of measures feasible in (8.26).

8.2.3.1 Riesz Functional

We adopt the notions related to multi-dimensional moment problem introduced in Section 6.2.3 from Chapter 6. To better explain this approach, we introduce additional notions related to the multi-dimensional moment problem.

Given an $\mathbb{R}^n$-valued random variable $\mathbf{x} \sim \nu$ and an integer vector $\boldsymbol{\alpha} \in \mathbb{N}^n$, the $\boldsymbol{\alpha}$-moment of $\mathbf{x}$ is defined as $\mathbb{E}[\mathbf{x}^{\boldsymbol{\alpha}}] = \int_{\mathbb{R}^n} \prod_{i=1}^n x_i^{\alpha_i} d\nu$. Moreover, we define the *order* of an $\boldsymbol{\alpha}$-moment to be $|\boldsymbol{\alpha}|$. Finally, a sequence $\mathbf{y} = \{y_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{N}^n}$ indexed by $\boldsymbol{\alpha}$ is called a *multi-sequence*. Given a multi-sequence $\mathbf{y} = \{y_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathbb{N}^n}$, we define the linear functional $L_{\mathbf{y}} : \mathbb{R}[\mathbf{x}] \to \mathbb{R}$ as

$$f(\mathbf{x}) = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^n} f_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}} \mapsto L_{\mathbf{y}}(f) = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^n} f_{\boldsymbol{\alpha}} y_{\boldsymbol{\alpha}}. \tag{8.31}$$

The introduction of the above functional, often known as the *Riesz functional* [206], is convenient to express the moments of random variables. More specifically, let $\mathbf{x}$ be an $\mathbb{R}^n$-valued random variable with corresponding probability measure ν and let f be a polynomial in $\mathbf{x}$. Then, the expectation of $f(\mathbf{x})$ is equal to

$$\int f(\mathbf{x}) d\nu = \int \sum_{\boldsymbol{\alpha}\in\mathbb{N}^n} f_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}} d\nu = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^n} f_{\boldsymbol{\alpha}} y_{\boldsymbol{\alpha}} = L_{\mathbf{y}}(f)$$

where $y_{\boldsymbol{\alpha}}$ is the $\boldsymbol{\alpha}$-moment of $\mathbf{x}$.

According to Putinar's Positivstellensatz (see Theorem 27 from Chapter 7), whether an infinite multi-sequence is K-feasible can be characterized using the moment and localizing matrices associated to this sequence. In the following subsection, we will leverage this result to construct approximate solutions of P and D.

8.2.3.2 Finite-dimensional approximations

A. SDP relaxation of P

As mentioned above, in the relaxed version of P, we aim to optimize over sequences of moments of a tuple of measures $(\widetilde{\mu}, \widehat{\mu}, \mu, \mu_T)$. We use $(\widetilde{\mathbf{y}}, \widehat{\mathbf{y}}, \mathbf{y}, \mathbf{y}_T)$ to denote the moment sequences of the corresponding measures, respectively. On the one hand, since μ is supported on $[0,T]\times\mathcal{X}$, the elements in the moment sequence $\mathbf{y}$ are of the form $y_{\boldsymbol{\alpha}}$ where $\boldsymbol{\alpha}\in\mathbb{N}\times\mathbb{N}^n$. On the other hand, since μ_T is supported on $\mathcal{X}$, the elements in $\mathbf{y}_T$ are of the form $y_{\boldsymbol{\alpha}}$ where $\boldsymbol{\alpha}\in\mathbb{N}^n$. Using the Riesz functional (8.31) on (8.30), we obtain

$$\begin{aligned}&L_{\mathbf{y}_T}(v(T,\cdot)) - L_{\mathbf{y}}(\mathcal{L}v) = L_{\mathbf{y}_0}(v(0,\cdot))\\ &\text{for all } v(t,\mathbf{x}) = t^a\mathbf{x}^{\boldsymbol{\alpha}} \text{ and } a+|\boldsymbol{\alpha}|\le 2r.\end{aligned} \tag{8.32}$$

Applying the Riesz functional on the first linear constraint in P, we have that

$$\begin{aligned}&L_{\widetilde{\mathbf{y}}}(w) + L_{\widehat{\mathbf{y}}}(w) = L_{\mathbf{y}}(w)\\ &\text{for all } w(t,\mathbf{x}) = t^a\mathbf{x}^{\boldsymbol{\alpha}} \text{ and } a+|\boldsymbol{\alpha}|\le 2r.\end{aligned} \tag{8.33}$$

Both equations in (8.33) are linear with respect to the elements in $\mathbf{y},\widetilde{\mathbf{y}},\widehat{\mathbf{y}},\mathbf{y}_T$; hence, it is possible to write them compactly into a linear equation, as follows:

$$A_r(\widetilde{\mathbf{y}},\widehat{\mathbf{y}},\mathbf{y},\mathbf{y}_T) = b_r. \tag{8.34}$$

From Theorem 27, since $\mathtt{supp}(\mu)\subseteq[0,T]\times\mathcal{X}$, the moment and localizing matrices of $\mathbf{y}$ with respect to $g_i^{\mathcal{X}}$ are positive semidefinite for all positive integers $r\in\mathbb{N}$. Let

$$d_i^{\mathcal{X}_u} = \frac{\deg\, g_i^{\mathcal{X}_u}}{2}\ \forall i\in[n_{\mathcal{X}_u}],\ d_j^{\mathcal{X}} = \frac{\deg\, g_j^{\mathcal{X}}}{2}\ \forall j\in[n_{\mathcal{X}}]$$

where deg denotes the degree of a polynomial. Given a fixed positive integer $r\in\mathbb{N}$, we construct the r-th order relaxation of $\mathtt{P}$, as follows:

$$\begin{aligned}
\mathtt{P_r} : \underset{(\widetilde{\mathbf{y}},\widehat{\mathbf{y}},\mathbf{y},\mathbf{y}_T)}{\text{maximize}}\ & L_{\widetilde{\mathbf{y}}}(g) \\
\text{subject to } & A_r(\widetilde{\mathbf{y}},\widehat{\mathbf{y}},\mathbf{y},\mathbf{y}_T) = b_r \\
& M_r(\widetilde{\mathbf{y}}) \succeq 0,\ M_{r-1}(t(T-t),\widetilde{\mathbf{y}}) \succeq 0 \\
& M_{r-d_i^{\mathcal{X}_u}}(g_i^{\mathcal{X}_u},\widetilde{\mathbf{y}}) \succeq 0, \forall i \in [n_{\mathcal{X}_u}] \\
& M_r(\widehat{\mathbf{y}}) \succeq 0, M_{r-1}(t(T-t),\widehat{\mathbf{y}}) \succeq 0 \\
& M_{r-d_i^{\mathcal{X}}}(g_i^{\mathcal{X}},\widehat{\mathbf{y}}) \succeq 0, \forall i \in [n_{\mathcal{X}}] \\
& M_r(\mathbf{y}) \succeq 0,\ M_{r-1}(t(T-t),\mathbf{y}) \succeq 0 \\
& M_{r-d_i^{\mathcal{X}}}(g_i^{\mathcal{X}},\mathbf{y}) \succeq 0, \forall i \in [n_{\mathcal{X}}] \\
& M_r(\mathbf{y}_T) \succeq 0, \\
& M_{r-d_i^{\mathcal{X}}}(g_i^{\mathcal{X}},\mathbf{y}_T) \succeq 0, \forall i \in [n_{\mathcal{X}}].
\end{aligned} \tag{8.35}$$

In this program, the decision variable is the 4-tuple of finite multi-sequences $(\widetilde{\mathbf{y}},\widehat{\mathbf{y}},\mathbf{y},\mathbf{y}_T)$. Furthermore, $\mathtt{P_r}$ is an SDP and, thus, can be solved using off-the-shelf software. In addition to relaxing the primal LP $\mathtt{P}$, it is also possible to relax the dual LP $\mathtt{D}$, as shown next.

B. SOS relaxation of $\mathtt{D}$

To formulate the relaxed program of $\mathtt{D}$, we begin by considering the dual of $\mathtt{P_r}$. Furthermore, as shown in $\mathtt{D}$, the decision variables are $v(t,\mathbf{x}) \in \mathcal{C}^1([0,T] \times \mathcal{X})$ and $w(t,\mathbf{x}) \in \mathcal{C}([0,T] \times \mathcal{X})$. The relaxed program is obtained by restricting the functions in (8.28) to polynomials of degrees up to $2r$, and then replacing the non-negativity constraint with sum-of-squares constraints [207]. To formalize this argument, we first need to introduce some notations.

Given a semi-algebraic set $A = \{\mathbf{x} \in \mathbb{R}^n \mid h_i(\mathbf{x}) \geq 0, h_i \in \mathbb{R}[\mathbf{x}], \forall i \in [m]\}$, we define the

r-th order quadratic module of A as

$$\begin{aligned} Q_r(A) = \big\{q \in& \mathbb{R}[\mathbf{x}]_r \mid \exists \text{ SOS } \{s_k\}_{k\in[m]\cup\{0\}} \subset \mathbb{R}[x]_r \\ &\text{s.t. } q = s_0 + \sum_{k\in[m]} h_k s_k\big\}. \end{aligned} \tag{8.36}$$

Following a process similar to [208], the relaxed dual program, denoted by $\mathtt{D}_r$, can be written as follows

$$\begin{aligned} \mathtt{D_r} : \text{minimize } & \int v(0,\cdot) d\mu_0 \\ & \text{subject to } w - g \in Q_{2r}([0,T]\times\mathcal{X}_u) \\ & -\mathcal{L}v - w \in Q_{2r}([0,T]\times\mathcal{X}) \\ & v(T,\cdot) \in Q_{2r}(\mathcal{X}) \\ & w \in Q_{2r}([0,T]\times\mathcal{X}). \end{aligned} \tag{8.37}$$

In this program, we optimize over the vector of polynomials $(w,v) \in \mathbb{R}[t,\mathbf{x}]_{2r} \times \mathbb{R}[t,\mathbf{x}]_{2r}$.

Notice that $\mathtt{P}_r$ and $\mathtt{D}_r$ provide approximate solutions to $\mathtt{P}$ and $\mathtt{D}$, respectively. In the next theorem, we show that there is no duality gap between $\mathtt{P}_r$ and $\mathtt{D_r}$ and that the optimal values of $\mathtt{P}_r$ and $\mathtt{D}_r$ converge to the optimal values of $\mathtt{P}$ and $\mathtt{D}$, respectively, as r increases.

Theorem 39. *Given a positive integer $r \in \mathbb{N}$, let $p_r^\star$ and $d_r^\star$ be the optimal values of $\mathtt{P}_r$ and $\mathtt{D}_r$, respectively. If $\mathcal{X}_u$ and $\mathcal{X}$ have nonempty interior, then $p_r^\star = d_r^\star$. Furthermore,*

$$d_r^* = p_r^\star \downarrow p^\star = d^\star. \tag{8.38}$$

Proof. See Appendix A.7. □

As a result of this theorem, $p_r^\star$ is a non-increasing function of r and it converges asymptotically to $p^\star$. From Theorem 37, $p^\star$ is equal to the expected time the system spends in the unsafe region, as expressed in (8.12).

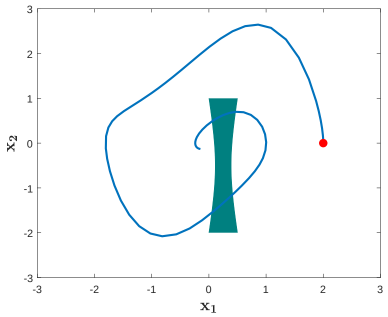


Figure 8-3: Trajectory $\mathbf{x}(t)$, where $t \in [0, 10]$, of the Van der Pol system (blue curve) with initial condition $\mathbf{x}_0 = [2, 0]^T$ (red circle). The unsafe region $\mathcal{X}_u$ is depicted by the nonconvex colored set.

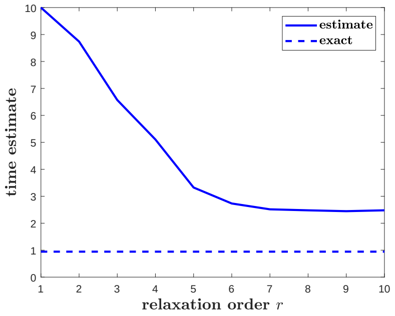


Figure 8-4: This figure shows the exact value (dashed line) and the approximation (solid line) to (8.12) using $\mathtt{D}_r$ with different values of r. The system dynamics under consideration is the Van der Pol system (8.39), whereas the initial distribution is $\mu_0 = \delta_{[2,0]^\top}$.

8.2.4 Numerical Examples

In this section, we provide a numerical example to illustrate our framework. We complete all numerical simulations using YALMIP [209] (for sum-of-squares programs) and MOSEK [210] (for semidefinite programs). In particular, we evaluate our framework on the Van der Pol oscillator—a second order nonlinear dynamical system whose dynamics is given by

$$\begin{aligned}\dot{x}_1 &= -x_2 \\ \dot{x}_2 &= x_1 + (x_1^2 - 1)x_2.\end{aligned} \tag{8.39}$$

Moreover, we consider the following parameter settings (see Figure 8-3): (i) the final time is set to be $T = 10$, (ii) the initial condition is set to be $\mathbf{x}(0) = \mathbf{x}_0 = [2, 0]^\top$, and (iii) the unsafe region is specified by a nonconvex two-dimensional semi-algebraic set $\mathcal{X}_u =$

$\{(x_1, x_2) \in \mathbb{R}^2 \mid 52(x_1 - 0.25)^2 - (x_2 + 0.5)^2 \leq 1, 0 \leq x_1 \leq 0.5, -2 \leq x_2 \leq 1\}$. To ease the numerical computations, we adopt proper scaling of the system's coordinates such that T and $\mathcal{X}$ are normalized to be $T = 1$ and $\mathcal{X} = [-1, 1] \times [-1, 1]$, respectively. In this case, (8.12) cannot be computed analytically. However, through numerical simulation, we obtain that the Van der Pol oscillator spends (approximately) 0.9446 seconds in the unsafe region $\mathcal{X}_u$. We demonstrate our upper bounds on this time using $\mathtt{D}_r$ with varying values of r in Figure 8-4.

Chapter 9

Conclusion

The analysis of the global behavior of networked systems presents the following three major challenges: (*i*) analyzing or characterizing the properties of networked systems generally requires full knowledge of the parameters describing the system's dynamics, yet an exact quantitative description of the parameters of the system may not be available due to measurement errors and/or modeling uncertainties; (*ii*) retrieving the whole structure of many real networks is very challenging due to both computation and security constraints, hence an exact analysis of the global behavior of many real-world networks is practically unfeasible; (*iii*) the dynamics describing the interactions between components are often stochastic, which leads to difficulty in analyzing individual behaviors in the network. In this thesis, we have addressed all three challenges using results from structural systems theory and measure theory, as summarized below.

In the first part of the thesis, we adopted graph-theoretic methods to handle the challenge brought by inexact models and/or imprecise measurements. Chapter 2 studied the problem of characterizing structural target controllability in undirected networks with unknown link weights. We achieved this goal by first characterizing the generic properties of symmetrically structured matrices. We then derived a necessary and sufficient condition for structural controllability of undirected networks with multiple control inputs. Based on these results, we provided a graph-theoretic necessary and sufficient condition for structural target controllability. Furthermore, we derived necessary and sufficient conditions for (symmetric) structural output controllability using graph and structural systems theory.

In Chapter 3, we have addressed the problem of designing the topology of a networked dynamical system in order to achieve (general) structural controllability. In particular, given a system digraph, we have developed an efficient methodology to find the minimum number of edges that must be added to the digraph to render a structurally controllable system. As part of our analysis, we have characterized the set of all possible solutions to this problem, and provided a polynomial-time algorithm to obtain an optimal solution. Additionally, we have presented scalable algorithms to solve our problem under additional assumptions that are commonly found in engineering applications. Finally, we have numerically illustrated our results in the context of random networked systems.

In Chapter 4, we leveraged results provided in Chapter 2 and Chapter 3 to examine the problem of selecting a set of undirected edges incurring into a minimum total cost in order to render a (symmetric) structural target controllable system. We presented a thorough analysis of this problem and showed that obtaining an optimal solution to the problem under consideration is NP-hard. Motivated by engineering applications, we also considered two special instances of this problem by imposing extra assumptions on the topology of the system and/or the cost function. We showed that these two special cases can be solved efficiently, and proposed polynomial-time algorithms to obtain an optimal solution. Finally, we demonstrated the validity of our algorithms on particular system graphs.

In Chapter 5, we studied structural stabilizability—a more general system property than controllability, in undirected networked dynamical systems. Using the results from Chapter 2, we proposed a computationally-efficient graph-theoretic method to estimate the maximum dimension of stabilizable subspace of an undirected network. Furthermore, we formulated the optimal actuator-disabling attack problem (respectively, optimal recovery problem), whose objective is to remove (respectively, add) actuators to minimize (respectively, maximize) the maximum dimension of stabilizable subspace, respectively.

We showed that these two problems are NP-hard. Despite this, we developed a $(1-1/e)$ approximation algorithm for the optimal recovery problem. Finally, we provided graph-theoretic conditions for structural stabilizability in arbitrary linear structural systems (i.e., systems without parameter constraints).

In the second part of this thesis, we utilized measure-theoretic techniques to estimate global properties of a network and to analyze stochastic networked processes. More specifically, in Chapter 6, we showed that, given enough local information (i.e., subgraph counts), it is possible to approximate the spectral radius of large-scale networks. In particular, we developed a novel mathematical framework to upper and lower bound the spectral radius of a digraph from the counts of a collection of small subgraphs. By leveraging recent results on the K-moment problem, we proposed a hierarchy of semidefinite programs of small size allowing us to compute sequences of upper and lower bounds on the spectral radius of a digraph using, solely, the counts of certain subgraphs. We illustrated the quality of our bounds using both random digraphs and real-world directed networks.

In Chapter 7, we analyzed the (exact) stochastic dynamics of the networked SIS, SI, and SIR epidemic models with heterogeneous spreading and recovery rates. The analysis of these models are, in general, very challenging since their state space grows exponentially with the number of nodes in the network. A common approach to overcome this challenge is to apply moment-closure techniques to approximate the exact stochastic dynamics via ordinary differential equations. However, most existing moment-closure techniques do not provide quantitative guarantees on the quality of the approximation, limiting the applicability of these techniques. To overcome this limitation, we have proposed a novel moment-closure framework which allows us to derive explicit quality guarantees. This framework is based on recent results from real algebraic geometry relating the multidimensional moment problem with semidefinite programming. We illustrated how this technique can be used to derive upper and lower bounds on the exact (stochastic) dynamics of the SIS, SI, and SIR models. Moreover, we provided a simplified version of our moment-closure technique to approximate the mean dynamics of the SIS model using a linear number of piecewise-affine differential equations. We illustrated the validity of our results via numerical simulations in the Zachary's Karate Club network. Moreover, in the second part of Chapter 7, we introduced a model of coinfection dynamics in multilayer networks, which we call the non-homogeneous $(SIS)^L$ model. We then studied the problem of simultaneously controlling the dynamics of several diseases in an arbitrary multilayer network by distributing vaccines throughout the network. Since the spread of the diseases is closely related to the eigenvalues of the adjacency matrices representing network layers, we reformulated our control problem as a spectral optimization problem and casted these spectral problems as a geometric program. In addition, we also examined the case when global structural information about the network is not available. We have proposed a linear program based on *Gershgorin's circle theorem* as an efficient relaxed solution to the global problem. Finally, we illustrated our results with numerical simulations in synthetic networks.

Our results in Chapter 6 and Chapter 7 mainly utilized results in real algebraic geometry that relates semidefinite programming to the multidimensional moment problem. In Chapter 8, we further considered two specific application of the multidimensonal moment problem. In the first application, we applied similar idea as in Chapter 6 to lower-bound the algebraic connectivity of (connected) undirected graphs using, solely, local structural information in the form of closed-walks. We provided experiments in both random and real-world networks to demonstrate the performance of our bounds. In the second application, we proposed a flexible safety verification notion for nonlinear autonomous systems described via polynomial dynamics and unsafe regions described via polynomial inequalities. Instead of verifying safety by checking whether the dynamics completely avoids the unsafe regions, we consider the system to be safe if it spends less than a certain amount of time in these regions. This more flexible notion can be of relevance in, for example, solar-powered vehicles where the vehicle should avoid spending too much time is dark areas. More generally, this framework can be useful in those situations where the system is able to tolerate the exposure to a deteriorating agent, such as excessive heat or radiation, for a limited amount of time. To solve this problem, we first proposed an infinite-dimensional LP over the space of measures whose solution is equal to the (expected) time our (nonlinear) system spends in the (possibly nonconvex) unsafe regions. We then approximated the solution of the LP through a monotonically converging sequence of upper bounds by solving a hierarchy of SDPs by leveraging results from the multidimensional moment problem. Finally, we validated our approach via a simple example involving a nonlinear Van der Pol oscillator.

List of Figures

List of Tables

Appendix A

Appendix

A.1 Proof of the results in Chapter 2

A.1.1 Proof of Lemma 1 and related results

Before we proceed to the proof of Lemma 1, we introduce Proposition 2, which lays the foundation for the proof of Lemma 1.

Proposition 2 ([30, §1.2])**.** *Given a (symmetrically) structured matrix* $\bar{M} \in \{0, \star\}^{n\times m}$ *and* $\mathcal{B}(\mathcal{S}_1, \mathcal{S}_2, \mathcal{E}_{\mathcal{S}_1,\mathcal{S}_2})$*, where* $\mathcal{S}_1 = \{v_1, \dots, v_m\}$*,* $\mathcal{S}_2 = \{v'_1, \dots, v'_n\}$*, and* $\mathcal{E}_{\mathcal{S}_1,\mathcal{S}_2}$ *is defined by* $\mathcal{E}_{\mathcal{S}_1,\mathcal{S}_2} = \{\{v_i, v'_j\} \colon [\bar{M}]_{ji} \neq 0, v_i \in \mathcal{S}_1, v'_j \in \mathcal{S}_2\}$*, then* $\text{t–rank}(\bar{M}) = n$ *if and only if* $|\mathcal{N}_{\mathcal{B}}(\mathcal{S})| \geq |\mathcal{S}|$ *for all* $\mathcal{S} \subseteq \mathcal{S}_2$*.*

Proof of Lemma 1. First, we show the sufficiency of the theorem. Notice that the generic-rank of $C_{\mathcal{T}}\bar{A}$ equals k, if and only if, there exists a k-by-k non-zero minor in $C_{\mathcal{T}}\bar{A}$; hence, it suffices to find that minor. Since $|\mathcal{N}(\mathcal{S})| \geq |\mathcal{S}|$, $\forall \mathcal{S} \subseteq \mathcal{X}_{\mathcal{T}}$, there exists k entries that lie on distinct rows and distinct columns of $C_{\mathcal{T}}\bar{A}$ according to Proposition 2. As a result, we can select rows indexed by $\mathcal{T} = \{i_1, \dots, i_k\}$ and columns indexed $j_1, \cdots, j_k$ in $\bar{A}$ such that $\{[\bar{A}]_{i_\ell j_\ell}\}_{\ell=1}^{k}$ lies on distinct rows and distinct columns. Next, we consider the following two cases.

On one hand, if $\{j_1, \dots, j_k\} = \{i_1, \dots, i_k\}$, then $M = C_{\mathcal{T}}\bar{A}C_{\mathcal{T}}^{\top}$ is a square submatrix of

$\bar{A}$. We consider a particular numerical realization $\tilde{A}$ of $\bar{A}$, as follows. Let $[\tilde{A}]_{ij} \neq 0$ for all $(i,j) \notin \{(i_\ell, j_\ell) : \ell \in [k]\}$, $[\tilde{A}]_{ij} = [\tilde{A}]_{ji}$, and $[\tilde{A}]_{ij} = 0$ otherwise. Subsequently, by computing the determinant , $\det(C_{\mathcal{T}}\tilde{A}C_{\mathcal{T}}^\top) = \text{sgn}(\sigma_1)\Pi_{\ell=1}^k[\tilde{A}]_{i_\ell j_\ell} + \text{sgn}(\sigma_2)\Pi_{\ell=1}^k[\tilde{A}]_{j_\ell i_\ell}$, where $\text{sgn}(\sigma_1)$ and $\text{sgn}(\sigma_2)$ are the signatures of the permutations $\sigma_1 = \{(i_\ell, j_\ell) : \ell \in [k]\}$, and $\sigma_2 = \{(j_\ell, i_\ell) : \ell \in [k]\}$, respectively. Notice that if $\text{sgn}(\sigma_1) = \text{sgn}(\sigma_2)$, then it follows that $\det(C_{\mathcal{T}}\tilde{A}C_{\mathcal{T}}^\top) \neq 0$. Furthermore, if $\{\tilde{A} : \det(C_{\mathcal{T}}\tilde{A}C_{\mathcal{T}}^\top) = 0\}$ is a proper variety, we have that M admits an k-by-k non-zero minor generically. Thus, the generic-rank of $C_{\mathcal{T}}\bar{A}$ equals to k.

On the other hand, when $\{j_1, \ldots, j_k\} \neq \{i_1, \cdots, i_k\}$, it sufficies to show there exists a numerical realization $\tilde{A}$ such that $\det([\tilde{A}]_{i_1,\cdots,i_k}^{j_1,\cdots,j_k}) \neq 0$. We consider a numerical realization $\tilde{A}$ by assigning distinct real values to $\star$-entries corresponding to $\{[\bar{A}]_{i_\ell j_\ell}\}_{\ell=1}^k$ while keeping $[\tilde{A}]_{ij} = [\tilde{A}]_{ji}$, and assigning 0 otherwise. Without loss of generality, we can permute $\{\ell\}_{\ell=1}^k$ such that for each $[\bar{A}]_{i_{\ell_r} j_{\ell_r}} \in \{[\bar{A}]_{i_{\ell_r} j_{\ell_r}}\}_{r=1}^p$, $[\bar{A}]_{j_{\ell_r} i_{\ell_r}}$ is not in matrix $[\bar{A}]_{i_1,\cdots,i_k}^{j_1,\cdots,j_k}$, and for each $[\bar{A}]_{i_{\ell_r} j_{\ell_r}} \in \{[\bar{A}]_{i_{\ell_r} j_{\ell_r}}\}_{r=p+1}^k$, $[\bar{A}]_{j_{\ell_r} i_{\ell_r}}$ is in matrix $[\bar{A}]_{i_1,\cdots,i_k}^{j_1,\cdots,j_k}$. We declaim that there is only one nonzero entry in either the i_{ℓ_r}th row or j_{ℓ_r}th column, $\forall r \in [p]$, otherwise it contradicts that $\{[\bar{A}]_{i_\ell j_\ell}\}_{\ell=1}^k$ are in distinct rows and distinct columns of $[\bar{A}]$. Thus, we compute $\det([\tilde{A}]_{i_1,\cdots,i_k}^{j_1,\cdots,j_k})$,

$$\det([\tilde{A}]_{i_1,\cdots,i_k}^{j_1,\cdots,j_k}) = (\prod_{r=1}^p [\tilde{A}]_{i_{\ell_r} j_{\ell_r}}) \cdot \det([\tilde{A}]_{i_{\ell_{p+1}},\cdots,i_{\ell_k}}^{j_{\ell_{p+1}},\cdots,j_{\ell_k}}) \neq 0, \tag{A.1}$$

where $\det([\tilde{A}]_{i_{\ell_{p+1}},\cdots,i_{\ell_k}}^{j_{\ell_{p+1}},\cdots,j_{\ell_k}}) \neq 0$ is true because $\{i_1, \cdots, i_k\} = \{j_1, \cdots, j_k\}$. Thus, there exists numerical realization such that $\det([\tilde{A}]_{i_1,\cdots,i_k}^{j_1,\cdots,j_k}) \neq 0$. Next, we show the necessity of the theorem by contrapositive. We assume that there exists $\mathcal{S} \subseteq \mathcal{X}_{\mathcal{T}}$, such that $|\mathcal{N}(\mathcal{S})| < |\mathcal{S}|$. Then, by Proposition 2, there does not exist k entries that lie on the distinct rows and distinct columns of $C_{\mathcal{T}}\bar{A}$, which implies g–rank$(C_{\mathcal{T}}\bar{A}) < k$. □

A.1.2 Proof of Corollary 1

Proof. Suppose $|\mathcal{N}(\mathcal{S})| \geq |\mathcal{S}|$, $\forall \mathcal{S} \subseteq \mathcal{X}_{\mathcal{T}}$, then, by Proposition 2, there exist k entries, $\{[\bar{A}, \bar{B}]_{i_\ell j_\ell}\}_{\ell=1}^k$, such that they are all $\star$-entries which lie on distinct rows and distinct columns of $[\bar{A}, \bar{B}]$. Among those k entries, suppose $\{[\bar{A}, \bar{B}]_{i_\ell j_\ell}\}_{\ell=1}^q$ are in columns of $\bar{A}$, and $\{[\bar{A},\bar{B}]_{i_\ell j_\ell}\}_{\ell=q+1}^{k}$ are in the columns of $\bar{B}$. By Lemma 1, there exists a numerical realization $\tilde{A}$, such that $\det([\tilde{A},\tilde{B}]_{i_1,\ldots,i_q}^{j_1,\ldots,j_q}) \neq 0$. Since $\bar{B}$ is a structured matrix, there exists a numerical realization $\tilde{B}$ such that $\det([\tilde{A},\tilde{B}]_{i_{q+1},\ldots,i_k}^{j_{q+1},\ldots,j_k}) \neq 0$ Hence, there exists a numerical realization $[\tilde{A},\tilde{B}]$ with

$$\begin{aligned}\det([\tilde{A},\tilde{B}]_{i_1,\ldots,i_k}^{j_1,\ldots,j_k}) &= \det([\tilde{A},\tilde{B}]_{i_1,\ldots,i_q}^{j_1,\ldots,j_q})\det([\tilde{A},\tilde{B}]_{i_{q+1},\ldots,i_k}^{j_{q+1},\ldots,j_k})\\ &\neq 0,\end{aligned}$$

which implies that g–rank$(C_{\mathcal{T}}\left[\bar{A},\bar{B}\right]) = k$. □

A.1.3 Proof of Lemma 2

We introduce Proposition 3, Proposition 4 and Lemma 16 to support the proof of Lemma 2.

Proposition 3 ([211, §2.1]). *Let $\varphi_1(s)$ and $\varphi_2(s)$ be polynomials in s with $\varphi_1(s) = \sum_{i=0}^{n_1} a_i s^{n_1-i}$, and $\varphi_2(s) = \sum_{i=0}^{n_2} b_i s^{n_2-i}$, respectively. Let $R(\varphi_1,\varphi_2)$ be defined as*

$$R(\varphi_1,\varphi_2) = \det\left(\begin{bmatrix} a_{n_1} & a_{n_1-1} & \cdots & a_0 & 0 & \cdots & 0\\ 0 & a_{n_1} & \cdots & a_1 & a_0 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & a_{n_1} & a_{n_1-1} & \cdots & a_0\\ hline0 & 0 & \cdots & & & \cdots & b_0\\ \vdots & \vdots & & \vdots & \vdots & & \vdots\\ 0 & b_{n_2} & \cdots & b_1 & b_0 & \cdots & 0\\ b_{n_2} & b_{n_2-1} & \cdots & b_0 & 0 & \cdots & 0 \end{bmatrix}\right). \tag{A.2}$$

If $a_{n_1} \neq 0$ and $b_{n_2} \neq 0$, then $\varphi_1(s)$ and $\varphi_2(s)$ have a nontrivial common factor if and only if the $R(\varphi_1,\varphi_2) = 0$.

Proposition 4 (Hoffman-Wielandt Theorem [168, §6.3]). *Given $n \times n$ symmetric matrices A and E, let $\lambda_1,\ldots,\lambda_n$ be the eigenvalues of A, and $\hat{\lambda}_1,\ldots,\hat{\lambda}_n$ be the eigenvalues of $A+E$. There is a permutation $\sigma(\cdot)$ of the integers $\{1,\ldots,n\}$ such that*

$$\sum_{i=1}^{n}(\hat{\lambda}_{\sigma(i)} - \lambda_i)^2 \leq \|E\|_F^2, \tag{A.3}$$

where $\|E\|_F = \sqrt{tr(EE^\top)}$.

Lemma 16. *Let $\bar{A}$ be an $n \times n$ symmetrically structured matrix, and let $G(\bar{A}) = \{\mathcal{X},\mathcal{E}_{\mathcal{X},\mathcal{X}}\}$ be the digraph associated with $\bar{A}$. Assume* t–rank$(\bar{A}) = k$*, and denote*

$\{[\bar{A}]_{i_\ell j_\ell}\}_{\ell=1}^{k}$ as the k entries that lie on distinct rows and distinct columns. We define $\mathcal{S} = \{x_{i_1}, \dots, x_{i_k}\} \subseteq \mathcal{X}$. Then, $G_{\mathcal{S}}$ can be covered by disjoint cycles.

Proof of Lemma 16. We approach the proof by contradiction. Suppose $G_{\mathcal{S}}$ cannot be covered by disjoint cycles, then at least one vertex $x_i \in \mathcal{S}$ can only be covered by cycles intersecting with other cycles in $G_{\mathcal{S}}$, which implies that there does not exist k edges in which no two edges share the same 'tail' or 'head' vertex in $G(\bar{A})$, i.e., there does not exist k entries that lie on distinct rows and distinct columns of $\bar{A}$, which, by Proposition 2, contradicts t–rank$(\bar{A}) = k$. □

Proof of Lemma 2. We expand the characteristic polynomial of a matrix $\tilde{A}$ as

$$\det(sI - \tilde{A}) = s^n + a_{n-1}s^{n-1} \cdots + a_{n-k}s^{n-k} + \cdots + a_0. \tag{A.4}$$

Besides, we have

$$a_q = (-1)^{n-q} \sum_{1 \le k_1 < \cdots < k_{n-q} \le n} \det([\tilde{A}]_{k_1,\dots,k_{n-q}}^{k_1,\dots,k_{n-q}}), \tag{A.5}$$

where $q = 0, 1, \dots, n-1$. Since t–rank$(\bar{A}) = k$, there exists a numerical realization $\tilde{A}$ and a set of indexes, $\{i_1, \dots, i_k\} \subseteq [n]$, such that $\det([\tilde{A}]_{i_1,\dots,i_k}^{i_1,\dots,i_k}) \neq 0$. Furthermore, $V_0 := \{\mathbf{p}_{\tilde{A}} \in \mathbb{R}^{n_{\bar{A}}} : a_{n-k} = 0\}$ is a proper variety. Since the maximum order of principle minor is at most the term rank of a matrix, we have $a_{n-k-1} = \cdots = a_0 = 0$. Thus, to characterize nonzero eigenvalues, we define the polynomial $\varphi_{\tilde{A}}(s)$ as

$$\varphi_{\tilde{A}}(s) = s^k + a_{n-1}s^{k-1} + \cdots + a_{n-k}. \tag{A.6}$$

In the rest of the proof, we show that there exists a numerical realization $\mathbf{p}_{\tilde{A}} \in V_0^c$ such that $\tilde{A}$ has k non-zero simple eigenvalues. Since t–rank$(\bar{A}) = k$, we define the set $\mathcal{S}$ as in Lemma 16. By Lemma 16, there exist disjoint cycles $\mathcal{C}_1, \dots, \mathcal{C}_l$ covering $G_{\mathcal{S}}$. Let us denote by $\mathcal{C}_i$ the i-th cycle in $\{\mathcal{C}_1, \dots, \mathcal{C}_l\}$. Moreover, without loss of generality, we let the length of cycle $\mathcal{C}_i$ be either $|\mathcal{C}_i| = 2q$, or $|\mathcal{C}_i| = 2q+1$, for some $q \in \mathbb{N}$. Note that by definition, there is a one-to-one correspondence between the edge in $G(\bar{A})$ and the $\star$-entry in $\bar{A}$. From this observation, we denote by $\bar{A}_i \in \{0, \star\}^{|\mathcal{C}_i| \times |\mathcal{C}_i|}$ the square submatrix formed by collecting rows and columns corresponding to the indexes of vertices in $\mathcal{V}_{\mathcal{C}_i}$ of the cycle $\mathcal{C}_i$. We let all the $\star$-entries of $\bar{A}$ be zero, except for $\star$-entries corresponding to edges in $\{\mathcal{E}_{\mathcal{C}_i}\}_{i=1}^{l}$. Hence, there exists a permutation matrix P and numerical realization $\tilde{A}$, such that $P\tilde{A}P^{-1}$ is a block diagonal matrix,

$$P\tilde{A}P^{-1} = \begin{bmatrix} \tilde{A}_1 & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \tilde{A}_2 & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \tilde{A}_l & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{A.7}$$

If $|\mathcal{C}_i| = 2q$, we can assume $\mathcal{C}_i = (x_{i_1}, x_{j_1}, x_{i_2}, x_{j_2}, \ldots, x_{i_q}, x_{j_q}, x_{i_1})$ without loss of generality. Since $G_{\mathcal{V}_{\mathcal{C}_i}}$ is a subgraph of the digraph $G(\bar{A})$ associated with the symmetrically structured matrix $\bar{A}$, there exist q disjoint cycles of length-2 covering $G_{\mathcal{V}_{\mathcal{C}_i}}$, i.e., cycles $(x_{i_1}, x_{j_1}, x_{i_1}), (x_{i_2}, x_{j_2}, x_{i_2}), \ldots, (x_{i_q}, x_{j_q}, x_{i_q})$. We assign distinct nonzero weights to $\star$-entries of $\bar{A}_i$ that correspond to edges in the q cycles of length-2, and assign zero weights to other $\star$-entries in $\bar{A}_i$. As a result, we have

$$\tilde{A}_i = \begin{bmatrix} \boxed{\begin{matrix} 0 & a_{i_1j_1} \\ a_{i_1j_1} & 0 \end{matrix}} & \begin{matrix} \cdots & 0 & 0 \\ \cdots & 0 & 0 \end{matrix} \\ \begin{matrix} \vdots & \vdots \end{matrix} & \begin{matrix} \ddots & \vdots & \vdots \end{matrix} \\ \begin{matrix} 0 & 0 \\ 0 & 0 \end{matrix} & \begin{matrix} \cdots \\ \cdots \end{matrix} \boxed{\begin{matrix} 0 & a_{i_qj_q} \\ a_{i_qj_q} & 0 \end{matrix}} \end{bmatrix},$$

where $a_{i_1j_1}, \ldots, a_{i_qj_q}$ are q nonzero distinct weights. Thus, $\tilde{A}_i$ has $2q$ simple nonzero eigenvalues.

If $|\mathcal{C}_i| = 1$, then the eigenvalue of $\tilde{A}_i \in \mathbb{R}^{1\times 1}$ can be placed to any value. If $|\mathcal{C}_i| = 2q+1$ and $q > 0$, then there are $2q$ vertices in $\mathcal{C}_i$ that can be covered by q cycles of length-2, and one vertex that cannot be covered by any length-2 cycle in a vertex-disjoint way in $G_{\mathcal{V}_{\mathcal{C}_i}}$. Assign distinct nonzero weights to $\star$-entries corresponding to the q cycles of length-2, and zero to other $\star$-entries in $\bar{A}_i$. As a result, the constructed numerical realization, $\tilde{A}_i$, has $2q$ nonzero simple eigenvalues and one zero eigenvalue. Denote by $\lambda_j(\tilde{A}_i)$ the jth eigenvalue of $\tilde{A}_i$, $j \in \{1, \ldots, |\mathcal{C}_i|\}$.

By Proposition 4, given a sufficiently small $\epsilon > 0$, $\exists \delta > 0$ and permutation $\sigma(\cdot)$ of integers $\{1, \ldots, |\mathcal{C}_i|\}$, such that for two numerical realizations of $\bar{A}_i$: $\tilde{A}_i$ and $\tilde{A}_{ip}$, if $||\tilde{A}_{ip} - \tilde{A}_i||_F < \delta$, then $\max\{|\lambda_{\sigma(j)}(\tilde{A}_{ip}) - \lambda_j(\tilde{A}_i)|\} < \epsilon$. Perturb $\star$-entries of $\tilde{A}_i$ corresponding to edges in

$\mathcal{E}_{\mathcal{C}_i}$, such that $\tilde{A}_{ip}$, which is derived by this perturbation of $\tilde{A}_i$, satisfies $||\tilde{A}_{ip} - \tilde{A}_i||_F < \delta$. Moreover, since t–rank$(\bar{A}_i) = 2q + 1$, by Lemma 1, g–rank$(\bar{A}_i) = 2q + 1$. The above analysis shows that we can perturb $\tilde{A}_i$, such that rank$(\tilde{A}_{ip}) = 2q + 1$, and

$$\min_{j \neq r, j, r \in \{1, \ldots, |\mathcal{C}_i|\}} |\lambda_j(\tilde{A}_{ip}) - \lambda_r(\tilde{A}_{ip})| >$$

$$\min_{j \neq r, j, r \in \{1, \ldots, |\mathcal{C}_i|\}} |\lambda_j(\tilde{A}_i) - \lambda_r(\tilde{A}_i)| - 2\epsilon.$$

It implies that there exists $\tilde{A}_{ip}$ which has $2q+1$ nonzero simple eigenvalues. Notice that $\tilde{A}_{ip}$ is also a numerical realization of $\bar{A}_i$. Hence, for either $|\mathcal{C}_i| = 2q$, or $|\mathcal{C}_i| = 2q + 1$, there exists a numerical realization $\tilde{A}_i$ such that $\tilde{A}_i$ has $|\mathcal{C}_i|$ nonzero simple eigenvalues. Also, there exists $\tilde{A}$ that has $\sum_{i=1}^{l} |\mathcal{C}_i| = k$ nonzero simple eigenvalues.

Denote by $\varphi'_{\tilde{A}}$ the derivative of $\varphi_{\tilde{A}}$ with respect to λ. If $\mathbf{p}_{\tilde{A}} \in V_0^c$, and $\tilde{A}$ has repeated nonzero modes, then $\varphi_{\tilde{A}}$ and $\varphi'_{\tilde{A}}$ have a common nontrivial zero (i.e., by Proposition 3, $R(\varphi_{\tilde{A}}, \varphi'_{\tilde{A}}) = 0$). Define $V_1 = \{\mathbf{p}_{\tilde{A}} \in \mathbb{R}^{n_{\bar{A}}} \colon a_{n-k} = 0 \text{ or } R(\varphi_{\tilde{A}}, \varphi'_{\tilde{A}}) = 0\}$, where $a_{n-k} = 0$ and $R(\varphi_{\tilde{A}}, \varphi'_{\tilde{A}}) = 0$ are both polynomials of $\star$-entries of $\bar{A}$. Since we have shown that there exists $\tilde{A}$ which has k nonzero simple eigenvalues, i.e., $\exists \mathbf{p}_{\tilde{A}} \in \mathbb{R}^{n_{\bar{A}}}$ such that $a_{n-k} \neq 0$ and $R(\varphi_{\tilde{A}}, \varphi'_{\tilde{A}}) \neq 0$, we conclude that V_1 is proper. □

Remark 20. *To characterize the generic rank of* $[\bar{A}, \bar{B}]$*, which is crucial in the derivation of Lemma 3, we should consider the proper variety in parameter space* $\mathbb{R}^{n_{\bar{A}}+n_{\bar{B}}}$*. Since each* $\star$*-entry of* $\bar{A}$ *is independent of those in* $\bar{B}$*,* V_1 *is also a proper variety in* $\mathbb{R}^{n_{\bar{A}}+n_{\bar{B}}}$*. Let us redefine* V_1 *as*

$$V_1 = \{[\mathbf{p}_{\tilde{A}}, \mathbf{p}_{\tilde{B}}] \in \mathbb{R}^{n_{\bar{A}}+n_{\bar{B}}} : a_{n-k} = 0 \text{ or } R(\varphi_{\tilde{A}}, \varphi'_{\tilde{A}}) = 0\}. \tag{A.8}$$

A.1.4 Proof of Lemma 3

We first introduce Lemma 17 in support of proving Lemma 3.

Lemma 17. *Consider an irreducible structural pair* $(\bar{A}, \bar{B})$*, where* $\bar{A} \in \{0, \star\}^{n \times n}$ *is a symmetrically structured matrix with* t–rank$(\bar{A}) = k$*. Let* $V_1 \subset \mathbb{R}^{n_{\bar{A}}+n_{\bar{B}}}$ *be defined as in* (A.8)*. There exists a proper variety* $V_2 \subset \mathbb{R}^{n_{\bar{A}}+n_{\bar{B}}}$ *such that if* $[\mathbf{p}_{\tilde{A}}, \mathbf{p}_{\tilde{B}}] \in V_1^c$*, then there exists a non-zero uncontrollable mode of* $\tilde{A}$ *if and only if* $[\mathbf{p}_{\tilde{A}}, \mathbf{p}_{\tilde{B}}] \in V_2$*.*

Sketch of Proof of Lemma 17. We will first prove that V_2 exists. Suppose $[\mathbf{p}_{\tilde{A}}, \mathbf{p}_{\tilde{B}}] \in V_1^c$, by a similar reasoning as in Lemma 2, all the k nonzero eigenvalues of $\tilde{A}$ are simple. Let λ be a nonzero eigenvalue of $\tilde{A}$, and $\varphi_{\tilde{A}}(s)$ be defined as in (A.6), then we have,

$$\varphi_{\tilde{A}}(\lambda) = \lambda^k + a_{n-1}\lambda^{k-1} + \cdots + a_{n-k} = 0. \tag{A.9}$$

Let us further assume that (λ, v) is an uncontrollable mode of $\tilde{A}$; in other words,

$$v^\top \tilde{A} = \lambda v^\top, \quad v^\top \tilde{B} = \mathbf{0}. \tag{A.10}$$

Since all the nonzero eigenvalues λ are simple, recall the fact in [35] that the left eigenvector $v^\top$ equals (apart from a constant scalar) any of the nonzero row of the adjugate matrix $\operatorname{adj}(\lambda I - \tilde{A})$. Hence,

$$\operatorname{adj}(\lambda I - \tilde{A})\tilde{B} = \mathbf{0}_{n\times m}. \tag{A.11}$$

Equations (A.9) and (A.11) imply that the two polynomials (A.12) and (A.13) have a common zero λ, namely,

$$\varphi_{\tilde{A}}(s) = s^k + a_{n-1}s^{k-1} + \cdots + a_{n-k} = 0, \tag{A.12}$$

$$\psi_{\tilde{A},\tilde{B}}(s) = \operatorname{tr}([\operatorname{adj}(sI - \tilde{A})\tilde{B}][\operatorname{adj}(sI - \tilde{A})\tilde{B}]^\top) = 0. \tag{A.13}$$

The variety V_2 is defined as follows,

$$V_2 = \{[\mathbf{p}_{\tilde{A}}, \mathbf{p}_{\tilde{B}}] \in \mathbb{R}^{n_{\bar{A}}+n_{\bar{B}}} : R(\varphi_{\tilde{A}}, \psi_{\tilde{A},\tilde{B}}) = 0\}, \tag{A.14}$$

where $R(\varphi_{\tilde{A}}, \psi_{\tilde{A},\tilde{B}}) = 0$ is a polynomial of the $\star$-entries in $\bar{A}$ and $\bar{B}$. The properness of V_2 can be shown by contradiction by adapting the proof in [35, Theorem 2]. Conversely, suppose $[\mathbf{p}_{\tilde{A}}, \mathbf{p}_{\tilde{B}}] \in V_2 \cap V_1^c$, by the definition of V_1 and V_2, $\varphi_{\tilde{A}}$ and $\psi_{\tilde{A},\tilde{B}}$ have a common zero $\lambda \neq 0$. Since λ is a zero of $\varphi_{\tilde{A}}$, λ is also an eigenvalue of $\tilde{A}$, which is an uncontrollable eigenvalue. □

Proof of Lemma 3. Define $V = V_1 \cup V_2$, where V_1 and V_2 are defined as in (A.8) and (A.14), respectively. We can prove V_1 is proper by a similar reasoning as the one in Lemma 2. By Lemma 17, V_2 is proper. Hence, $V = V_1 \cup V_2$ is proper. If $[\mathbf{p}_{\tilde{A}}, \mathbf{p}_{\tilde{B}}] \in V^c$, $\tilde{A}$ has k nonzero simple controllable modes. □

A.1.5 Proof of Theorem 1

We first introduce Lemma 18, which lays the foundation for the proof of Theorem 1.

Lemma 18. *Consider a structural pair $(\bar{A}, \bar{B})$, and a target set $\mathcal{T}$ with the corresponding state vertex set $\mathcal{X}_{\mathcal{T}}$ in $G(\bar{A}, \bar{B})$. We define $C_{\mathcal{T}}$ according to (2.4). Given a numerical realization $(\tilde{A}, \tilde{B})$, we define controllability matrix $Q(\tilde{A}, \tilde{B})$ as in Definition 1. Then, for any numerical realization $(\tilde{A}, \tilde{B})$, we have that* $\operatorname{rank}(C_{\mathcal{T}} Q(\tilde{A}, \tilde{B})) \leq |\mathcal{N}(\mathcal{X}_{\mathcal{T}})|$.

Proof of Lemma 18. Consider a numerical realization $(\tilde{A}, \tilde{B})$, from the Cayley-Hamilton theorem, we have that

$$\begin{aligned} \operatorname{rank}(C_{\mathcal{T}}[\tilde{B}, \tilde{A}Q(\tilde{A}, \tilde{B})]) &= \operatorname{rank}(C_{\mathcal{T}}[\tilde{B}, \tilde{A}\tilde{B}, \ldots, \tilde{A}^{n-1}\tilde{B}, \tilde{A}^n\tilde{B}]) \\ &= \operatorname{rank}([C_{\mathcal{T}}Q(\tilde{A}, \tilde{B}), C_{\mathcal{T}}\tilde{A}^n\tilde{B}]) \\ &= \operatorname{rank}(C_{\mathcal{T}}Q(\tilde{A}, \tilde{B})). \end{aligned} \tag{A.15}$$

In $G(\bar{A}, \bar{B})$, let m_1, m_2 be the number of input, state vertices in $\mathcal{N}(\mathcal{X}_{\mathcal{T}})$, respectively. Then, (A.15) yields,

$$\begin{aligned} \operatorname{rank}(C_{\mathcal{T}}Q(\tilde{A}, \tilde{B})) &= \operatorname{rank}(C_{\mathcal{T}}[\tilde{B}, \tilde{A}Q(\tilde{A}, \tilde{B})]) \\ &\leq \operatorname{rank}(C_{\mathcal{T}}\tilde{B}) + \operatorname{rank}(C_{\mathcal{T}}\tilde{A}Q(\tilde{A}, \tilde{B})) \\ &\leq m_1 + \min(\operatorname{rank}(C_{\mathcal{T}}\tilde{A}),\ \operatorname{rank}(Q(\tilde{A}, \tilde{B}))) \\ &\leq m_1 + m_2 \\ &= |\mathcal{N}(\mathcal{X}_{\mathcal{T}})|. \end{aligned}$$

This completes the proof. □

Proof of Theorem 1. To show the necessity of the theorem, suppose that there exists a vertex $x_i \in \mathcal{X}$ that is not input-reachable, then the i-th row of controllability matrix will be zero row, which implies that $\operatorname{rank}(Q(\tilde{A}, \tilde{B})) < n$, for any numerical realization of the pair $(\bar{A}, \bar{B})$. On the other hand, suppose there exists a set $\mathcal{S} \subseteq \mathcal{X}$, such that $|\mathcal{N}(\mathcal{S})| < |\mathcal{S}|$, then by Lemma 18, $\operatorname{rank}(Q(\tilde{A}, \tilde{B})) < n$, for any numerical realization of the pair $(\bar{A}, \bar{B})$. Hence, the necessity is proved.

To show the sufficiency, we proceed as follows. First, since $|\mathcal{N}(S)| \geq |\mathcal{S}|$, $\forall \mathcal{S} \subseteq \mathcal{X}$, it follows from Corollary 1 that g–rank$([\bar{A}, \bar{B}]) = n$. Because all the state vertices are input-reachable, $(\bar{A}, \bar{B})$ is irreducible. If we denote the term-rank of $\bar{A}$ as k, then by Lemma 3, there exists a proper variety $V \subset \mathbb{R}^{n_{\bar{A}}+n_{\bar{B}}}$ such that, if $[\mathbf{p}_{\tilde{A}}, \mathbf{p}_{\tilde{B}}] \in V^c$ then $\tilde{A}$ has k nonzero, simple and controllable modes. Let λ be an eigenvalue of $\tilde{A}$. On one hand, if $\lambda \neq 0$, then λ is controllable by Lemma 3. On the other hand, if $\lambda = 0$, since g–rank$([\bar{A}, \bar{B}]) = n$, then there exists a proper variety $W \subset \mathbb{R}^{n_{\bar{A}}+n_{\bar{B}}}$, such that if $[\mathbf{p}_{\tilde{A}}, \mathbf{p}_{\tilde{B}}] \in W^c \cap V^c$, then rank$([\tilde{A}, \tilde{B}]) = n$. As a result, $\lambda = 0$ is controllable by the eigenvalue PBH test. Since all the modes of $\tilde{A}$ are controllable generically, $(\bar{A}, \bar{B})$ is structurally controllable. □

A.1.6 Proof of Theorem 2

Proof. The necessity of Conditions *1)* and *2)* can be proved in a similar approach as the proof in Theorem 1. What remains to be shown is their sufficiency. It suffices to show that Conditions *1)* and *2)* result in that generically the left null space of target controllability matrix is trivial.

Suppose there exists an input-unreachable state vertex $x_i \in \mathcal{X} \backslash \mathcal{X}_{\mathcal{T}}$. Since all the vertices in $\mathcal{X}_{\mathcal{T}}$ are input-reachable, for $\forall x_j \in \mathcal{X}_{\mathcal{T}}$, there is no path from x_j to x_i, and there is also no path from x_i to x_j due to the symmetry in $G(\bar{A})$. This implies in model (2.2) that the ith state has no impact on the dynamics of $\mathcal{T}$ corresponding states. Omitting the ith state from the system will not change the dynamics of $\mathcal{T}$ corresponding states. Hence, we could assume that $(\bar{A}, \bar{B})$ is irreducible. By Lemma 3, there exists a proper variety $V \subset \mathbb{R}^{n_{\bar{A}}+n_{\bar{B}}}$, such that if $[\mathbf{p}_{\tilde{A}}, \mathbf{p}_{\tilde{B}}] \in V^c$, then all the nonzero modes of $\tilde{A}$ are controllable. In the rest of the proof, we assume $[\mathbf{p}_{\tilde{A}}, \mathbf{p}_{\tilde{B}}] \in V^c$. Denote by $e_1, \ldots, e_l$ the left eigenvectors corresponding to zero modes of $\tilde{A}$, and $e_{l+1}, \ldots, e_n$ the left eigenvectors for nonzero modes. Denote the left null space of a matrix M as $\boldsymbol{N}(M^\top)$.

From Lemma 3, we have that if $[\mathbf{p}_{\tilde{A}}, \mathbf{p}_{\tilde{B}}] \in V^c$, then $\boldsymbol{N}((Q(\tilde{A}, \tilde{B}))^\top) \subseteq \text{span}\{e_1^\top, \ldots, e_l^\top\}$. For the target set $\mathcal{T}$, define the matrix $C_{\mathcal{T}}$ according to (2.4). By the assumption $|\mathcal{N}(\mathcal{S})| \geq |\mathcal{S}|$, $\forall \mathcal{S} \subseteq \mathcal{X}_{\mathcal{T}}$, and Corollary 1, we have that g–rank$(C_{\mathcal{T}}[\bar{A}, \bar{B}]) = |\mathcal{T}|$, which implies that there exists a proper variety $W \subset \mathbb{R}^{n_{\bar{A}}+n_{\bar{B}}}$, such that if $[\mathbf{p}_{\tilde{A}}, \mathbf{p}_{\tilde{B}}] \in V^c \cap W^c$, then $\mathrm{rank}(C_{\mathcal{T}}[\tilde{A},\tilde{B}]) = |\mathcal{T}|$, i.e., $\boldsymbol{N}((C_{\mathcal{T}}[\tilde{A},\tilde{B}])^\top) = \mathbf{0}$. Define $\hat{I} \in \mathbb{R}^{n\times n}$ as

$$[\hat{I}]_{ij} = \begin{cases} 1, & \text{if } j = i,\ i \in \mathcal{T}, \\ 0, & \text{otherwise.} \end{cases} \tag{A.16}$$

We claim that there does not exist a nontrivial vector $e \in \mathbb{C}^n$ such that $\hat{I}e = e$, $e^\top \tilde{A} = 0e^\top$ and $e^\top \tilde{B} = \mathbf{0}$. Otherwise, $e^\top[\tilde{A},\tilde{B}] = \mathbf{0}$, which contradicts $\boldsymbol{N}((C_{\mathcal{T}}[\tilde{A},\tilde{B}])^\top) = \mathbf{0}$.

Hence, if $[\mathbf{p}_{\tilde{A}}, \mathbf{p}_{\tilde{B}}] \in V^c \cap W^c$, then there is no nontrivial vector $v \in \mathbb{C}^{|\mathcal{T}|}$, such that $v^\top C_{\mathcal{T}} \in \mathrm{span}\{e_1^\top, \ldots, e_l^\top\}$. Thus, generically, $\boldsymbol{N}((C_{\mathcal{T}}Q(\tilde{A},\tilde{B}))^\top) = \mathbf{0}$. The $(\bar{A},\bar{B})$ is structurally target controllable with respect to $\mathcal{T}$. □

A.1.7 Proofs of Theorem 3 and Theorem 4

Proof of Theorem 3. ($\Longleftarrow$)Suppose there exists a target set $\mathcal{T} = \{t_i\}_{i=1}^k \subseteq [n]$ such that there is no right-unmatched vertex in $\mathcal{B}(\mathcal{X}_{\mathcal{T}}, \mathcal{Y}, \mathcal{E}_{\mathcal{X}_{\mathcal{T}},\mathcal{Y}})$, and $(\bar{A},\bar{B})$ is structurally target controllable with respect to $\mathcal{T}$. We construct $\tilde{C} \in \{0,1\}^{k\times n}$ such that $[\tilde{C}]_{it_i} = 1$ and $\sum_{j=1}^n [\tilde{C}]_{ij} = 1, \forall i \in [k]$. Since $(\bar{A},\bar{B})$ is structurally target controllable with respect to $\mathcal{T}$, there exist numerical realizations $\tilde{A},\tilde{B}$ such that $\tilde{C}\cdot Q(\tilde{A},\tilde{B})$ is full row rank. Thus, $(\bar{A},\bar{B},\bar{C})$ is structurally output controllable.

($\Longrightarrow$)We approach the proof by contraposition. Suppose for all target sets $\mathcal{T} \subseteq [n]$ with respect to which $(\bar{A},\bar{B})$ is structurally target controllable, there exists at least one right-unmatched vertex in $\mathcal{B}(\mathcal{X}_{\mathcal{T}}, \mathcal{Y}, \mathcal{E}_{\mathcal{X}_{\mathcal{T}},\mathcal{Y}})$. Then, by taking a similar reasoning used in the proof of Lemma 18, we can show that $\mathrm{rank}(\tilde{C}\cdot Q(\tilde{A},\tilde{B})) \leq k, \forall \tilde{C} \in \mathbb{R}^{k\times n}$, which implies $(\bar{A},\bar{B},\bar{C})$ is not structurally output controllable. □

Sketch of Proof of Theorem 4. The NP-hardness can be proved by reducing a general instance of 3-dimensional matching problem [212][p.46] to an instance of the structural output controllability problem. More specifically, the elements in the three dimensions $\mathcal{S}_1 \times \mathcal{S}_2 \times \mathcal{S}_3$ of the 3-dimensional matching problem are recast as vertices in $\mathcal{U} \times \mathcal{X} \times \mathcal{Y}$, where $\mathcal{U}, \mathcal{X}$, and $\mathcal{Y}$ are input, state and output vertices, respectively. The links in $\mathcal{S}_1 \times \mathcal{S}_2$ and in $\mathcal{S}_2 \times \mathcal{S}_3$ are recast as edges in $\mathcal{E}_{\mathcal{U},\mathcal{X}}$, and $\mathcal{E}_{\mathcal{X},\mathcal{Y}}$, respectively. We let $[\bar{A}]_{ij} = 0$, for $\forall i, j \in [|\mathcal{X}|]$; $[\bar{B}]_{ij} = \star$ if $(u_j, x_i) \in \mathcal{E}_{\mathcal{U},\mathcal{X}}$ and $[\bar{B}]_{ij} = 0$ otherwise; $[\bar{C}]_{ij} = \star$ if $(x_j, y_i) \in \mathcal{E}_{\mathcal{X},\mathcal{Y}}$ and $[\bar{C}]_{ij} = 0$ otherwise. By Theorem 3, the constructed structural system $(\bar{A}, \bar{B}, \bar{C})$ is structurally output controllable if there exists a 3-dimensional matching in $\mathcal{S}_1 \times \mathcal{S}_2 \times \mathcal{S}_3$. Since such a reduction can be completed in polynomial time, the problem of verifying both conditions in Theorem 3 is NP-hard. □

A.2 Proof of the results in Chapter 3

A.2.1 Proof of Theorem 6

Proof of Theorem 6. First, we show that if the set of edges $\tilde{\mathcal{E}}$ contains S_M and S_B as subsets, then it must be a feasible edge-addition configuration. We notice that, given the system digraph $G(\bar{A}, \bar{B}) = (\mathcal{X} \cup \mathcal{U}, \mathcal{E}_{\mathcal{X},\mathcal{X}} \cup \mathcal{E}_{\mathcal{U},\mathcal{X}})$, it suffices to show that $S_M \cup S_B$ satisfies both conditions in Theorem 5 when the graph $G_{aug} \equiv (\mathcal{X} \cup \mathcal{U}, \mathcal{E}_{\mathcal{X},\mathcal{X}} \cup \mathcal{E}_{\mathcal{U},\mathcal{X}} \cup S_M \cup S_B)$ is considered. Hereafter, we denote the bipartite representation of G_{aug} by $\mathcal{B}_{aug} \equiv \mathcal{B}(\mathcal{X}^+ \cup \mathcal{U}^+, \mathcal{X}^-, \mathcal{E}_{\mathcal{X}^+,\mathcal{X}^-} \cup \mathcal{E}_{\mathcal{U}^+,\mathcal{X}^-} \cup S_M^{\pm} \cup S_B^{\pm})$, where $S_M^{\pm} = \{s(e) \colon e \in S_M\}$ and $S_B^{\pm} = \{s(e) \colon e \in S_B\}$.

To verify Condition (a) of Theorem 5, we decompose the set of state vertices $\mathcal{X}$, into $\mathcal{R}_1$ and $\mathcal{N}$ based on their reachability as in Definition 4. Specifically, $\mathcal{R}_1$ contains all the reachable state vertices and $\mathcal{N}$ contains all the unreachable state vertices. Since $\mathcal{N} = \bigcup_{h=1}^{r} \Delta(\mathcal{N}_{t_h})$, every state vertex $v \in \mathcal{N}$ must be contained in some $\Delta(\mathcal{N}_{t_h})$ for some iteration step h. By the recursive construction of the bridging set S_B as described in Definition 4, $\mathcal{N}_{t_h}$ is reachable provided that $\mathcal{N}_{t_{h-1}}$ is also reachable. Thus, we conclude that all $v \in \mathcal{N}$ become reachable in G_{aug}.

To verify Condition (b) of Theorem 5, let M be a maximum matching associated with the system bipartite graph. Next, we propose to consider a bipartite graph $\mathcal{B}_M \equiv \mathcal{B}(\mathcal{X} \cup \mathcal{U}, \mathcal{E}_{\mathcal{X},\mathcal{X}} \cup \mathcal{E}_{\mathcal{U},\mathcal{X}} \cup S_M^{\pm})$, which is a sub-graph of the bipartite graph $\mathcal{B}_{aug}$. By the construction of S_M, $M \cup S_M$ is a matching in $\mathcal{B}_M$. Furthermore, it is a maximum matching since it has no right-unmatched vertices in $\mathcal{B}_M$. Since $\mathcal{B}_{aug}$ has the same set of vertices as $\mathcal{B}_M$, it follows that $M \cup S_M$ is also a maximum matching associated with $\mathcal{B}_{aug}$. Subsequently, $M \cup S_M$ satisfies Condition (b) in Theorem 5 for the system bipartite graph $\mathcal{B}_{aug}$.

Therefore, if $S_M \cup S_B$ is added to the system digraph, the resulting system is structurally controllable, which implies that $S_M \cup S_B$ is a feasible edge-addition configuration.

Next, we show that if $\tilde{\mathcal{E}}$ is a feasible edge-addition configuration, then it must contain the union of the two sets as described in the theorem. Assume, by contradiction, that there is no such S_B in $\tilde{\mathcal{E}}$, then there is a source SCC containing only state vertices that is unreachable. This implies that none of its states are reachable, which precludes the Condition (a) in Theorem 5 to hold; hence, a contradiction is attained. On the other hand, assume that for any maximum matchings M associated with $\mathcal{B}(\bar{A}, \bar{B})$, we have $S_M \setminus \tilde{\mathcal{E}} \neq \emptyset$, then there exists at least one right-unmatched vertex corresponding to the head of an edge in $S_M^{\pm} \setminus M$, which precludes Condition (b) in Theorem 5 to hold; hence, a contradiction is attained. Thus, a set $\tilde{\mathcal{E}}$ is a feasible edge-addition configuration if and only if it contains S_M and S_B as subsets. □

A.2.2 Proof of Corollary 2

Proof of Corollary 2. From Theorem 6, any feasible edge-addition configuration contains S_M, for some maximum matching M associated with the system bipartite graph, and S_B, the bridging edges as subsets, i.e., $\tilde{\mathcal{E}} \supseteq S_M \cup S_B$. Consequently, an optimal edge-addition configuration should satisfy $|\tilde{\mathcal{E}}^*| \geq |S_M| = n_r$ and $|\tilde{\mathcal{E}}^*| \geq |S_B| = r$. □

A.2.3 Proof of Theorem 7 and Theorem 8

Proof of Theorem 7. Briefly, the proof requires the following steps. First, we show that an optimal edge-addition configuration $\tilde{\mathcal{E}}^*$ must satisfy $|\tilde{\mathcal{E}}^*| \geq n_r + r - q$. Then, we construct a feasible edge-addition configuration such that its cardinality achieves $n_r + r - q$.

From Theorem 6, a feasible edge-addition configuration must satisfy $\tilde{\mathcal{E}} \supseteq S_M \cup S_B$. As a result, the cardinality of a feasible edge-addition configuration should satisfy $|\tilde{\mathcal{E}}| \geq |S_M \cup S_B|$, which implies that $|\tilde{\mathcal{E}}| \geq |S_M| + |S_B| - |S_M \cap S_B|$. Notice that, $S_M = n_r$ and $|S_B| = r$, then $|\tilde{\mathcal{E}}| \geq n_r + r - |S_M \cap S_B|$. Thus, an optimal edge-addition configuration, which we denote as $\tilde{\mathcal{E}}^*$, must satisfy $|\tilde{\mathcal{E}}^*| \geq n_r + r - \max_{M,S_B} |S_M \cap S_B|$, where the maximum is taken over all possible maximum matchings M of the system bipartite graph and possible bridging sets S_B for the system digraph. To obtain the value of $\max_{M,S_B} |S_M \cap S_B|$, we recall that maximizing the intersection between S_M and S_B gives the maximum number of right-unmatched vertices across all possible maximum matchings associated with $\mathcal{B}(\bar{A}, \bar{B})$ in the unreachable source SCCs, i.e., the unreachable source assignability number q, from Definition 6. Therefore, we have that $\max_{M,S_B} |S_M \cap S_B| = q$, which implies that $|\tilde{\mathcal{E}}^*| \geq n_r + r - q$. Next, we show that there exists a feasible edge-addition configuration that achieves $p^* = n_r + r - q$, which we approach by construction.

Given the system digraph $G(\bar{A}, \bar{B})$, we partition its state vertices based on reachability. Specifically, we denote $\mathcal{R}_1$ as the set of all reachable state vertices and $\mathcal{N}$ as the set of all unreachable state vertices. Moreover, we use $\mathcal{N}_1, \ldots, \mathcal{N}_r \subseteq \mathcal{N}$ to denote the vertex sets of r source SCCs that are unreachable, as in Definition 4. Furthermore, let G_r be the $\mathcal{R}_1$-induced subgraph of $G(\bar{A}, \bar{B})$.

Next, we obtain a maximum matching $\bar{M}$ that attains the USAN using Algorithm 2. Without loss of generality, we assume there are q unreachable-assignable source SCCs whose vertex sets are denoted as $\mathcal{N}_1, \ldots, \mathcal{N}_q$ with $q \leq r$. Let $U_L^{\mathcal{X}}(\bar{M})$ and $U_R(\bar{M})$ be the set of left-unmatched and right-unmatched state vertices associated with $\bar{M}$, respectively. We can obtain a digraph $G(\mathcal{V}(s^{-1}(\bar{M})), \mathcal{E}(s^{-1}(\bar{M})))$ from $\bar{M}$, where $\mathcal{E}(s^{-1}(\bar{M}))) = \{s^{-1}(e) : e \in \bar{M}\}$ and $\mathcal{V}(s^{-1}(\bar{M}))$, the vertices used by the edges belonging to $\mathcal{E}(s^{-1}(\bar{M}))$. In particular, the set of edges $\mathcal{E}(s^{-1}(\bar{M}))$ is spanned by a disjoint union of paths $\{\mathcal{P}_i\}_{i \in \mathcal{I}}$ and cycles $\{\mathcal{C}_j\}_{j \in \mathcal{J}}$, where $\mathcal{I}$ and $\mathcal{J}$ denote their indices. Furthermore, to construct an optimal edge-addition configuration, we define the following sets according to the correspondence between the maximum matching attaining the USAN q and the path and cycle decomposition captured by $G(\mathcal{V}(s^{-1}(\bar{M})), \mathcal{E}(s^{-1}(\bar{M})))$. Let $\mathcal{V}_L$ be the set of ending vertices of paths in $\{\mathcal{P}_i\}_{i\in\mathcal{I}}$ whose starting vertex is in $\mathcal{U}$. Let $\mathcal{S}$ be the set containing q starting vertices corresponding to disjoint paths in $\{\mathcal{P}_i\}_{i\in\mathcal{I}}$ and belonging to different unreachable source SCCs. Lastly, let $\mathcal{S}^{\pm} = \{x_i^+ \colon x_i \in \mathcal{V}_L\}$, which by construction is a subset of left-unmatched vertices associated with $\bar{M}$. Thus, either $U_L^{\mathcal{X}}(\bar{M}) \cap \mathcal{S}^{\pm} \neq \emptyset$ or $U_L^{\mathcal{X}}(\bar{M}) \cap \mathcal{S}^{\pm} = \emptyset$ holds.

We now begin to construct a feasible edge-addition configuration that achieves p^* under the assumption that $U_L^{\mathcal{X}}(\bar{M}) \cap \mathcal{S}^{\pm} \neq \emptyset$ holds. We first initialize $\tilde{\mathcal{E}}^*$ to be an empty set. Then, at the initialization ($k = 1$), we add an edge (v_1, z_1) into $\tilde{\mathcal{E}}^*$, where $v_1^+ \in U_L^{\mathcal{X}}(\bar{M}) \cap \mathcal{S}^{\pm}$ and z_1^- is a right-unmatched vertex associated with $\bar{M}$ in some unreachable source SCCs, i.e., $z_1 \in \mathcal{N}_l$ for some $l \in \{1, \ldots, q\}$. Since $v_1^+ \in \mathcal{S}^{\pm}$, it follows that $v_1 \in \mathcal{R}_1$. Consequently, if we add the edge (v_1, z_1) to the system digraph, then the vertex z_1 becomes reachable, which implies that all the state vertices in $\Delta(\mathcal{N}_l)$ become reachable as well. On the other hand, if $z_1^- \in U_R(\bar{M})$, then there must exist a path in $G(\mathcal{V}(s^{-1}(\bar{M})), \mathcal{E}(s^{-1}(\bar{M})))$ departing from z_1. In addition, the end of this path is a left-unmatched state vertex $v_2^+ \in U_L^{\mathcal{X}}(\bar{M})$ with $v_2^+ \neq v_1^+$. In particular, $v_2 \in \Delta(\mathcal{N}_l)$ since it is reachable from z_1. Then, we can add another edge departing from v_2^+ to another right-unmatched vertex z_2^- in a different unreachable source SCC, i.e., to add the edge (v_2, z_2) to $\tilde{\mathcal{E}}$. We iterate this procedure for another $q - 1$ steps, i.e., $k = 2, \ldots, q$, until all q unreachable-assignable SCCs become reachable by adding edges into $\tilde{\mathcal{E}}^*$.

Now, without loss of generality, let $\tilde{\mathcal{E}}^* = \{(v_k, z_k) \colon k = 1, \ldots, q\}$, where $v_k^+ \in U_L^{\mathcal{X}}(\bar{M})$ and $z_k^- \in U_R(\bar{M})$ for all $k = 1, \ldots, q$, respectively. Nonetheless, there are $r - q$ remaining unreachable source SCCs, i.e., $\mathcal{N}_{q+1}, \ldots, \mathcal{N}_r$. To ensure reachability of all state vertices, it suffices to add edges from the set of reachable state vertices to each one of the remaining unreachable source SCCs. Consequently, the complementary set of edges to account in $\tilde{\mathcal{E}}^*$ is a set of bridging edges containing r edges by Definition 4. However, as implied by Theorem 6, to construct a feasible edge-addition configuration, we still need to include S_M as a subset. Towards this service, we notice that q right-unmatched vertices, i.e., those in the unreachable-assignable SCCs, have been matched during the iterative procedure. Consequently, it suffices to add $n_r - q$ edges to ensure that all the remaining right-unmatched state vertices are matched, i.e., those in $U_R(\bar{M}) \setminus \{z_1^-, \ldots, z_q^-\}$. As such, we have constructed a set of edges considered to be added, i.e., $\tilde{\mathcal{E}}^*$, that contains a set of bridging edges and $S_{\bar{M}}$ for the maximum matching $\bar{M}$. As a result, $\tilde{\mathcal{E}}^*$ is a feasible edge-addition configuration by Theorem 6. In addition, it contains $n_r + r - q$ edges, which implies that it is an optimal edge-addition configuration – the construction considered in this paragraph leads to Step 4 of Algorithm 3.

Next, we discuss the case when $U_L^{\mathcal{X}}(\bar{M}) \cap \mathcal{S}^{\pm} = \emptyset$. First, we define $\mathcal{G}_r^{\pm} = \{x_i^- : x_i \in \mathcal{R}_1\}$ as the set of left-unmatched state vertices in G_r. As a consequence, two particular cases may happen: either $U_L^{\mathcal{X}}(\bar{M}) \cap G_r^{\pm} = \emptyset$ or $U_L^{\mathcal{X}}(\bar{M}) \cap G_r^{\pm} \neq \emptyset$ holds. Consider the first case, where $U_L^{\mathcal{X}}(\bar{M}) \cap G_r^{\pm} = \emptyset$, since $U_L^{\mathcal{X}}(\bar{M}) \cap \mathcal{S}^{\pm} = \emptyset$, then the subgraph of $G(\mathcal{V}(s^{-1}(\bar{M})), \mathcal{E}(s^{-1}(\bar{M})))$ constrained to the vertices in G_r consists only of cycles. Therefore, and without loss of generality, we let c_r be the number of those cycles, whose set of vertices are denoted as $\mathcal{C}_i, i = 1, \ldots, c_r$. According to the assumption $\|\bar{B}\|_0 \neq 0$, there exists an edge $(u, v) \in \mathcal{E}_{\mathcal{U},\mathcal{X}}$, with $u \in \mathcal{U}$ and $v \in \mathcal{V}$. Additionally, v belongs to the vertex set of some cycle, i.e., $v \in \mathcal{C}_j$ for some $j \leq c_r$, which we represent by the ordered sequence $(v, v_1, \ldots, v_k, v)$. If we replace the cycle $(v, v_1, \ldots, v_k, v)$ by the path $(u, v, v_1, \ldots, v_k)$, then the new digraph will correspond to another maximum matching $\hat{M}$ associated with $\mathcal{B}(\bar{A}, \bar{B})$ with a reachable left-unmatched state vertex v_k. Additionally, $U_L^{\mathcal{X}}(\hat{M}) \cap \mathcal{S}^{\pm} \neq \emptyset$, and, as a result, we may reduce the case with assumptions $U_L^{\mathcal{X}}(\bar{M}) \cap \mathcal{S}^{\pm} = \emptyset$ and $U_L^{\mathcal{X}}(\bar{M}) \cap G_r^{\pm} = \emptyset$ to the case previously discussed by constructing a new maximum matching $\hat{M}$ – this procedure corresponds to steps 3 – 9 in Algorithm 3.

Now, we suppose that $U_L^{\mathcal{X}}(\bar{M}) \cap \mathcal{S}^{\pm} = \emptyset$ and $U_L^{\mathcal{X}}(\bar{M}) \cap G_r^{\pm} \neq \emptyset$ hold simultaneously. Then, there exists $v_1 \in U_L^{\mathcal{X}}(\bar{M}) \cap G_r^{\pm}$ and $v_r \in U_R(\bar{M})$ such that $(v_r, \ldots, v_1)$ is a path whose edges are associated with those in $\bar{M}$ through a signal-notation mapping. In particular, $v_r \notin \mathcal{U}$. If v_r is not a vertex in some unreachable source SCCs, then we may apply the procedure introduced in the case when $U_L^{\mathcal{X}}(\bar{M}) \cap \mathcal{S}^{\pm} \neq \emptyset$ to construct a feasible edge-addition configuration containing p^* edges. Nonetheless, if v_r is a vertex in some unreachable source SCCs, then a modification of the iterative construction must be adopted. Specifically, recall that previously, at the basis step of iteration, we add

(v_1, z_1) into $\tilde{\mathcal{E}}^*$, in which $z_1 \in \mathcal{N}_l$ is arbitrarily chosen. Now, if z_1 is chosen to be equal to v_r, then (v_1, v_r) is added into $\tilde{\mathcal{E}}^*$ and follow-up iteration steps cannot be performed since the end of the path starting at z_1 is v_1. Consequently, we must adopt the following modification: if $q = 1$, then we must add an edge (v_1, v_r) into $\tilde{\mathcal{E}}^*$; otherwise, we add an edge (v_1, z_1) into $\tilde{\mathcal{E}}^*$ with $z_1^- \in U_R(\bar{M})$ being a vertex in some unreachable source SCCs and $z_1 \neq v_r$ at the basis step. In other words, when constructing the first q steps of a feasible edge-addition configuration, we force $z_i^- \in U_R(\bar{M})$, $z_i \in \mathcal{N}_l$ and $z_i \neq v_r$ for all $i = 1, \ldots, q-1$ and $z_q = v_r$, whereas the rest of the construction readily follows as previously discussed. As such, we can obtain a feasible edge-addition configuration achieving p^* if $U_L^{\mathcal{X}}(\bar{M}) \cap \mathcal{S}^{\pm} = \emptyset$ and $U_L^{\mathcal{X}}(\bar{M}) \cap G_r^{\pm} \neq \emptyset$ simultaneously – this construction procedure is summarized in steps 20 – 32 in Algorithm 3.

Therefore, we conclude that if $\|\bar{B}\|_0 > 0$, we can construct a feasible edge-addition configuration achieving $p^* = n_r + r - q$. □

Proof of Theorem 8. The correctness of the algorithm follows from the proof of Theorem 7. To determine the computational complexity of the algorithm, we consider the computational complexity incurred by each one of the major steps in the algorithm. Specifically, Step 1 requires the computation of strongly connected components, which can be achieved by applying the depth-first search algorithm twice with complexity $\mathcal{O}(|\mathcal{X} \cup \mathcal{U}| + |\mathcal{E}_{\mathcal{X},\mathcal{X}} \cup \mathcal{E}_{\mathcal{U},\mathcal{X}}|)$ [154]. Finding a minimum-weighted maximum matching in Step 2 incurs in $\mathcal{O}(|\mathcal{X} \cup \mathcal{U}|^3)$, and can be achieved as described in Algorithm 2, and we can guarantee that exists at least one left-unmatched vertex of $\bar{M}$ that is reachable in $\mathcal{O}(|\mathcal{X}|)$. In Step 3, we iteratively construct an optimal edge-addition configuration as described in the proof of Theorem 7, which can be attained in $\mathcal{O}(|\mathcal{X}| + |\mathcal{U}|)$, since it searches over the computed maximum matching and the source SCCs in the system digraph. Finally, in Step 4, we add the remaining edges to ensure conditions in Theorem 6, which incurs in $\mathcal{O}(|\mathcal{X}|)$. In summary, the computational complexity of Algorithm 8 is dominated by the second step, which implies an overall computational complexity in $\mathcal{O}(|\mathcal{X} \cup \mathcal{U}|^3)$. □

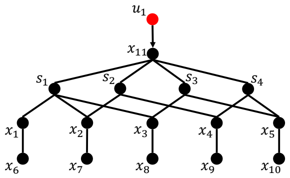


Figure A-1: Example of the construction of $\mathcal{G}(\mathcal{X}\cup\mathcal{U},\mathcal{E}_u,\mathcal{E}_{\mathcal{U},\mathcal{X}})$ from sets $\mathcal{U}_{\mathcal{S}}=\{\mathcal{S}_\ell\}_{\ell=1}^p$ in the proof of Theorem 9. Suppose we have $\mathcal{U}_{\mathcal{S}}=\{1,2,3,4,5\}$, $\mathcal{S}_1=\{1,2,3\},\mathcal{S}_2=\{2,4\},\mathcal{S}_2=\{3,5\},\mathcal{S}_4=\{4,5\}$. Then $\mathcal{X}=\{x_i\}_{i=1}^{11}\cup\{s_i\}_{i=1}^4$, $\mathcal{T}=\{1,2,\ldots,10\}$ and $\mathcal{U}=\{u_1\}$ are the set of state vertices, target set and the set of input vertex, respectively.

A.3 Proof of the results in Chapter 4

A.3.1 Proof of Theorem 9

Proof. Given a structural pair $(\hat{A},\bar{B})$ and target set $\mathcal{T}\subseteq[n]$, where $\hat{A}$ is symmetrically structured, we can verify in polynomial time whether the pair $(\hat{A},\bar{B})$ is structurally target controllable with respect to $\mathcal{T}$ using Theorem 2. Thus, the Problem 4 is in NP. To show the NP-hardness of Problem 4, we reduce a general min-set-cover problem instance to an instance of Problem 4. More specifically, a general min-set-cover problem instance admits following elements, (i) a universe $\mathcal{U}_{\mathcal{S}}=[n]$; (ii) a collection of sets $\{\mathcal{S}_\ell\}_{\ell=1}^p$, where $\mathcal{U}_{\mathcal{S}}=\bigcup_{\ell=1}^p\mathcal{S}_\ell$. Based on (i) and (ii), we construct a mixed graph $\mathcal{G}(\mathcal{X}\cup\mathcal{U},\mathcal{E}_u,\mathcal{E}_{\mathcal{U},\mathcal{X}})$ as follows: we let $\mathcal{X}=\{x_i\}_{i=1}^{2n+1}\cup\{s_\ell\}_{\ell=1}^p$, $\mathcal{U}=\{u_1\}$, $\mathcal{E}_u=\{\{x_i,x_{i+n}\}\}_{i=1}^n\cup\{\{x_i,s_\ell\}\colon i\in\mathcal{S}_\ell,\ell\in[p]\}\cup\{\{s_\ell,x_{2n+1}\}\}_{\ell=1}^p$ and $\mathcal{E}_{\mathcal{U},\mathcal{X}}=\{(u_1,x_{2n+1})\}$, and we let $\bar{B}\in\{0,\star\}^{(2n+1+\ell)\times 1},\bar{B}=[0,\ldots,0,\star]^\top$, be the structural pattern of input matrix. We define a target set $\mathcal{T}=[2n]$ such that the target vertex set is $\mathcal{X}_{\mathcal{T}}=\{x_i\}_{i=1}^{2n}$. See Figure A-1 for an illustration of such construction. We define a cost function $c(e),\forall e\in\mathcal{X}\times\mathcal{X}$, as

$$c(e)=\begin{cases}1, & \text{if } e\in\mathcal{E}_u,\\ \infty, & \text{otherwise.}\end{cases}\tag{A.17}$$

From our construction of $\mathcal{G}(\mathcal{X}\cup\mathcal{U},\mathcal{E}_u,\mathcal{E}_{\mathcal{U},\mathcal{X}})$, we have $|\mathcal{N}(\mathcal{S})|\geq|\mathcal{S}|,\forall\mathcal{S}\subseteq\mathcal{X}$, and $\forall x_i\in\mathcal{X}$ are reachable. The structural pair, whose mixed graph representation is denoted by $\mathcal{G}(\mathcal{X}\cup\mathcal{U},\mathcal{E}_u,\mathcal{E}_{\mathcal{U},\mathcal{X}})$, is structurally controllable, i.e., the constructed instance of Problem 4 with $c(e)\in\{1,\infty\}$ is well-defined.

Next, we show a minimum solution of the constructed instance of Problem 4 yields a minimum solution of min-set-cover problem. Suppose we have a minimum solution $\mathcal{E}_u(\bar{A}^\star)$ of the constructed instance of Problem 4, then we claim that the minimum cost is $2n+\mu$, for some $\mu \geq 1$, otherwise the two conditions in Theorem 2 cannot be simultaneouly satisfied when considering the target vertex set $\mathcal{X}_{\mathcal{T}} = \{x_i\}_{i=1}^{2n}$. Let $\{\ell_i\}_{i=1}^{\mu}$ be the indexes of the vertices which are in $\{s_\ell\}_{\ell=1}^{p}$ and are incident to edges in $\mathcal{E}_u(\bar{A}^\star)$. Since all the target vertices are reachable from input vertex, it implies that in the min-set-cover problem the subsets $\{\mathcal{S}_{\ell_i}\}_{i=1}^{\mu}$ constitute a feasible solution to the given instance of min-set-cover problem. Furthermore, suppose there exists a union of subsets $\bigcup_{i'=1}^{\mu'} \mathcal{S}_{\ell_{i'}} = \mathcal{U}_{\mathcal{S}}$ and $\mu' < \mu$, then from $\{\mathcal{S}_{\ell_{i'}}\}_{i'=1}^{\mu'}$ we can construct a set of sets $\{\mathcal{S}'_{\ell_{i'}}\}_{i'=1}^{\mu'}$ such that (i) for $\forall i', j' \in [\mu']$, $\mathcal{S}'_{\ell_{i'}} \cap \mathcal{S}'_{\ell_{j'}} = \emptyset$ and (ii) for $\forall i' \in [\mu']$, $\mathcal{S}'_{\ell_{i'}} \subseteq \mathcal{S}_{\ell_{i'}}$. Based on the set $\{\mathcal{S}'_{\ell_{i'}}\}_{i'=1}^{\mu'}$, we can construct a feasible solution to Problem 4, $\mathcal{E}_u(\hat{A}) = \{\{x_i, x_{i+n}\}\}_{i=1}^{n} \cup \{\{s_{\ell_{j'}}, x_{2n+1}\}\}_{j'=1}^{\mu'} \cup \{\{x_i, s_{\ell_{j'}}\} \colon i \in \mathcal{S}_{\ell_{j'}}, j' \in [\mu']\}$ with a total cost $2n+\mu' < 2n+\mu$. Then it contradicts that $\mathcal{E}_u(\bar{A}^\star)$ is a minimum solution to the constructed instance of Problem 4. Hence, a minimum solution of the constructed instance of Problem 4 leads to a minimum solution of min-set-cover problem. Consequently, we have shown that a general instance of min-set-cover problem can be reduced to an instance of Problem 4 in polynomial time. Therefore, the Problem 4 is thus NP-hard. □

A.3.2 Proof of Theorem 10

Proof. See the Proof of Theorem 9. □

A.3.3 Proof of Theorem 11 and related results

Proof of Lemma 4. ($\Longrightarrow$) Before we derive the two statements of Lemma 4, we first prove that t–rank$(\bar{A}) \geq n-1$, when $(\bar{A}, \bar{B})$ is structurally controllable under Assumption 2. By Theorem 1 and Proposition 2, $(\bar{A}, \bar{B})$ is structurally controllable if and only if $\forall x_i \in \mathcal{X}$ is reachable and t–rank$([\bar{A}, \bar{B}]) = n$. Since $\bar{B} \in \{0, \star\}^{n\times 1}, ||\bar{B}||_0 = 1$, we have t–rank($\bar{B}$) = 1. Suppose t–rank($\bar{A}$) $< n-1$, then,

$$\begin{aligned}\text{t–rank}([\bar{A},\bar{B}]) &\leq \text{t–rank}(\bar{A}) + \text{t–rank}(\bar{B}) = \text{t–rank}(\bar{A}) + 1\\ &< (n-1)+1 = n,\end{aligned}$$

which contradicts to the fact that t–rank($[\bar{A},\bar{B}]$) $= n$. Thus, t–rank($\bar{A}$) $\geq n-1$.

Since t–rank($\bar{A}$) $\geq n-1$, it follows that either $G_{\mathcal{X}\setminus\{x_1\}}$ or $G_{\mathcal{X}}$, subgraphs of $G(\bar{A})$, can be covered by vertex-disjoint cycles. Furthermore, since $[\bar{A}]_{jj} = 0, \forall j \in [n]$, there is no cycle of length 1 in $G_{\mathcal{X}}$. Hence, the first statement is true. Since $(\bar{A},\bar{B})$ is structurally controllable and $||\bar{B}||_0 = 1$, we have $\forall x_i \in \mathcal{X}$ is reachable, i.e., $G_{\mathcal{X}}$ is strongly connected.

($\Longleftarrow$) Suppose either $G_{\mathcal{X}\setminus\{x_1\}}$ or $G_{\mathcal{X}}$ can be covered by vertex-disjoint cycles and $G_{\mathcal{X}}$ is strongly connected, then we have t–rank($[\bar{A},\bar{B}]$) $= n$ and $\forall x_i \in \mathcal{X}$ is reachable. By Theorem 1, $(\bar{A},\bar{B})$ is structurally controllable. □

Proof of Theorem 11. The general idea in this proof is to characterize the optimal solution in the two cases: $|\mathcal{M}_1| = |\mathcal{M}_2|$ or $|\mathcal{M}_1| \neq |\mathcal{M}_2|$.

Case 1: Suppose $|\mathcal{M}_1| = |\mathcal{M}_2|$, then we approach the proof by first showing that computing $\mathcal{E}_{u_1}$ and $\mathcal{E}_{u_2}$ is polynomially solvable and then we show $|\mathcal{E}_{u_1}| < |\mathcal{E}_{u_2}|$ if $\mathcal{E}_{u_2}$ exists.

(i) Suppose $\mathcal{E}_{u_1}$ exists, then t–rank($\bar{A}$) $= n$ and there exist vertex-disjoint directed cycles $\{\mathcal{C}_i\}_{i=1}^{\ell}$ covering $\mathcal{X}$ in $G(\bar{A})$. We let $\mathcal{E}_{odd} = \{e \in \mathcal{E}_{\mathcal{C}_i} \colon |\mathcal{V}_{\mathcal{C}_i}| = 2q+1, q \in \mathbb{N}\}$ and $\mathcal{E}_{even} = \{e \in \mathcal{E}_{\mathcal{C}_i} \colon |\mathcal{V}_{\mathcal{C}_i}| = 2q, q \in \mathbb{N}\}$ be the union set of directed edges in cycles of odd-length and even-length in $\{\mathcal{C}_i\}_{i=1}^{\ell}$, respectively. The minimum number of undirected edges needed to cover an odd-length cycle $\mathcal{C}_{odd}$ is $|\mathcal{V}_{\mathcal{C}_{odd}}|$, i.e., the total number of vertices in this cycle $\mathcal{C}_{odd}$. Similarly, the minimum number of undirected edges needed to cover an even-length cycle $\mathcal{C}_{even}$ is $|\mathcal{V}_{\mathcal{C}_{even}}|$. However, there exist $|\mathcal{V}_{\mathcal{C}_{even}}|/2$ vertex-disjoint length-2 cycles covering vertices in $\mathcal{C}_{even}$, as shown in the proof of Lemma 2. Thus, the minimum number of undirected edges needed to make the vertices in $\mathcal{C}_{even}$ be covered by vertex-disjoint cycles is $|\mathcal{V}_{\mathcal{C}_{even}}|/2$, i.e., for each length-2 cycle we need one undirected edge. We denote by $\mathcal{E}'_u$ the minimum-cardinality set of undirected edges needed such that the digraph $G(\mathcal{X}, \{(x_i,x_j) \colon \{x_i,x_j\} \in \mathcal{E}'_u\})$can be covered by vertex-disjoint cycles. Let

$\mathcal{E}_u''$ be a set of undirected edges such that each directed cycle in $G(\mathcal{X}, \{(x_i, x_j) \colon \{x_i, x_j\} \in \mathcal{E}_u'\})$ is connected with some other directed cycles. We condense each cycle as a vertex, then finding undirected edges to make such vertices connected is equivalent to finding a minimum undirected spanning tree. Let n_{even}, n_{odd} be the total number of vertex-disjoint length-2 and odd-length cycles in $G(\mathcal{X}, \{(x_i, x_j) \colon \{x_i, x_j\} \in \mathcal{E}_u'\})$, respectively. Since there are $n_{even} + n_{odd}$ cycles, then we need $n_{even} + n_{odd} - 1$ undirected edges to connect the $n_{even} + n_{odd}$ cycles together, i.e., $|\mathcal{E}_u''| = n_{even} + n_{odd} - 1$. In total, we need undirected edges $|\mathcal{E}_{u_1}| = |\mathcal{E}_u'| + |\mathcal{E}_u''|$, i.e.,

$$\begin{aligned} |\mathcal{E}_{u_1}| &= |\mathcal{E}_u'| + |\mathcal{E}_u''| \\ &= (|\mathcal{E}_{even}|/2 + |\mathcal{E}_{odd}|) + (n_{even} + n_{odd} - 1), \end{aligned} \tag{A.18}$$

Since the total number of length-2 cycles n_{even} equals to $|\mathcal{E}_{even}|/2$, (A.18) yields

$$\begin{aligned} |\mathcal{E}_{u_1}| &= (|\mathcal{E}_{even}| + |\mathcal{E}_{odd}|) + (n_{odd} - 1) \\ &= |\mathcal{X}| + n_{odd} - 1. \end{aligned} \tag{A.19}$$

Therefore, minimizing $|\mathcal{E}_{u_1}|$ is equivalent to minimizing n_{odd}, the total number of odd-length disjoint cycles in $G(\bar{A})$. As such, we have recast this problem equivalently to maximizing the total number of length-2 disjoint cycles in $G(\bar{A})$, which is equivalent to finding a maximum undirected matching in $\mathcal{G}(\bar{A})$. The above approach to compute $\mathcal{E}_{u_1}$ can be used to find minimum undirected edges building vertex-disjoint directed cycles covering $\mathcal{X} \setminus \{x_1\}$ and make $G_{\mathcal{X}\setminus\{x_1\}}$ strongly connected. The only difference is that we should add an additional undirected edge connecting x_1 and a vertex in $\mathcal{X} \setminus \{x_1\}$, which makes $\forall x_i \in \mathcal{X}$ reachable.

(ii) We show $\mathcal{E}_{u_1}$ is an optimal solution by showing $|\mathcal{E}_{u_2}| > |\mathcal{E}_{u_1}|$ when $\mathcal{E}_{u_2}$ exists. Since $\mathcal{M}_1$ is a maximum undirected matching in $\mathcal{G}(\bar{A})$, there exists at most $|\mathcal{M}_1|$ length-2 vertex disjoint cycles in $G(\bar{A})$. Each state vertex $\forall x_i \in \mathcal{X}$ is either covered by a length-2 or odd-length vertex-disjoint cycle in $G(\bar{A}^\star)$. Through the derivation of (A.19), we notice that there are $|\mathcal{X}| - 2|\mathcal{M}_1|$ odd-length vertex-disjoint cycles in $G(\bar{A})$. Hence, $|\mathcal{E}_{u_1}| = |\mathcal{X}| + (|\mathcal{X}| - 2|\mathcal{M}_1|) - 1$.

Subsequently, we compute $\mathcal{E}_{u_2}$, the minimum number of undirected edges needed such that $\mathcal{X} \setminus \{x_1\}$ can be covered by vertex-disjoint cycles and $\mathcal{G}(\mathcal{X}, \mathcal{E}_{u_2})$ is strongly connected. There are at most $(|\mathcal{M}_1|-1)$ length-2 vertex-disjoint cycles in $G_{\mathcal{X}\setminus\{x_1\}}$ (Otherwise, $|\mathcal{M}_2| = |\mathcal{M}|_1+1$). Thus, we need at least $(|\mathcal{X}\setminus\{x_1\}|+(|\mathcal{X}\setminus\{x_1\}|-2(|\mathcal{M}_1|-1))-1)$ undirected edges such that $\mathcal{X}\setminus\{x_1\}$ is covered by disjoint directed cycles and they are strongly connected, and one edge connecting x_1 with some vertex in $\mathcal{X}\setminus\{x_1\}$ to ensure the reachability of all the $x_i \in \mathcal{X}$. In total, we have

$$\begin{aligned}|\mathcal{E}_{u_2}| &\geq (|\mathcal{X}\setminus\{x_1\}| + (|\mathcal{X}\setminus\{x_1\}| - 2(|\mathcal{M}_1|-1)) - 1) + 1 \\ &= 2|\mathcal{X}| - 2|\mathcal{M}_1| = |\mathcal{E}_{u_1}| + 1 > |\mathcal{E}_{u_1}|.\end{aligned}$$

Therefore, $\mathcal{E}_{u_1}$ is an optimal solution when $|\mathcal{M}_1| = |\mathcal{M}_2|$. **Case 2:** Suppose $|\mathcal{M}_1| \neq |\mathcal{M}_2|$, we first prove $|\mathcal{M}_2| = |\mathcal{M}_1| + 1$, then we show $\{u_1, x_1\} \in \mathcal{M}_2$, and finally we prove $|\mathcal{E}_{u_2}| < |\mathcal{E}_{u_1}|$ if $\mathcal{E}_{u_1}$ exists.

(i) We first prove $|\mathcal{M}_2| = |\mathcal{M}_1| + 1$. Suppose $|\mathcal{M}_2| \neq |\mathcal{M}_1|$, then it follows that $|\mathcal{M}_2| \geq |\mathcal{M}_1|+1$ because $\mathcal{G}(\bar{A})$ is a subgraph of $\mathcal{G}(\bar{A})_a$. We want to show that $|\mathcal{M}_2| \leq |\mathcal{M}_1|+1$. We prove it by contradiction. Suppose there exists a maximum undirected matching $\mathcal{M}_2$ in $\mathcal{G}(\bar{A}_a)$ such that $|\mathcal{M}_2| \geq |\mathcal{M}_1|+2$. There are only two cases, $\mathcal{M}_2$ includes either $\{u_1, x_1\}$, or $\{x_1, x_j\}$ for some $x_j \in \mathcal{X}$. In the first case, let $\mathcal{M}' = \mathcal{M}_2 \setminus \{\{u_1, x_1\}\}$. $\mathcal{M}'$ is also a maximum undirected matching in $\mathcal{G}(\bar{A})$, and we have

$$|\mathcal{M}'| = |\mathcal{M}_2| - 1 \geq |\mathcal{M}_1| + 1,$$

which contradicts that $\mathcal{M}_1$ is a maximum undirected matching in $\mathcal{G}(\bar{A})$. Additionally, in the second case, let $\mathcal{M}' = \mathcal{M}_2 \setminus \{\{u_1, x_1\}\}$. $\mathcal{M}'$ is also a maximum undirected matching in $\mathcal{G}(\bar{A})$, and we have

$$|\mathcal{M}'| = |\mathcal{M}_2| \geq |\mathcal{M}_1| + 2,$$

which also contradicts that $\mathcal{M}_1$ is a maximum undirected matching in $\mathcal{G}(\bar{A})$. Thus, $|\mathcal{M}_2| = |\mathcal{M}_1| + 1$ is true.

(ii) Next, we show $\{u_1, x_1\} \in \mathcal{M}_2$ if $|\mathcal{M}_2| \neq |\mathcal{M}_1|$. Suppose $\{u_1, x_1\} \notin \mathcal{M}_2$, then let $\mathcal{M}' = \mathcal{M}_2 \setminus \{\{u_1, x_1\}\}$. $\mathcal{M}'$ is a maximum undirected matching in $\mathcal{G}(\bar{A})$. It yields

$$|\mathcal{M}'| = |\mathcal{M}_2| = |\mathcal{M}_1| + 1,$$

which contradicts that $\mathcal{M}_1$ is a maximum undirected matching in $\mathcal{G}(\bar{A})$. Thus, we have $\{u_1, x_1\} \in \mathcal{M}_2$.

(iii) Finally, we show $|\mathcal{E}_{u_1}| > |\mathcal{E}_{u_2}|$ when $\mathcal{E}_{u_1}$ exists. Suppose $\mathcal{E}_{u_1}$ exits, then there exist $(|\mathcal{X}| - 2|\mathcal{M}_1|)$ odd length directed vertex-disjoint cycles in $G(\bar{A})$. By equation (A.19), $|\mathcal{E}_{u_1}| = |\mathcal{X}| + (|\mathcal{X}| - 2|\mathcal{M}_1|) - 1$. From (i) and (ii), we conclude that $\mathcal{M}_2 \setminus \{\{u_1, x_1\}\}$ is a maximum undirected matching in $\mathcal{G}(\mathcal{X} \setminus \{x_1\}, \{\{x_i, x_j\} \in \mathcal{E}_u(\bar{A}) \colon i, j \neq 1\})$. By a similar reasoning of (A.19), we need $|\mathcal{X} \setminus \{x_1\}| + (|\mathcal{X} \setminus \{x_1\}| - 2|\mathcal{M}_1|) - 1$ undirected edges such that $\mathcal{X} \setminus \{x_1\}$ can be covered by vertex-disjoint cycles and $G_{\mathcal{X} \setminus \{x_1\}}$ is strongly connected. Additionally, we need to add an edge connecting x_1 and a vertex in $\mathcal{X} \setminus \{x_1\}$ to let $\forall x_i \in \mathcal{X}$ be reachable. In total, $|\mathcal{E}_{u_2}| = 1 + |\mathcal{X} \setminus \{x_1\}| + (|\mathcal{X} \setminus \{x_1\}| - 2|\mathcal{M}_1|) - 1$, which implies $|\mathcal{E}_{u_2}| = |\mathcal{E}_{u_1}| - 1$. Thus, $\mathcal{E}_{u_2}$ is an optimal solution when $|\mathcal{M}_1| \neq |\mathcal{M}_2|$. □

A.3.4 Proof of Theorem 12

Proof. Suppose $|\mathcal{M}_1| = |\mathcal{M}_2|$, then the Step 2 returns a set $\mathcal{E}'_u$ of undirected edges such that $\mathcal{X}$ can be covered by $(|\mathcal{X}| - 2|\mathcal{M}_1|)$ odd-length and $|\mathcal{E}'_u \cap \mathcal{M}_1|$ length-2 vertex-disjoint directed cycles. Let $\hat{\mathcal{E}}'_u = \mathcal{E}'_u \cap \mathcal{M}_1$, and $\check{\mathcal{E}}'_u = \mathcal{E}'_u \setminus \hat{\mathcal{E}}'_u$. The $\hat{\mathcal{E}}'_u$ includes all the undirected edges constructing length-2 diected cycles and $\check{\mathcal{E}}'_u$ includes all the undirected edges constructng odd-length directed cycles. Subsequently, Step 4 returns a set $\mathcal{M}_t$ of undirected edges which connect $|\hat{\mathcal{E}}'_u|$ length-2 cycles and $(|\mathcal{X}| - 2|\mathcal{M}_1|)$ odd-length cycles and we have $|\mathcal{M}_t| = |\hat{\mathcal{E}}'_u| + (|\mathcal{X}| - 2|\mathcal{M}_1|) - 1$. In total,

$$\begin{aligned}
|\mathcal{E}_u| &= |\mathcal{E}'_u| + |\mathcal{M}_t| = |\mathcal{E}'_u| + (|\hat{\mathcal{E}}'_u| + (|\mathcal{X}| - 2|\mathcal{M}_1|) - 1) \\
&= (|\hat{\mathcal{E}}'_u| + |\check{\mathcal{E}}'_u|) + (|\hat{\mathcal{E}}'_u| + (|\mathcal{X}| - 2|\mathcal{M}_1|) - 1) \\
&= (2|\hat{\mathcal{E}}'_u| + |\check{\mathcal{E}}'_u|) + (|\mathcal{X}| - 2|\mathcal{M}_1| - 1) \\
&= |\mathcal{X}| + (|\mathcal{X}| - 2|\mathcal{M}_1| - 1) = 2|\mathcal{X}| - 2|\mathcal{M}_1| - 1,
\end{aligned} \tag{A.20}$$

where $2|\hat{\mathcal{E}}'_u| + |\check{\mathcal{E}}'_u| = |\mathcal{X}|$ because there are $2|\hat{\mathcal{E}}'_u|$ vertices covered by $|\hat{\mathcal{E}}'_u|$ length-2 directed cycles and $|\check{\mathcal{E}}'_u|$ vertices covered by odd-length directed cycles. The (A.20) shows that $\mathcal{E}_u$ is an optimal solution to Problem 5. Similarly, we can prove Algorithm 2 returns an optimal solution when $|\mathcal{M}_2| \neq |\mathcal{M}_1|$. Algorithm 2 involves computing a maximum undirected matching, a minimum cost perfect bipartite matching, and a minimum undirected spanning tree. The overall complexity is $\mathcal{O}(|\mathcal{X}|^3)$. □

A.3.5 Proof of Theorem 13 and related results

Proof of Lemma 5. We prove the first statement by contradiction. Suppose the total number of right-unmatched vertices become $r-q$ with respect a maximum matching $\tilde{\mathcal{M}}$ in $\mathcal{B}_2$, where $q \geq 3$, then we have $|\tilde{\mathcal{M}}| = n-(r-q)$. However, $\tilde{\mathcal{M}'} = \tilde{\mathcal{M}} \backslash \{(x_i, x_j), (x_j, x_i)\}$ is also a matching in $\mathcal{B}_1$ and $|\tilde{\mathcal{M}'}| = n-r+q-2$, which contradicts that $\mathcal{M}$ is a maximum matching in $\mathcal{B}_1$.

We now prove the second statement. Let $\{x_{i_\ell}\}_{\ell=1}^{k} \subseteq \mathcal{X}$ be the right-matched vertices with respect to matching $\mathcal{M}$ in $\mathcal{B}_1$. We approach the proof by constructing a matching $\mathcal{M}'$ such that $\{x_{i_\ell}\}_{\ell=1}^{k} \cup \{x_i, x_j\}$ are right-matched. Suppose x_i is right-unmatched, then x_i is either a starting vertex in a path $\mathcal{P}_1 = (x_i, x_{i_1} \cdots, x_{i_{(2p)}})$, in which there are $2p+1$ vertices for some $p \in \mathbb{N}$, or an isolated vertex, which can be considered as a starting vertex of a trivial path $\mathcal{P}_1$ which has only one vertex and no edge, in $G(\mathcal{X} \cup \mathcal{U}, \mathcal{M})$, otherwise it contradicts that x_i is right-unmatched with respect to $\mathcal{M}$. Similarly, we have x_j is either a starting vertex of a path $\mathcal{P}_2 = (x_j, x_{j_1} \cdots, x_{j_{(2q)}})$ or an isolated vertex in $G(\mathcal{X} \cup \mathcal{U}, \mathcal{M})$. Thus, we construct

$$\begin{aligned}\mathcal{M}' =& (\mathcal{M} \cup \{(x_i, x_j), (x_j, x_i)\} \cup \{(x_{i_{2\ell}}, x_{i_{2\ell-1}})\}_{\ell=1}^{p} \\ & \cup \{(x_{j_{2\ell}}, x_{j_{2\ell-1}})\}_{\ell=1}^{q}) \setminus (\{(x_{i_{2\ell}}, x_{i_{2\ell+1}})\}_{\ell=1}^{p-1} \\ & \cup \{(x_{j_{2\ell}}, x_{j_{2\ell+1}})\}_{\ell=1}^{q-1} \cup \{(x_i, x_{i_1}), (x_j, x_{j_1})\}).\end{aligned}$$

It is true that $|\mathcal{M}'| = |\mathcal{M}| + 2$, and x_i, x_j are right-matched with respect to $\mathcal{M}'$ in $\mathcal{B}_2$. □

Proof of Theorem 13. We first prove that $b = \text{ceil}(\ell + \max(\frac{r-q_1-2q_2}{2}, 0))$ is a lower bound for the optimal solution of Problem 6. Next, we show that the lower bound can be improved to $b' = \text{ceil}(\ell + \max(\frac{r-q_1-(2q_2-1)}{2}, 0))$ when there is no reachable right-unmatchable target vertex and $r - q_2 > 0$.

Since there are ℓ unreachable T-SCCs, the set $\mathcal{E}_u$ of undirected edges to be added satisfies $|\mathcal{E}_u| \geq \ell$. Moreover, we can reduce the total number of right-unmatched vertices to at least $\max(r - 2q_2 - q_1, 0)$ after we make all the unreachable Class-2, 1 and 0 T-SCCs reachable by adding $q_2+q_1+q_0 = \ell$ edges connecting each unreachable T-SCC with some reachable vertices. In addition, we need to add at least ceil(max($(r - 2q_2 - q_1)/2, 0)$) undirected edges to make the remained right-unmatched vertices matched. Thus,

$$
\begin{aligned}
|\mathcal{E}_u| &\geq q_0 + q_1 + q_2 + \text{ceil}(\max(\frac{r - q_1 - 2q_2}{2}, 0)) \\
&= \text{ceil}(l + \max(\frac{r - q_1 - 2q_2}{2}, 0)) = b.
\end{aligned}
\tag{A.21}
$$

The b is a lower bound for an optimal solution to Problem 6.

Next, we consider the case when there is no reachable right-unmatched target vertex and $r - q_2 > 0$. In this case, there must exist at least one Class-2 T-SCC which has more than one right-unmatched target vertex, otherwise it contradicts to $r - q_2 > 0$. We then show that the maximum number of right-unmatched target vertices can be reduced is $q_1 + 2q_2 - 1$. To fulfill the purpose of making q_1 Class-1 T-SCC reachable and minimizing right-unmatched vertices simultaneously, we should add q_1 undirected edges which connect each Class-1 T-SCC with a right-unmatched target vertex in Class-2 T-SCCs. Besides, we need to add q_2 undirected edges such that q_2 Class-2 T-SCCs are made reachable. Since there is no reachable right-unmatched target vertex, there must exist an edge in those newly added edges which connects a right-unmatched target vertex in a Class-2 unreachable T-SCC and a reachable right-matched vertex, which implies that among the all $q_1 + q_2$ newly added edges, at least $q_1 + 1$ edge cannot reduce the total number of right-unmatched target vertex by 2. Therefore, the maximum total number of right-unmatched target vertices can be reduced is $(q_1 + 1) + 2(q_2 - 1) = q_1 + 2q_2 - 1$, and b' is the lower bound for the minimum cost of Problem 6. □

A.3.6 Proof of Theorem 14

Proof. Let r, ℓ, q_1 and q_2 be defined in the statement of Theorem 13. Suppose we have q_2 Class-2 T-SCCs. By running Algorithm 6, there are only two cases: either all class-2 T-SCCs go through Steps (5) to (7), or there exists a j_0th, where $j_0 \geq 1$, Class-2 T-SCC which does not. In the first case, we have $r - 2q_2 \geq 0$ because when running Algorithm 6 each Class-2 T-SCC at its own iteration is made reachable by adding an edge which also eliminates 2 right-unmatched target vertices in Step (5), a result from Lemma 5; In the later case, we have two subcases: either the 1st Class-2 T-SCC goes through Steps (5) to (7), or it does not. In the first subcase, since we have sorted Class-2 T-SCCs in a decreasing order of the total number of right-unmatched target vertices in each T-SCC, we declaim that $|\mathcal{X}''_i| \leq |\mathcal{X}''_{j_0-1}| = 1, \forall i \geq j_0$, otherwise the $(j_0 - 1)$th Class-2 T-SCC has at least 2 right-unmatched target vertices, implying that the j_0th Class-2 T-SCC goes through Steps (5) to (7), which is a contradiction. After Algorithm 6 iterating over all the Class-2 T-SCCs, all the Class-2 T-SCCs will be made reachable and all right-unmatched target vertices will be matched because each $i \geq j_0$th Class-2 T-SCC has $|\mathcal{X}''_i| \leq 1$ right-unmatched target vertex and each newly added edges in the iteration of $i \geq j_0$th Class-2 T-SCC can make the right-unmatched target vertex in the ith T-SCC right-matched, as shown in the Step (10), Step (17) and Step (19). Hence, the right-unmatched target vertices will be reduced to $\max(r - 2q_2, 0) = 0$. In the later subcase, since the edge added at the 1st Class-2 T-SCC's iteration only eliminates one right-unmatched target vertex, after iterating over all the Class-2 T-SCCs, by a similar reasoning taken in the above analysis, we conclude that the remainder number of right-unmatched target vertices will be $\max(r - (1 + 2(q_2 - 1)), 0) = \max(r - (2q_2 - 1), 0)$.

From Step (29) to (38), since each newly added edge connects a vertex in the feature set of this Class-1 T-SCC with either a reachable right-unmatched vertex if exists, or a reachable vertex otherwise, all the Class-1 T-SCCs will be made reachable, and the total number of right-unmatched target vertices will be reduced to $\max(r - q_1 - (2q_2 - t), 0)$.

Then, from Step (39) to (42), all the Class-0 T-SCCs will be made reachable. So far, we have made all the $(q_2 + q_1 + q_0)$ unreachable T-SCCs reachable by adding $q_2 + q_1 + q_0$ undirected edges. From Step (43) to (50), we make all the $\max(r - q_1 - (2q_2 - t), 0)$ remainder right-unmatched target vertices matched by adding $\text{ceil}(\max(r - q_1 - (2q_2 - t), 0)/2)$ edges. We see that $\sum_{e \in \mathcal{E}_u(\bar{A}^\star)} c(e) = \text{ceil}(\ell + \max(r - q_1 - (2q_2 - t), 0)/2)$, i.e., Algorithm 6 returns an optimal solution. Since in Algorithm 6 the Step (15) computes a maximum matching and Algorithm 5 has complexity $\mathcal{O}(|\mathcal{X}|^3)$, we conclude Algorithm 6 has complexity $\mathcal{O}(|\mathcal{X}|^3)$. □

A.4 Proof of the results in Chapter 5

Proof of Lemma 6. First, we notice that sufficiency follows from the fact that structural controllability ensures that almost surely there exists numerical realization ensuring controllability, which implies that any desired state can be attained by a finite sequence of inputs. Therefore, if there was not one such sequence, then the uncontrollable subspace is nonempty, and the only way to ensure that we can take the state to the origin is when the subspace is stable. a control input driving the states to the origin in finite time. Necessity follows by contrapositive argument. Suppose that $(\bar{A}, \bar{B})$ is irreducible but not structurally controllable, then by Theorem 1 in Chapter 2, there exists a set $\mathcal{S} \subseteq \mathcal{X}$ such that $|\mathcal{N}(\mathcal{S})| < |\mathcal{S}|$, which implies that g-rank$([\bar{A}, \bar{B}]) < n$. For $\forall[\mathbf{p}_{\tilde{A}}, \mathbf{p}_{\tilde{B}}] \in \mathbb{R}^{n_{\bar{A}}+n_{\bar{B}}}$, $\exists v \in \mathbb{C}^n$, such that $v^T[\tilde{A}, \tilde{B}] = 0$, i.e., $v^T\tilde{A} = v^T 0$. Consequently, there exists a zero eigenvalue which is not controllable, hence not stabilizable. □

Proof of Lemma 7. (If) Let us construct a numerical realization $\tilde{A}_{22}$ by assigning zero value to off-diagonal $\star$-entries of $\bar{A}_{22}$, and negative values to $\star$-entries on the diagonal. In this case, matrix $\tilde{A}_{22}$ is negative definite diagonal matrix.

(Only if) We approach the proof by contrapositive. Let m be the dimension of $\bar{A}_{22}$, and $\{v_i\}_{i=1}^m$ be the standard basis in $\mathbb{R}^m$. Suppose there exists a fixed zero $[\bar{A}_{22}]_{ii} = 0$, then $v_i^T\tilde{A}_{22}v_i = [\tilde{A}_{22}]_{ii} = [\bar{A}_{22}]_{ii} = 0$, for all numerical realizations of $\bar{A}_{22}$; hence, $\tilde{A}_{22}$ is not negative definite. □

Proof of Theorem 15. (If) Without loss of generality, suppose $(\bar{A}, \bar{B})$ can be transformed to the form of (2.3). Suppose for $\forall\mathcal{S} \subseteq \mathcal{X}_r$, $|\mathcal{N}(\mathcal{S})| \geq |\mathcal{S}|$, then the input reachable subsystem $(\bar{A}_{11}, \bar{B}_1)$ is structurally controllable. If for $\forall x_i \in \mathcal{X}_u$, x_i has self-loop in $G(\bar{A}, \bar{B})$, then $[\bar{A}]_{ii}$ is a $\star$-entry. Let us assign negative numerical weights to all the $\star$-entries of $\bar{A}$ that correspond to the self-loop of all $x_i \in \mathcal{X}_u$. Then, the input-unreachable part of the system, $\tilde{A}_{22}$, is a negative definite diagonal matrix. Thus, we have shown that there exists a numerical realization $(\tilde{A}, \tilde{B})$, such that the uncontrollable part is asymptotically stable. Hence, the system is structurally stabilizable.

(Only if) The necessity can be proved by contrapositive. Suppose there exists a state vertex $x_i \in \mathcal{X}_u$ that $[\bar{A}]_{ii} = 0$, then, by Lemma 7 any numerical realization $(\tilde{A}, \tilde{B})$ has an uncontrollable non-negative eigenvalue. Furthermore, assume there exists $\mathcal{S} \subseteq \mathcal{X}$ such that $|\mathcal{N}(\mathcal{S})| < |\mathcal{S}|$, then by Lemma 6, $(\bar{A}, \bar{B})$ is not structurally stabilizable. □

Proof of Lemma 8. Suppose t–rank$([\bar{A}, \bar{B}]) = k$, then there exists a set $\mathcal{T} \in [n]$, such that for $\forall \mathcal{S} \subseteq \mathcal{X}_{\mathcal{T}} = \{x_i \in \mathcal{X} \colon i \in \mathcal{T}\}$, $|\mathcal{N}(\mathcal{S})| \geq |\mathcal{S}|$. By Theorem 2 in Chapter 2, $(\bar{A}, \bar{B})$ is structurally target controllable with respect to $\mathcal{T}$, which implies that there exists a numerical realization $(\tilde{A}, \tilde{B})$ with $[\mathbf{p}_{\tilde{A}}, \mathbf{p}_{\tilde{B}}] \in V^c \cap W^c$, where V and W are proper varieties in $\mathbb{R}^{n_{\bar{A}} + n_{\bar{B}}}$, such that the dimension of the controllable subspace is k, i.e., almost surely the dimension of controllable subspace of a numerical realization $(\tilde{A}, \tilde{B})$ is k. We have the generic dimension of controllable subspace of $(\bar{A}, \bar{B})$, $d_c =$ t–rank$([\bar{A}, \bar{B}])$. □

Proof of Theorem 16. Without loss of generality, there exists only two cases: either $(\bar{A}, \bar{B})$ is irreducible or not. In the first case, by Lemma 3 in Chapter 2, Lemma 6 and Lemma 8, the generic dimension of controllable subspace of $(\bar{A}, \bar{B})$ is t–rank$([\bar{A}, \bar{B}])$, and if t–rank$([\bar{A}, \bar{B}]) < n$, then for any numerical realization $(\tilde{A}, \tilde{B})$, there are $(n - k)$ zero uncontrollable eigenvalues. Therefore, the maximum dimension of stabilizable subspace of $(\bar{A}, \bar{B})$ is equal to t–rank$([\bar{A}, \bar{B}])$. In the other case, permute $(\bar{A}, \bar{B})$ to the form of (2.3) and let k be the number of $\star$-entries in the diagonal of $\bar{A}_{22}$. We can construct a numerical realization $\tilde{A}_{22}$ with m-dim$(\bar{A}_{22}, 0) = k$ by assigning negative values to the nonzero diagonals of $\bar{A}_{22}$ and zeros to other free entries of $\bar{A}_{22}$. By Theorem 15, we have that m-dim$(\bar{A}, \bar{B}) \geq$ t-rank$([\bar{A}_{11}, \bar{B}_1]) + k$. □

Proof of Theorem 17. We prove the NP-hardness of Problem 9 by (polynomially) reducing Min-k-Union problem to instances of Problem 9.

Suppose that we have a universe set $\mathcal{U}_{\mathcal{S}} = \{\mathcal{S}_\ell\}_{\ell=1}^{p}$, and an integer $k \in \mathbb{Z}^+$, for which we need to select k subsets in $\{\mathcal{S}_\ell\}_{\ell=1}^{p}$ such that $|\bigcup_{i=1}^{k} \mathcal{S}_{\ell_i}|$ is minimized. Let $n = |\mathcal{U}_{\mathcal{S}}|$ and define the state vertex set as $\mathcal{X} = \{x_i\}_{i=1}^{2n}$, and input vertex set as $\mathcal{U} = \{u_i\}_{i=1}^{p}$. Next, we can construct a set of directed edges between state vertices, $\mathcal{E}_{\mathcal{X},\mathcal{X}} = \{(x_i, x_{i+n}), (x_{i+n}, x_i)\}_{i=1}^{n}$, and a set of directed edges between input and state vertices,

$\mathcal{E}_{\mathcal{U},\mathcal{X}} = \{(u_i, x_j) \colon i \in [p], j \in \mathcal{S}_i\}$ – see Figure A-2 as an example for such a construction. In the constructed graph $G(\mathcal{X} \cup \mathcal{U}, \mathcal{E}_{\mathcal{X},\mathcal{X}} \cup \mathcal{E}_{\mathcal{U},\mathcal{X}})$, we have $|\mathcal{N}(\mathcal{S})| \geq |\mathcal{S}|, \forall \mathcal{S} \subseteq \mathcal{X}$, and all $x_i \in \mathcal{X}$ are reachable.

From the graph $G(\mathcal{X} \cup \mathcal{U}, \mathcal{E}_{\mathcal{X},\mathcal{X}} \cup \mathcal{E}_{\mathcal{U},\mathcal{X}})$, we construct the symmetrically structured matrix $\bar{A} \in \{0, \star\}^{2n \times 2n}$ such that $[\bar{A}]_{ij} = \star$ if $\{x_j, x_i\} \in \mathcal{E}_{\mathcal{X},\mathcal{X}}$ and $[\bar{A}]_{ij} = 0$ otherwise. We also construct $\bar{B} \in \{0, \star\}^{2n \times p}$ such that $[\bar{B}]_{ij} = \star$ if $\{u_j, x_i\} \in \mathcal{E}_{\mathcal{U},\mathcal{X}}$ and $[\bar{B}]_{ij} = 0$ otherwise. We can verify that the maximum stabilizable subspace for the constructed symmetrically structured matrix $\bar{A} \in \{0, \star\}^{2n \times 2n}$ is n and m-dim$(\bar{A}, \bar{B}(\mathcal{J})) = n + \dfrac{1}{2}|\cup_{j \in \mathcal{J}} \mathcal{S}_j|$. Let the attack budget be $c = p - k$. In our constructed instance of Problem 9, we aim to remove c actuators from $\{u_i\}_{i=1}^{p}$ such that the maximum dimension of the stabilizable subspace is minimized. Subsequently, we claim that an optimal solution of the constructed instance of Problem 9 enables us to retrieve an optimal solution to the Min-k-Union problem.

Suppose we have a feasible solution $\mathcal{U}_r = \{u_{\ell_i}\}_{i=1}^{p-k}$. Then if we consider $\mathcal{L} = [p] \setminus \{\ell_i\}_{i=1}^{p-k}$, we have that $\mathcal{L}$ is a feasible solution of Min-k-Union problem. Moreover, suppose $\mathcal{U}_r^* = \{u_{\ell_i}\}_{i=1}^{p-k}$ is a minimum solution to Problem 9, but $\mathcal{L} = [p] \setminus \{\ell_i\}_{i=1}^{p-k}$ is not an optimal solution of Min-k-Union problem, then $\mathcal{L}' = \{\eta_i\}_{i=1}^{k}$ would be a solution to Min-k-Union problem such that $|\bigcup_{i=1}^{k} \mathcal{S}_{\eta_i}| < |\bigcup_{i \in \mathcal{L}} \mathcal{S}_i|$. Next, let $\mathcal{U}_r' = \{u_i \in \mathcal{U} \colon i \in [p] \setminus \mathcal{L}'\}$ and notice that the maximum stabilizable subspace by removing $\mathcal{U}_r'$ is smaller than the maximum stabilizable subspace when removing $\mathcal{U}_r^*$, which contradicts $\mathcal{U}_r^*$ is an optimal solution.

□

Proof of Theorem 18. Consider an instance of Problem 9 under Assumption 3. We associate the structural pair $(\bar{A}, \bar{B})$, where $\bar{A} \in \{0, \star\}^{n \times n}$ and $\bar{B} \in \{0, \star\}^{n \times m}$, with a digraph $G(\bar{A}, \bar{B})$. Denote by $\{\mathcal{D}_i\}_{i=1}^{p}$ the set of vertices in the ith SCC in $\mathcal{D}(\bar{A}, 0)$.

Firstly, without loss of generality, we could assume that

$$\mathcal{D}_i = \{x_{d_{i-1}+1}, x_{d_{i-1}+2}, \cdots, x_{d_{i-1}+|\mathcal{D}_i|}\},$$

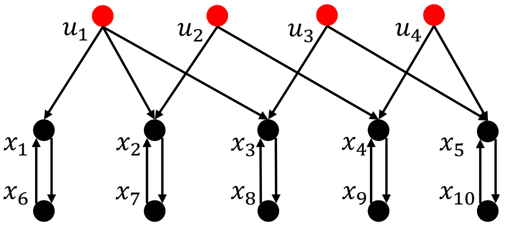


Figure A-2: Example of the construction of $G(\mathcal{X} \cup \mathcal{U}, \mathcal{E}_{\mathcal{X},\mathcal{X}} \cup \mathcal{E}_{\mathcal{U},\mathcal{X}})$ in the proof of Theorem 17. Suppose we have a finite universe set $\mathcal{U}_{\mathcal{S}} = \bigcup_{\ell=1}^{4} \mathcal{S}_\ell$, where $\mathcal{U}_{\mathcal{S}} = \{1,2,3,4,5\}, \mathcal{S}_1 = \{1,2,3\}, \mathcal{S}_2 = \{2,4\}, \mathcal{S}_3 = \{3,5\}, \mathcal{S}_4 = \{4,5\}$. From the given set $\mathcal{U}_{\mathcal{S}} = \bigcup_{\ell=1}^{4} \mathcal{S}_\ell$. We construct the state vertex set $\mathcal{X} = \{x_i\}_{i=1}^{10}$, and the input vertex set $\mathcal{U} = \{u_i\}_{i=1}^{4}$. The black and red vertices in Figure A-2 are the state and input vertices in $G(\mathcal{X} \cup \mathcal{U}, \mathcal{E}_{\mathcal{X},\mathcal{X}} \cup \mathcal{E}_{\mathcal{U},\mathcal{X}})$, respectively.

with $d_0 = 0$ and $d_i = d_{i-1} + |\mathcal{D}_i|$. Secondly, for the ith SCC, we define the set $\mathcal{R}_i = \{r_{i-1}+1, r_{i-1}+2, \ldots, r_{i-1} + (|\mathcal{D}_i| - \text{m-dim}(\bar{A}_{\mathcal{D}_i}, 0))\}$, with $r_0 = 0$ and $r_i = r_{i-1} + |\mathcal{D}_i| - \text{m-dim}(\bar{A}_{\mathcal{D}_i}, 0)$, where $\bar{A}_{\mathcal{D}_i} \in \{0, \star\}^{|\mathcal{D}_i| \times |\mathcal{D}_i|}$ is the symmetrically structured matrix corresponding to the ith SCC. Finally, for the jth control input, we construct the set $\mathcal{S}_j = \{\bigcup_{i \in \mathcal{I}} \mathcal{R}_i \mid \mathcal{I}$ is the set of SCCs reachable from $u_j\}$. By the above construction and Assumption 3, we have that

$$\text{m-dim}(\bar{A}, \bar{B}(\mathcal{J})) = \text{m-dim}(\bar{A}, 0) + |\bigcup_{j \in \mathcal{J}} \mathcal{S}_j|. \tag{A.22}$$

We let $\mathcal{U}_{\mathcal{S}} = \bigcup_{i=1}^{m} \mathcal{S}_i$. Suppose the budget in Problem 9 is k, then Problem 9 is to find $(m-k)$ sets $\mathcal{S}_{\ell_1}, \cdots, \mathcal{S}_{\ell_{m-k}}$ among $\{\mathcal{S}_i\}_{i=1}^{m}$ such that $|\bigcup_{i=1}^{m-k} \mathcal{S}_{\ell_i}|$, i.e., the number of reachable state vertices, is minimized. By Definition 13, we see that in this case Problem 9 is equivalent to the Min-k-Union problem, in which we are given sets $\{\mathcal{S}_i\}_{i=1}^{m}$ and we aim to find $(m-k)$ sets $\{\mathcal{S}_{\ell_i}\}_{i=1}^{m-k}$, $\{\ell_i\}_{i=1}^{m-k} \subseteq \{1, 2, \cdots, (m-k)\}$, such that $|\bigcup_{i=1}^{m-k} \mathcal{S}_{\ell_i}|$ is minimized. If there exists a $\rho(m)$-approximation algorithm for the Min-k-Union problem, i.e., $|\bigcup_{j \in \mathcal{J}} \mathcal{S}_j| \leq \rho(m) |\bigcup_{j \in \mathcal{J}^*} \mathcal{S}_j|$, then,

$$\begin{aligned} \text{m-dim}(\bar{A}, \bar{B}(\mathcal{J})) &= \text{m-dim}(\bar{A}, 0) + |\bigcup_{j \in \mathcal{J}} \mathcal{S}_j| \\ &\leq \text{m-dim}(\bar{A}, 0) + \rho(m) \cdot (|\bigcup_{j \in \mathcal{J}^*} \mathcal{S}_j|) \\ &\leq \rho(m) \cdot \text{m-dim}(\bar{A}, \bar{B}(\mathcal{J}^*)), \end{aligned} \tag{A.23}$$

where $\mathcal{J}^*$ is an optimal solution to the Min-k-Union problem. From the above reasoning, we have that $\bar{B}(\mathcal{J}^*)$ is also an optimal solution to Problem 10 and $\text{m-dim}(\bar{A}, \bar{B}(\mathcal{J})) \leq \rho(m) \cdot \text{m-dim}(\bar{A}, \bar{B}(\mathcal{J}^*))$. □

Sketch of Proof of Theorem 19. We can prove the NP-hardness by reducing a general instance of the Max-k-Union problem to an instance of Problem 10. Suppose we have a ground set $\mathcal{U}_{\mathcal{S}} = \{\mathcal{S}_\ell\}_{\ell=1}^p$, and an integer $k \in \mathbb{N}$. The constrained maximum set coverage problem is to select k subsets in $\mathcal{U}_{\mathcal{S}}$ such that $|\bigcup_{i=1}^k \mathcal{S}_{\ell_i}|$ is maximized. Following a similar construction and reasoning taken in the proof of Theorem 17, we can prove that the Max-k-Union problem can be reduced to Problem 10 in polynomial time. □

Proof of Theorem 20. Consider a structural pair $(\bar{A}, \bar{B})$, where $\bar{A} \in \{0, \star\}^{n\times n}$ is symmetrically structured and $\bar{B} \in \{0, \star\}^{n\times m}$ is structured. We let $\mathcal{U}$ denote the input vertices corresponding columns of $\bar{B}$, and let $\mathcal{U}_{can}$, where $|\mathcal{U}_{can}| = m'$, be the set of new actuators that can be added to the system. We associate with the set $\mathcal{U}_{can}$ the structured matrix $\bar{B}_{\mathcal{U}_{can}} \in \{0, \star\}^{n\times m'}$. Define a function $f \colon \mathcal{J} \subseteq [m'] \to \text{m–dim}(\bar{A}, [\bar{B}, \bar{B}_{\mathcal{U}_{can}}(\mathcal{J})])$. We first prove that $f(\mathcal{J})$ is a submodular function, and then we show that Algorithm 7 returns a $(1 - 1/e)$ approximation solution.

Before we proceed, we construct a few sets. We denote by $\{\mathcal{D}_i\}_{i=1}^p$ the set of vertices in the ith input-unreachable SCC in $G(\bar{A}, \bar{B})$. Without loss of generality, we assume that $\mathcal{D}_i = \{x_{d_{i-1}+1}, x_{d_{i-1}+2}, \cdots, x_{d_{i-1}+|\mathcal{D}_i|}\}$, with $d_0 = 0$ and $d_i = d_{i-1} + |\mathcal{D}_i|$. Then, for the i-th unreachable SCC, we define the set $\mathcal{R}_i = \{r_{i-1}+1, r_{i-1}+2, \ldots, r_{i-1} + (\text{t-rank}(\bar{A}_{\mathcal{D}_i}) - \text{m-dim}(\bar{A}_{\mathcal{D}_i}, 0))\}$, with $r_0 = 0$ and $r_i = r_{i-1} + \text{t-rank}(\bar{A}_{\mathcal{D}_i}) - \text{m-dim}(\bar{A}_{\mathcal{D}_i}, 0)$, where $\bar{A}_{\mathcal{D}_i} \in \{0, \star\}^{|\mathcal{D}_i|\times|\ \mathcal{D}_i|}$ is the symmetrically structured matrix corresponding to the i-th unreachable SCC. Finally, for the j-th control input u_j in $\mathcal{U}_{can}$, we construct the set $\mathcal{S}_j = \{\bigcup_{i\in\mathcal{I}} \mathcal{R}_i \mid \mathcal{I}$ is the set of unreachable SCCs in $\mathcal{D}(\bar{A}, \bar{B})$ but reachable from $u_j\}$.

Let $q(\mathcal{J})$ be the total number of state vertices which are right-unmatched in $\mathcal{B}(\bar{A}, \bar{B})$

but right-matched in $\mathcal{B}(\bar{A}, [\bar{B}, \bar{B}_{\mathcal{U}_{can}}(\mathcal{J})])$. By definition, we have

$$f(\mathcal{J}) = \text{m-dim}(\bar{A}, \bar{B}) + |\bigcup_{j\in\mathcal{J}} \mathcal{S}_j| + q(\mathcal{J}). \tag{A.24}$$

which implies $f(\mathcal{J})$ is a monotonically increasing function of $\mathcal{J} \subseteq [m']$.

Furthermore, consider two sets $\mathcal{J}_1$, $\mathcal{J}_2$, where $\mathcal{J}_1 \subseteq \mathcal{J}_2 \subseteq [m']$. Suppose $j \in [m'] \setminus \mathcal{J}_2$ and denote by $\mathcal{J}_1' = \mathcal{J}_1 \cup \{j\}$ and $\mathcal{J}_2' = \mathcal{J}_2 \cup \{j\}$, then

$$f(\mathcal{J}_1') - f(\mathcal{J}_1) = |\bigcup_{j\in\mathcal{J}_1'} \mathcal{S}_j| - |\bigcup_{j\in\mathcal{J}_1} \mathcal{S}_j| + q(\mathcal{J}_1') - q(\mathcal{J}_1), \tag{A.25}$$

and

$$f(\mathcal{J}_2') - f(\mathcal{J}_2) = |\bigcup_{j\in\mathcal{J}_2'} \mathcal{S}_j| - |\bigcup_{j\in\mathcal{J}_2} \mathcal{S}_j| + q(\mathcal{J}_2') - q(\mathcal{J}_2). \tag{A.26}$$

On one hand, suppose $q(\mathcal{J}_1') - q(\mathcal{J}_1) = 1$, then $q(\mathcal{J}_2') - q(\mathcal{J}_2) = 1$ or 0; On the other hand, suppose $q(\mathcal{J}_1') - q(\mathcal{J}_1) = 0$, then $q(\mathcal{J}_2') - q(\mathcal{J}_2) = 0$. Recall that the set coverage function is a submodular function, i.e., $|\bigcup_{j\in\mathcal{J}_1'} \mathcal{S}_j| - |\bigcup_{j\in\mathcal{J}_1} \mathcal{S}_j| \geq |\bigcup_{j\in\mathcal{J}_2'} \mathcal{S}_j| - |\bigcup_{j\in\mathcal{J}_2} \mathcal{S}_j|$. Therefore, we have that

$$f(\mathcal{J}_1') - f(\mathcal{J}_1) \geq f(\mathcal{J}_2') - f(\mathcal{J}_2), \tag{A.27}$$

which implies that $f(\mathcal{J})$ is a monotonically increasing submodular function.

Because $f(\mathcal{J})$ is a monotonically increasing submodular function, by a similar technique taken in the proof of [213, Proposition 5.1], we can show that Algorithm 7 returns a $(1 - 1/e)$-approximation solution to Problem 10. □

A.5 Proof of the results in Chapter 6

Proof of Theorem 21. Given $k \in \mathbb{N}$, we have

$$\begin{aligned} \operatorname{Tr}(A^k) &= \sum_{i=1}^{n} [A^k]_{ii}, \\ &= \sum_{i=1}^{n} \sum_{j_1,\dots,j_{k-1}} [A]_{ij_1} \cdots [A]_{j_{k-1}i}. \end{aligned} \tag{A.28}$$

In particular, since $[A]_{ii} = 0$ for all $i \in [n]$, we must have $i \neq j_1, j_{k-1} \neq i$ and $j_\ell \neq j_{\ell+1}$ for all $\ell < k-1$ in the above summation, since the term $[A]_{ij_1} \cdots [A]_{j_{k-1}i}$ vanished otherwise. We use $i \to j_1 \cdots j_{k-1} \to i$ to represent a closed walk of length k satisfying $[A]_{ij_1} \cdots [A]_{j_{k-1}i} \neq 0$. Notice that there may exist repetitive indices in $i \to j_1 \cdots j_{k-1} \to i$; hence, we may have that $|\{i, j_1, \dots, j_{k-1}, i\}| \leq k$. Subsequently, we have:

$$\sum_{j_1,\dots,j_{k-1}} [A]_{ij_1} \cdots [A]_{j_{k-1}i} = \sum_{s=2}^{k} \sum_{|\{i,j_1,\dots,j_{k-1},i\}|=s} [A]_{ij_1} \cdots [A]_{j_{k-1}i}. \tag{A.29}$$

In other words, we can classify closed walks into subgraphs with orders less or equal to k. In particular, these subgraphs are weakly-connected. Combining (A.29) and (A.28), we have

$$\operatorname{Tr}(A^k) = \sum_{i=1}^{n} \sum_{s=2}^{k} \sum_{|\{i,j_1,\dots,j_{k-1},i\}|=s} [A]_{ij_1} \cdots [A]_{j_{k-1}i}. \tag{A.30}$$

Below, we analyze how the counts of order-k, weakly-connected subgraphs contribute to (A.29).

Let us consider a subgraph $G_{sub} \subseteq G$ with order $s \leq k$. Without loss of generality, we may relabel the vertices of G_{sub} by $[s]$. Consider a closed walk of length k in G_{sub} such that the closed walk traverses each edge of G_{sub} at least once. Let $\eta_{i,k}(G_{sub})$ be the number of these closed walks starting at $i \in [n]$. Then, each subgraph G_{sub} contributes $\sum_{i=1}^{s} \eta_{i,k}(G_{sub})$ number of walks in the summation in (A.30). Moreover, the number $\eta_{i,k}(G_{sub})$ is the same for all $G_h \in \mathtt{Iso}(G_{sub})$. Let $\eta_k(G_{sub}) = \sum_{i=1}^{s} \eta_{i,k}(G_{sub})$. Then, each class of subgraph contributes $\texttt{Count}(G_{sub}, G)\eta_k(G_{sub})$ to $\mathrm{Tr}(A^k)$. As a result,

$$\mathrm{Tr}(A^k) = \sum_{s=2}^{k} \sum_{G_{sub}\in\Omega_s} \texttt{Count}(G_{sub}, G)\eta_k(G_{sub}).$$

In particular, let A_s be the adjacency matrix of G_{sub}, if $\mathrm{Tr}[A_s^k] = 0$, then

$$\eta_{i,k}(G_{sub}) = 0,$$

for all $i \in [s]$. □

Proof of Theorem 23. First, consider the spectral distribution μ_A and generate from μ_A an infinite multi-sequence $\mathbf{y}_{2,\infty}$ whose elements are given by $y_{\boldsymbol{\alpha}} = \mathbb{E}_{\mu_A}[\mathbf{x}^{\boldsymbol{\alpha}}]$ for all $\boldsymbol{\alpha} \in \mathbb{N}^2$. The discussions before Theorem 23 show that, given a fixed $r \in \mathbb{N}$, there exits a finite subsequence in $\mathbf{y}_{2,\infty}$ satisfying (6.15)–(6.17). Furthermore, according to Corollary 3, this subsequence satisfies $\tilde{M}_r \succeq 0, \tilde{L}_r(g_1) \succeq 0, \tilde{L}_r(g_2) \succeq 0, \tilde{L}_r(g_3) \succeq 0, \tilde{L}_r(g_4) \succeq 0$. In other words, all the constraints in (6.24) are satisfied. Thus, we can induce from $\mathbf{y}_{2,\infty}$ a *finite* subsequence of moments that is feasible with respect to (6.24). Consequently, the minimization in Theorem 23 leads to a lower bound on λ_n.

Similarly, for $r > 1$, we let $\mathcal{F}_r$ be the set of feasible solutions to (6.24). Since $\tilde{M}_r \succeq 0$, it follows that all its principal submatrices are positive semidefinite. Thus, $\tilde{M}_{r-1} \succeq 0$. Similar statements hold for $\tilde{L}_r(g_1), \tilde{L}_r(g_2), \tilde{L}_r(g_3)$, and $\tilde{L}_r(g_4)$. Thus, we have that $\mathcal{F}_r \subseteq \mathcal{F}_{r-1}$ and, consequently, $\rho^\star_{l,2r+1} \geq \rho^\star_{l,2r-1}$. □

Proof of Theorem 24. It suffices to replace μ_A in the proof of Theorem 23 by $\tilde{\mu}_A$. The rest of the proof of this theorem follows exactly the same logic as the proof of Theorem 23. □

Proof of Theorem 25 and Theorem 26. By replacing μ_A in the proof of Theorem 23 by $\tilde{\nu}_{A_I}$, and $\tilde{\nu}_{A_R}$, respectively, we can obtain that $p^\star_r \geq \lambda_n(A_R)$ and $\omega^\star_r \geq \lambda_n(A_I)$ for every

$r \in \mathbb{N}$. Combining these bounds with

$$\omega_{\max}(A) \leq \lambda_n \left(\frac{j(A - A^\top)}{2} \right),$$

and

$$\lambda_n(A) \leq \lambda_n \left(\frac{A + A^\top}{2} \right),$$

the result follows. □

A.6 Proof of the results in Chapter 7

In this appendix, we provide proofs for the lemmas and theorems in this paper. Throughout this section, denote the m-by-n matrix of all ones by $J_{m,n} \in \mathbb{R}^{m \times n}$.

Lemma 19. *Given a directed graph $G = (\mathcal{V}, \mathcal{E})$, $i_1, \ldots, i_k \in [n]$, and $\boldsymbol{\alpha} \in \mathbb{N}^k$. If*

$$dx_i = -x_i dP_{\delta_i} + (1 - x_i) \sum_{j \in \mathcal{N}_i^-} x_j dP_{\beta_{ij}}, \tag{A.31}$$

for all $i \in \mathcal{V}$, then

$$\frac{d\mathbb{E}[\phi_{\boldsymbol{\alpha}}(\mathbf{x})]}{dt} = \frac{d\mathbb{E}[\phi_{\mathbf{1}}(\mathbf{x})]}{dt}. \tag{A.32}$$

Proof of Lemma 19. To show (A.32), we first write (A.31) in the form of (7.8). Notice that there are $|\mathcal{V}| + |\mathcal{E}|$ Poisson counters in total, thus we define $h_\ell : \mathbb{R}^n \to \mathbb{R}^n$, for $\ell \in [|\mathcal{V}| + |\mathcal{E}|]$. Each h_ℓ is defined as follows: (i) when $\ell \in [n]$, we let $h_\ell(\mathbf{x}) = [0, \ldots, -x_\ell, \ldots, 0]^\top$, and (ii) when $\ell > n$, we order the edges $(j, i) \in \mathcal{E}$ and assign then with a label ℓ; hence, $h_\ell(\mathbf{x}) = [0, \ldots, (1 - x_i)x_j, \ldots, 0]^\top$, i.e., each $(j, i) \in \mathcal{E}$ is associated with a function h_ℓ.

With these definitions, it follows from (7.9) that

$$\begin{aligned} d\phi_{\mathbf{1}}(\mathbf{x}) = & - \Pi_{s=1}^{k} x_{i_s} \left(\sum_{s=1}^{k} dP_{\delta_{i_s}} \right) \\ & + \sum_{s=1}^{k} \sum_{\ell \in \mathcal{N}_s^-} x_{i_1} \cdots (1 - x_{i_s}) x_\ell \cdots x_{i_k} dP_{\beta_{s\ell}}. \end{aligned} \tag{A.33}$$

Notice that the random variables x_i's are supported on $[0,1]$, for all $i \in [n]$, therefore $\mathbb{E}[\phi_{\boldsymbol{\alpha}}(\mathbf{x})]$ exists and is finite for every $\boldsymbol{\alpha} \in \mathbb{N}^n$. Subsequently, from (A.33), we have that:

$$\begin{aligned} \frac{d\mathbb{E}[\phi_{\mathbf{1}}(\mathbf{x})]}{dt} = & - \sum_{s=1}^{k} \delta_{i_\ell} \mathbb{E}\left[\phi_{\mathbf{1}}(\mathbf{x})\right] \\ & + \sum_{s=1}^{k} \sum_{\ell \in \mathcal{N}_{i_s}^-} \beta_{i_s \ell} \mathbb{E}[x_{i_1} \cdots (1 - x_{i_s}) x_\ell \cdots x_{i_k}]. \end{aligned} \tag{A.34}$$

On the other hand,

$$\begin{aligned} d\phi_{\boldsymbol{\alpha}}(\mathbf{x}) = & -\phi_{\boldsymbol{\alpha}}(\mathbf{x}) \left(\sum_{s=1}^{k} dP_{\delta_{i_s}} \right) \\ + & \sum_{s=1}^{k} \sum_{\ell \in \mathcal{N}_s^-} x_{i_1}^{\alpha_1} \cdots \left((x_{i_s} + (1 - x_{i_s})\, x_\ell)^{\alpha_s} - x_{i_s}^{\alpha_s} \right) \cdots x_{i_k}^{\alpha_k} dP_{\beta_{s\ell}}. \end{aligned} \tag{A.35}$$

Since $x_i \in \{0,1\}$ for all $i \in [n]$, the term $(x_{i_s} + (1 - x_{i_s})\, x_\ell)^{\alpha_s} - x_{i_s}^{\alpha_s}$ equals to

$$\sum_{\kappa=0}^{\alpha_s - 1} \binom{\alpha_s}{\kappa} x_{i_s}^{\kappa} \left((1 - x_{i_s}) x_\ell\right)^{\alpha_s - \kappa}.$$

However, since $x_{i_s} \in \{0,1\}$, the above term can be further simplified to $((1 - x_{i_s}) x_\ell)^{\alpha_s}$. Therefore, taking expectation of (A.35) leads to:

$$\begin{aligned} \frac{d\mathbb{E}[\phi_{\boldsymbol{\alpha}}(\mathbf{x})]}{dt} = & - \sum_{\ell=1}^{k} \delta_{i_\ell} \mathbb{E}\left[\phi_{\boldsymbol{\alpha}}(\mathbf{x})\right] \\ & + \sum_{s=1}^{k} \sum_{\ell \in \mathcal{N}_{i_s}^-} \beta_{i_s \ell} \mathbb{E}[x_{i_1}^{\alpha_1} \cdots (1 - x_{i_s})^{\alpha_s} x_\ell^{\alpha_s} \cdots x_{i_k}^{\alpha_k}] \\ = & \frac{d\mathbb{E}[\phi_{\mathbf{1}}(\mathbf{x})]}{dt}, \end{aligned} \tag{A.36}$$

where the second equality is due to x_i are binary random variables. □

With the above lemma, we proceed to prove Theorem 29.

Proof of Theorem 29. From Lemma 19, we have that

$$\begin{aligned}\frac{d\mathbb{E}[\phi_{\boldsymbol{\alpha}}(\mathbf{x})]}{dt} &= -\sum_{\ell=1}^{k} \delta_{i_\ell} \mathbb{E}\left[\phi_{\boldsymbol{\alpha}}(\mathbf{x})\right] \\ &+ \sum_{s=1}^{k} \sum_{\ell \in \mathcal{N}_{i_s}^{-}} \beta_{i_s \ell} \mathbb{E}[x_{i_1}^{\alpha_1} \cdots (1 - x_{i_s})^{\alpha_s} x_{\ell}^{\alpha_s} \cdots x_{i_k}^{\alpha_k}].\end{aligned} \tag{A.37}$$

Meanwhile, $\frac{d\mathbb{E}[\phi_{\boldsymbol{\alpha}}(\mathbf{x})]}{dt} = \frac{d\mathbb{E}[\phi_{\mathbf{1}}(\mathbf{x})]}{dt}$ holds for all $\boldsymbol{\alpha}$, thus rearranging the term $\mathbb{E}[x_{i_1} \cdots (1 - x_{i_s}) x_\ell \cdots x_{i_k}]$ leads us to (7.11). □

In order to show Lemma 15, we introduce the following lemma.

Lemma 20. *Consider a matrix $A = [a_{ij}] \in \mathbb{R}^{n \times n}$, a sequence of integers $d_1, \ldots, d_n \in \mathbb{N}$, and a mapping $f : \mathbb{R}^{n \times n} \to \mathbb{R}^{\sum_{i=1}^n d_i \times \sum_{i=1}^n d_i}$ defined as*

$$f(A) = \begin{bmatrix} a_{11} J_{d_1,d_1} & a_{11} J_{d_1,d_2} & \cdots & a_{1n} J_{d_1,d_n} \\ * & a_{22} J_{d_2,d_2} & \cdots & a_{2n} J_{d_2,d_n} \\ \vdots & \vdots & \ddots & \vdots \\ * & * & * & a_{nn} J_{d_n,d_n} \end{bmatrix},$$

where J_{pq} is the $p \times q$ matrix of all ones. Then, if $A \succeq 0$, we have that $f(A) \succeq 0$. ◇

Proof of Lemma 20. To proof the Lemma, let us define $T_{d_1,\ldots,d_n} \in \mathbb{R}^{\sum_{i=1}^n d_i \times n}$ as

$$T_{d_1,\ldots,d_n} = \begin{bmatrix} \mathbf{1}_{d_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ * & \cdots & \mathbf{1}_{d_n} \end{bmatrix},$$

i.e., a block-diagonal matrix with its diagonal blocks specified by $\mathbf{1}_{d_1}, \cdots, \mathbf{1}_{d_n}$.

Next, we notices that given a matrix $A = [a_{ij}] \in \mathbb{R}^{n\times n}$,

$$\begin{aligned} f(A) &= \begin{bmatrix} a_{11}J_{d_1,d_1} & a_{11}J_{d_1,d_2} & \cdots & a_{1n}J_{d_1,d_n} \\ * & a_{22}J_{d_2,d_2} & \cdots & a_{2n}J_{d_2,d_n} \\ \vdots & \vdots & \ddots & \vdots \\ * & * & * & a_{nn}J_{d_n,d_n} \end{bmatrix} \\ &= T_{d_1,\ldots,d_n} A T_{d_1,\ldots,d_n}^\top . \end{aligned}$$

Consequently, $A \succeq 0$ implies that $f(A) = T_{d_1,\ldots,d_n} A T_{d_1,\ldots,d_n}^\top \succeq 0$.

Suppose that there exists $\mathbf{v}$ such that $\mathbf{v}^\top A \mathbf{v} < 0$, then we construct $\mathbf{w} \in \mathbb{R}^{\sum_{i=1}^n d_i}$ as follows iteratively. The first d_1 entries of $\mathbf{w}$ are all equals to $\mathbf{v}_1/d_1$, the i-th d_i entries of $\mathbf{w}$ are all equals to $\mathbf{v}_i/d_i$. Thus, $T_{d_1,\ldots,d_n}^\top \mathbf{w} = \mathbf{v}$. Thus, $f(A)$ is negative definite. Consequently, we have shown that $A \succeq 0$ if and only if $f(A) \succeq 0$. □

With the help of Lemma 20, we are able to prove Lemma 15.

Proof of Lemma 15. Notice that $\tilde{S}_k = [0,1]^k$ is a compact, semi-algebraic set, and it satisfies the Putinar's condition, it follows that $\mathbf{y}_\infty(\mathcal{I}_k)$ is $\tilde{S}_k$-feasible if and only if the conditions in (7.6) are satisfied. Subsequently, it suffices to show that the matrices in (7.18) implies the positive semi-definiteness of the moment and localizing matrices specified according to Theorem 1 and vice versa.

Consider $r \in \mathbb{N}$ and $r \geq \bar{k}$, the construction of (7.13) together with the definition of $\mathbf{y}_\infty(\mathcal{I}_k)$ implies that $M_{\bar{k}}(\mathbf{y}(\mathcal{I}_k))$ is the $\bar{k}$-th order principal submatrix of $M_r(\mathbf{y}_\infty(\mathcal{I}_k))$. Since $M_r(\mathbf{y}_\infty(\mathcal{I}_k)) \succeq 0$, we have that $M_{\bar{k}}(\mathbf{y}(\mathcal{I}_k)) \succeq 0$ as well. Similarly, the positive semi-definiteness of localizing matrices of $\mathbf{y}_\infty(\mathcal{I}_k)$ implies that both $L^1_{\bar{k}}(\mathbf{y}(\mathcal{I}_k), s)$ and $L^0_{\bar{k}}(\mathbf{y}(\mathcal{I}_k), s)$ are positive semi-definite for all $s \in [k]$.

Conversely, if $M_{\bar{k}}(\mathbf{y}(\mathcal{I}_k))$ is positive semi-definite, we aim to show that $M_r(\mathbf{y}_\infty(\mathcal{I}_k))$ is also positive semi-definite for all $r \in \mathbb{N}$. Notice that it suffices to show the above relationship holds for $r > k$. To achieve this goal, we proceed by permuting the entries in $M_r(\mathbf{y}_\infty(\mathcal{I}_k))$. Without loss of generality, we assume that $\mathcal{I}_k = \{1,\ldots,k\}$. Given a a set $\mathcal{W} \subseteq \mathcal{I}_k$, we use $x_{\mathcal{W}} = \Pi_{i \in \mathcal{W}} x_i$. Next, we consider

$$\mathbf{v}_r(\mathbf{x}_{\mathcal{I}_k}) = \left[1, x_1, \ldots, x_k, x_1^2, \ldots x_k^2, \ldots, x_1^r, \ldots, x_k^r\right],$$

and

$$\mathbf{v}'_r(\mathbf{x}_{\mathcal{I}_k}) = \Bigg[1, \quad \overbrace{x_1, \ldots, x_1^r,}^{\text{monomials involving only } x_1} \quad \ldots,$$
$$\underbrace{x_{\mathcal{W}}, \ldots}_{\text{monomials involving only } x_w \text{ for } w \in \mathcal{W}} \quad , \ldots, x_{\mathcal{I}_k}\Bigg].$$

Thus, $\mathbf{v}'_r(\mathbf{x}_{\mathcal{I}_k})$ is a permutation on the entries in $\mathbf{v}_r(\mathbf{x}_{\mathcal{I}_k})$. Let $N_r = \binom{k+r}{r}$, and S_r be the permutation group on the set $[N_r]$, there exists $\pi : S_r \to S_r$ such that for all $i \in [N_r]$ we have $\mathbf{v}'_r(\mathbf{x}_{\mathcal{I}_k})_i = \pi(\mathbf{v}_r(\mathbf{x}_{\mathcal{I}_k})_j$ for some $j \in [N_r]$.

Consider the following N_r-dimensional matrix $\hat{M}_r$ whose entries are defined by:

$$[\hat{M}_r]_{\boldsymbol{\alpha},\boldsymbol{\beta}} = y_\infty(\mathcal{I}_k)_{\pi^{-1}(\boldsymbol{\alpha}+\boldsymbol{\beta})} \tag{A.38}$$

for all $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{N}^k_{r/2}$. Thus, there exists a permutation matrix $P \in \{0,1\}^{N_r \times N_r}$ such that $\hat{M}_r = PM_r(\mathbf{y}_\infty(\mathcal{I}_k))P^{-1}$. Moreover, $\hat{M}_r$ is in the following form:

$$\begin{bmatrix} 1 & \mu_1 \mathbf{1}^\top_{r/2} & \cdots & * \\ * & \mu_1 J_{r/2,r/2} & \cdots & \vdots \\ * & * & \ddots & \vdots \\ * & * & * & \mu_{\mathcal{I}_k} J_{q,q} \end{bmatrix}, \tag{A.39}$$

for some $q \in \mathbb{N}$. Similarly, we can permute the matrix $M_{\bar{k}}(\mathbf{y}(\mathcal{I}_k))$ into the above form. Thus, there exists M such that $T_{d_1,\ldots,d_{\bar{k}}} M T^\top_{d_1,\ldots,d_{\bar{k}}} = \hat{P} M_{\bar{k}}(\mathbf{y}(\mathcal{I}_k))\hat{P}^{-1}$ for some sequence of $d_1, \ldots, d_{\bar{k}}$. Furthermore, there exists another sequence $\{d_1, \ldots, d_{r/2}\}$ such that $T_{d_1,\ldots,d_{r/2}} M T^\top_{d_1,\ldots,d_{r/2}} = \hat{M}_r$. Consequently, applying Lemma 20 twice, we have that if $M_{\bar{k}}(\mathbf{y}(\mathcal{I}_k)) \succeq 0$ then $M_r(\mathbf{y}_\infty(\mathcal{I}_k)) \succeq 0$. The above claim holds for arbitrary r, thus the result follows. The relationship between finite and infinite dimensional localizing matrices can be shown using the above procedure. □

Proof of Theorem 30. To show the monotone relationship $\hat{\mu}_{\mathcal{I}}(t) \geq \mu_{\mathcal{I}}(t) \geq \check{\mu}_{\mathcal{I}}(t)$ holds for all $\mathcal{I}$ and $t \geq 0$, we apply the multi-variate comparison lemma, i.e., Theorem 1.2 from [214]. More specifically, we aim to show that when $\hat{\mu}_{\mathcal{I}} = \mu_{\mathcal{I}}$, $\dot{\hat{\mu}}_{\mathcal{I}} \geq \dot{\mu}_{\mathcal{I}}$ for all $\hat{\mu}_{\mathcal{J}} \geq \mu_{\mathcal{J}}, \mathcal{J} \neq \mathcal{I}$ and $\check{\mu}_{\mathcal{J}} \leq \mu_{\mathcal{J}}, \forall |\mathcal{J}| \leq k$.

On one hand, when $|\mathcal{I} \cup \{\ell\}\}| \leq k$, we have that all the terms with positive coefficients are bounded above by upper estimates $\hat{\mu}_{\mathcal{I}\cup\ell\setminus\{i_s\}}$, whereas the terms with negative coefficients are bounded below by $\check{\mu}_{\mathcal{I}\cup\{\ell\}}$; thence $\dot{\hat{\mu}}_{\mathcal{I}} \geq \dot{\mu}_{\mathcal{I}}$ holds.

On the other hand, when $|\mathcal{I} \cup \{\ell\}\}| = k+1$, it suffices to show that $\mu_{\mathcal{I}\cup\{\ell\}}$ is feasible in the SDPs (7.25). Consider the random variable $\mathbf{x}_{\mathcal{I}\cup\{\ell\}} = [x_{i_1}, \ldots, x_{\ell}]^{\top}$, its underlying measure at time t is supported on $\tilde{S}_{|\mathcal{I}|+1}$, which is a compact and semi-algebraic set. Let $\mathbf{y}$ be the infinite multi-sequence consisting of all the moments of $\mathbf{x}_{\mathcal{I}\cup\{\ell\}}$. From $\mathbf{y}$, we can readily construct moment matrix $M_r(\mathbf{y})$, and localizing matrices $L_r(g_j\mathbf{y})$, for any given $r \in \mathbb{N}$. Consequently, according to Theorem 1, these matrices are positive semi-definite. Moreover, according to Lemma 15, the positive semi-definiteness of these matrices are equivalent to positive semi-definiteness of $M_{\tilde{k}}(\mathbf{y}(\mathcal{I} \cup \{\ell\}))$, $L^1_{\tilde{k}}(\mathbf{y}(\mathcal{I} \cup \{\ell\}), s)$, and $L^0_{\tilde{k}}(\mathbf{y}(\mathcal{I} \cup \{\ell\}), s)$ for all $s \in [k]$. Consequently, $\{\mu_{\mathcal{J}}\}_{\mathcal{J} \subseteq \mathcal{I}\cup\{\ell\}}$ is a feasible solution to both 7.23 and (7.24). Meanwhile, the eigenvalues of (7.21) and (7.22) are monotonic in terms of their entries, which implies that $\{\mu_{\mathcal{J}}\}_{\mathcal{J} \subseteq \mathcal{I}\cup\{\ell\}}$ is also feasible with respect to both the minimization and maximization problems (7.25). Furthermore, this holds for all $\hat{\mu}_{\mathcal{J}} \geq \mu_{\mathcal{J}}, \mathcal{J} \neq \mathcal{I}$ and $\check{\mu}_{\mathcal{J}} \leq \mu_{\mathcal{J}}, \forall |\mathcal{J}| \leq k$. Summarizing the above claims, we have that if $\hat{\mu}_{\mathcal{I}}(0) \geq \mu_{\mathcal{I}}(0)$, then $\hat{\mu}_{\mathcal{I}}(t) \geq \mu_{\mathcal{I}}(t)$ holds for all $\mathcal{I}$ and $t \geq 0$. The above argument readily applies to the comparison between $\mu_{\mathcal{I}}(t)$ and $\check{\mu}_{\mathcal{I}}(t)$. □

Proof of Theorem 31. To show the results, we use the fact that a matrix is positive semi-definite, if and only if, all its principal minors are non-negative [215] and apply it to (7.26)– (7.29), respectively. Consider constraint (7.26), we must require the determinant of $M_1(\mathbf{y}(\mathcal{I}_2))$ to be non-negative. Notice that, $\det(M_1(\mathbf{y}(\mathcal{I}_2))) = -\mu_{ij}^2 + 2\mu_i\mu_j\mu_{ij} - \mu_i\mu_j(\mu_i + \mu_j - 1)$, which is quadratic in terms of μ_{ij}. As a result,

$\det(M_1(\mathbf{y}(\mathcal{I}_2))) \geq 0$ is equivalent to:

$$\mu_{ij} \in \left[\mu_i\mu_j - h(\mu_i,\mu_j), \mu_i\mu_j + h(\mu_i,\mu_j)\right],$$

where $h(x,y) = \sqrt{x(1-x)y(1-y)}$ for all $x, y \in [0,1]$. Moreover, we require all principal minors of $M_1(\mathbf{y}(\mathcal{I}_2))$ to be non-negative. In particular,

$$\det\left(\begin{bmatrix} \mu_i & \mu_{ij} \\ \mu_{ij} & \mu_j \end{bmatrix}\right) \geq 0$$

indicates that $\mu_{ij} \leq \sqrt{\mu_i\mu_j}$.

Similarly, we compute all principal minors of localizing matrices and require them to be non-negative. Therefore, from $L_1^0(\mathbf{y}(\mathcal{I}_2), i), L_1^0(\mathbf{y}(\mathcal{I}_2), j) \succeq 0$, we obtain $\mu_{ij} \leq \min\{\mu_i, \mu_j\}$. From $L_1^1(\mathbf{y}(\mathcal{I}_2), i), L_1^1(\mathbf{y}(\mathcal{I}_2), j) \succeq 0$, we have $\mu_{ij} \geq \mu_i + \mu_j - 1$.

Therefore, $\{\mu_i, \mu_j, \mu_{ij}\}$ is a feasible moment sequence provided that all the above constraints on μ_{ij} are satisfied simultaneously. This is equivalent to $\mu_{ij} \in [l_{ij}, u_{ij}]$, where

$$u_{ij} = \min\{\mu_i, \mu_j, \sqrt{\mu_i\mu_j}, \mu_i\mu_j + h(\mu_i,\mu_j)\}, \tag{A.40}$$

and

$$l_{ij} = \max\{\mu_i\mu_j - h(\mu_i,\mu_j), \mu_i + \mu_j - 1, 0\}. \tag{A.41}$$

Notice that u_{ij} can be further simplified as follows. When $\mu_i \leq \mu_j$, then (A.40) is equivalent to $\min\{\mu_i, \sqrt{\mu_i\mu_j}, \mu_i\mu_j + h(\mu_i,\mu_j)\}$. Under this circumstance, $\mu_i = \sqrt{\mu_i}\sqrt{\mu_i} \leq \sqrt{\mu_i\mu_j}$. Moreover,

$$\begin{aligned} \mu_i \leq \mu_j &\Rightarrow (1-\mu_j)\mu_i \leq (1-\mu_i)\mu_j, \\ &\Rightarrow (1-\mu_j)^2\mu_i^2 \leq \mu_i\mu_j(1-\mu_i)(1-\mu_j), \\ &\Rightarrow \mu_i \leq \mu_i\mu_j + h(\mu_i,\mu_j). \end{aligned}$$

As a result, when $\mu_i \leq \mu_j$, $u_{ij} = \mu_i$. Similarly, when $\mu_j \leq \mu_i$, $u_{ij} = \mu_j$. Subsequently, (A.40) can be simplified into:

$$u_{ij} = \min\{\mu_i, \mu_j\}.$$

We adopt an analogous procedure for simplifying (A.41). When $\mu_i + \mu_j \geq 1$, then $l_{ij} = \max\{\mu_i\mu_j - h(\mu_i, \mu_j), \mu_i + \mu_j - 1\}$. In this case, we have:

$$\begin{aligned}\mu_i + \mu_j - 1 \geq 0 &\Rightarrow (1-\mu_i)(1-\mu_j) \leq \mu_i\mu_j,\\ &\Rightarrow (1-\mu_i)(1-\mu_j) \leq h(\mu_i, \mu_j),\\ &\Rightarrow \mu_i + \mu_j - 1 \geq \mu_i\mu_j - h(\mu_i, \mu_j).\end{aligned}$$

Consequently, we obtain that:

$$l_{ij} = \max\{0, \mu_i + \mu_j - 1\}.$$

Since $\beta_{ij} > 0$, maximizing over $-\sum_{j=1}^{n} \beta_{ij}\mu_{ij}$ is equivalent to minimize μ_{ij}. In particular, the minimum of μ_{ij} is attained at l_{ij}. Thus, the upper bound is obtained. Similarly, to obtain the lower bound, we maximize μ_{ij} for all $i, j \in [n]$. The optimal solution of these two problems are $\overline{\mu}^{\star}_{ij}$ and $\underline{\mu}^{\star}_{ij}$, respectively.

Next, we aim to show that $\hat{\mu}_i(t) \geq \mu_i(t) \geq \check{\mu}_i(t)$. To achieve this goal, we utilize Proposition 1.4 from [214]. It suffice to show that when $\mu_i(t) = \hat{\mu}_i(t)$, $\dot{\hat{\mu}}_i \geq \dot{\mu}_i$ for all $\hat{\mu}_j \geq \mu_j$. Consider the difference between $\dot{\hat{\mu}}_i$ and $\dot{\mu}_i$

$$\begin{aligned}\frac{d\hat{\mu}_i}{dt} - \frac{d\mu_i}{dt} &= \sum_j \beta_{ij}\left[\hat{\mu}_j - \mu_j + \mu_{ij} - \overline{\mu}_{ij}\right]\\ &= \sum_j \left[\hat{\mu}_j - \mu_j + \mu_{ij} - \max\{0, \hat{\mu}_j + \mu_i - 1\}\right]\end{aligned}$$

The above equality is due to the assumption that $\hat{\mu}_i = \mu_i$. Consider the following cases: (i) $\hat{\mu}_j = \mu_j$, then the right-hand-side is larger than zero according to (7.24); (ii) $\hat{\mu}_j > \mu_j$ and $\hat{\mu}_j + \mu_i - 1 \leq 0$, the RHS is non-negative trivially; and (iii) $\hat{\mu}_j > \mu_j$ and $\hat{\mu}_j + \mu_i - 1 > 0$, it follows that:

$$\begin{aligned}\frac{d\hat{\mu}_i}{dt} - \frac{d\mu_i}{dt} &= \sum_j \left[\hat{\mu}_j - \mu_j + \mu_{ij} - \hat{\mu}_j - \mu_i + 1\right] \\ &\geq \sum_j \left[\mu_{ij} - \max\{0, \mu_i + \mu_j - 1\}\right] \geq 0.\end{aligned}$$

As a consequence, $\hat{\mu}_i(t) \geq \mu_i(t)$ according to [214]. Similar analysis holds for comparing $\dot{\hat{\mu}}_i$ and $\dot{\mu}_i$. □

Proof of Theorem 32. Similar to the treatment in Lemma 19, we define $h_\ell : \mathbb{R}^{3n} \to \mathbb{R}^{3n}$, for $\ell \in [|\mathcal{V}| + |\mathcal{E}|]$. Each h_ℓ is defined as follows: (i) when $\ell \in [n]$, we let

$$h_\ell(\mathbf{x}) = [0, 0, 0, \ldots, 0, -x_{\ell,I}, x_{\ell,I}, \ldots, 0, 0, 0]^\top,$$

and (ii) for every $(i, j) \in \mathcal{E}$, we let

$$h_{(i,j)}(\mathbf{x}) = [0, 0, 0, \ldots, -x_{i,S}x_{j,I}, x_{i,S}x_{j,I}, \ldots, 0, 0, 0]^\top.$$

With these definitions, according to (7.9), when $\ell \in [n]$, we have that $\phi_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathbf{x} + h_\ell(\mathbf{x})) - \phi_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathbf{x})$ equals to

$$\Pi_{k\in[n],k\neq\ell} x_{k,S}^{\alpha_k} x_{k,I}^{\beta_k} x_{k,R}^{\gamma_k} \left[x_{\ell,S}^{\alpha_\ell} 0^{\beta_\ell} (x_{\ell,R} + x_{\ell,I})^{\gamma_\ell} - x_{\ell,S}^{\alpha_\ell} x_{\ell,I}^{\beta_\ell} x_{\ell,R}^{\gamma_\ell}\right].$$

Consequently, when $\beta_\ell \neq 0$, the term above equals to $\phi_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathbf{x})$. Meanwhile, $x_{\ell,I}$ and $x_{\ell,R}$ are binary variables. Moreover, $x_{\ell,R} + x_{\ell,I} \leq 1$ for all $t \geq 0$. Thus, $(x_{\ell,I} + x_{\ell,R})^{\gamma_\ell} - x_{\ell,I}^{\gamma_\ell} = x_{\ell,I}$. Subsequently, we have

$$\phi_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathbf{x} + h_\ell(\mathbf{x})) - \phi_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathbf{x}) = \begin{cases} -\phi_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathbf{x}), \text{ if } \beta_\ell \neq 0, \\ \Pi_{k\in[n],k\neq\ell} x_{k,S}^{\alpha_k} x_{k,I}^{\beta_k} x_{k,R}^{\gamma_k} x_{\ell,S}^{\alpha_\ell} x_{\ell,I}, \text{ if } \beta_\ell = 0 \text{ and } \gamma_\ell \neq 0, \\ 0, \text{ otherwise.} \end{cases} \tag{A.42}$$

Similarly, for a given $(j, i) \in \mathcal{E}$, we have

$$\begin{aligned}
&\phi_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathbf{x} + h_{(j,i)}(\mathbf{x})) - \phi_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathbf{x}) = \\
&\Pi_{k\in[n],k\neq i} x_{k,S}^{\alpha_k} x_{k,I}^{\beta_k} x_{k,R}^{\gamma_k} x_{i,S}^{\alpha_i} x_{i,R}^{\gamma_i} \\
&\times \left[(1 - x_{j,I})^{\alpha_i} (x_{i,I} + x_{i,S} x_{j,I})^{\beta_i} - x_{i,I}^{\beta_i}\right].
\end{aligned} \tag{A.43}$$

On one hand, when $\alpha_i \neq 0$, we observe that if $x_{j,I} = 0$, then $(1-x_{j,I})^{\alpha_i}(x_{i,I}-x_{i,S}x_{j,I})^{\beta_i} - x_{i,I}^{\beta_i}$ equals to zero, whereas if $x_{j,I} = 1$, the term equals to $-x_{i,I}^{\beta_i}$.

On the other hand, when $\alpha_i = 0$ and $\beta_i = 0$, (A.43) equals to 0. Finally, we consider the case when $\alpha_i = 0$ and $\beta_i \neq 0$. In this context, it suffices to examine $(x_{i,I}+x_{i,S}x_{j,I})^{\beta_i} - x_{i,I}^{\beta_i}$. Notice that when $x_{i,S} = 0$ the sum equals to 0, and $x_{j,I}$ otherwise. Thus, it can be simplified into $x_{i,S}x_{j,I}$. Summarizing the above cases, let

$$Q = \phi_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}} / x_{i,I}^{\beta_i}$$

we obtain that:

$$\begin{aligned}
&\phi_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathbf{x} + h_{(j,i)}(\mathbf{x})) - \phi_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathbf{x}) = \\
&\begin{cases}
-\phi_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathbf{x}) x_{j,I}, \text{ if } \alpha_i \neq 0, \\
Q x_{i,S} x_{j,I}, \text{ if } \alpha_i = 0 \text{ and } \beta_i \neq 0, \\
0, \text{ otherwise.}
\end{cases}
\end{aligned} \tag{A.44}$$

Finally, (7.37) is obtained by taking expectation on the sum of $|\mathcal{V}| + |\mathcal{E}|$ equations (A.42) and (A.44). □

A.7 Proof of the Results in Chapter 8

Proof of Theorem 37. First, we show that when the initial distribution μ_0 and the system dynamics (7.2) are given, the Liouville equation (8.23) has a unique solution (μ, μ_T) up to a subset of $[0, T] \times \mathcal{X}$ of Lebesgue measure zero and (μ, μ_T) coincide with the average occupation measure defined by (8.21) and the average final measure defined by

(8.22). Let (μ, μ_T) be a pair of measures satisfying (8.23). From [201, Lemma 3], μ can be disintegrated as $d\mu(t, \mathbf{x}) = d\mu_t(\mathbf{x})dt$ where dt is the Lebesgue measure on $[0, T]$. $\mu_t(\mathbf{x})$ is a stochastic kernel on $\mathcal{X}$ given t and can be interpreted as the distribution of the states at time t following the evolution of (7.2) with $\mathbf{x}_0 \sim \mu_0$. $\mu_t(\mathbf{x})$ is uniquely defined dt-almost everywhere. As proved in [201, Lemma 3], μ_t satisfies a continuity equation which implies μ and μ_T coincide with the average occupation measure and the average final measure generated by the family of absolutely continuous admissible trajectories of (7.2) starting from μ_0.

Then solving $\mathtt{P}$ can be decomposed into two steps: first find a feasible $(\mu, \mu_T) \in \mathcal{M}_+([0, T] \times \mathcal{X}) \times \mathcal{M}_+(\mathcal{X})$ to the Liouville equation $\delta_T \otimes \mu_T = \delta_0 \otimes \mu_0 + \mathcal{L}^* \mu$ and then solve the following optimization problem:

$$\mathtt{Q} : \sup_{\widetilde{\mu}} \{ \int g d\widetilde{\mu} : \widetilde{\mu} \leq \mu ; \widetilde{\mu} \in \mathcal{M}_+([0, T] \times \mathcal{X}_u) \}. \tag{A.45}$$

Since $\mathcal{X}$ and $\mathcal{X}_u$ are compact with $\mathcal{X}_u \subseteq \mathcal{X}$, by [199, Theorem 3.1] the restriction $\widetilde{\mu}^*$ of μ to $\mathcal{X}_u$ defined by (8.27) is the unique optimal solution to $\mathtt{Q}$ and $\sup \mathtt{Q} = \max \mathtt{Q} = \int g d\widetilde{\mu}^* = \int_{\mathcal{X}_u} g d\mu$.

As the feasible μ in $\mathtt{P}$ coincides with the average occupation measure in (8.21), $\widetilde{\mu}^*$ is also the $\widetilde{\mu}$-component of an optimal solution to $\mathtt{P}$ and $\sup \mathtt{P} = \max \mathtt{P} = \int g d\widetilde{\mu}^*$. When $g \equiv 1$, we have $\max \mathtt{P} = \mu([0, T] \times \mathcal{X}_u)$ with μ being the average occupation measure defined in (8.21). □

Proof of Theorem 38. The proof follows the same lines as that of [201, Theorem 2]. Define

$$\mathbf{C} = \mathcal{C}([0, T] \times \mathcal{X}_u) \times \mathcal{C}([0, T] \times \mathcal{X}) \times \mathcal{C}([0, T] \times \mathcal{X}) \times \mathcal{C}(\mathcal{X})$$

$$\mathbf{M} = \mathcal{M}([0, T] \times \mathcal{X}_u) \times \mathcal{M}([0, T] \times \mathcal{X}) \times \mathcal{M}([0, T] \times \mathcal{X}) \times \mathcal{M}(\mathcal{X})$$

and let $\mathcal{K}$ and $\mathcal{K}'$ denote the positive cones of $\mathbf{C}$ and $\mathbf{M}$, respectively. By Riesz-Markov-Kakutani representation theorem [203], $\mathcal{K}'$ is the topological dual of the cone $\mathcal{K}$. The infinite dimensional linear program P can be written as:

$$\begin{aligned} &\sup \langle \gamma, c\rangle \\ &\text{s.t.} \mathcal{A}'\gamma = \beta, \quad \gamma \in \mathcal{K}' \end{aligned} \tag{A.46}$$

where the supremum is taken over the vector $\gamma = (\widetilde{\mu}, \widehat{\mu}, \mu, \mu_T)$, the linear operator $\mathcal{A}' : \mathcal{K}' \to \mathcal{C}^1([0,T]\times\mathcal{X})^* \times \mathcal{M}([0,T]\times\mathcal{X})$ is defined by $\mathcal{A}'\gamma = (\delta_T \otimes \mu_T - \mathcal{L}^*\mu, \mu - \widetilde{\mu} - \widehat{\mu})$ and $\beta = (\delta_0 \otimes \mu_0, 0) \in \mathcal{C}^1([0,T]\times\mathcal{X})^* \times \mathcal{M}([0,T]\times\mathcal{X})$. The vector of functions in the objective is $c = (g, 0, 0, 0)$. Define the duality bracket between a vector of measures $\nu \in (\mathcal{M}(\mathcal{S}))^p$ and a vector of functions $h \in (\mathcal{C}(\mathcal{S}))^p$ over a topological space $\mathcal{S}$ by $\langle h, \nu\rangle = \sum_{i=1}^p \int_{\mathcal{S}} [h]_i d[\nu]_i$. Then $\langle \gamma, c\rangle = \int g d\widetilde{\mu}$.

The dual to (A.46) can be interpreted as:

$$\begin{aligned} &\inf \langle \beta, z\rangle \\ &\text{s.t.} \mathcal{A}z - c \in \mathcal{K} \end{aligned} \tag{A.47}$$

where the infimum is over $z = (v, w) \in \mathcal{C}^1([0,T]\times\mathcal{X}) \times \mathcal{C}([0,T]\times\mathcal{X})$, the linear operator $\mathcal{A} : \mathcal{C}^1([0,T]\times\mathcal{X}) \times \mathcal{C}([0,T]\times\mathcal{X}) \to \mathbf{C}$ is given by $\mathcal{A}z = (w, w, -\mathcal{L}v - w, v(T,\cdot))$ and satisfies the adjoint property $\langle \mathcal{A}'\gamma, z\rangle = \langle \gamma, \mathcal{A}z\rangle$. The linear program (A.47) is exactly (8.28).

From [216, Theorem 3.10], there is no duality gap between (A.46) and (A.47) if the supremum of (A.46) is finite and the set $P = \{(\mathcal{A}'\gamma, \langle \gamma, c\rangle) \mid \gamma \in \mathcal{K}'\}$ is closed in the weak* topology of $\mathcal{K}'$. Since $\widetilde{\mu}$ is dominated by the average occupation measure μ and its underlying support is compact, the supremum of (A.46) is finite. To prove that P is closed, consider a sequence $\gamma_k = (\widetilde{\mu}^k, \widehat{\mu}^k, \mu^k, \mu_T^k) \in \mathcal{K}'$ such that $\mathcal{A}'\gamma_k \to a$ and $\langle \gamma_k, c\rangle \to b$ as $k \to \infty$ for some $(a, b) \in \mathcal{C}^1([0,T]\times\mathcal{X})^* \times \mathcal{M}([0,T]\times\mathcal{X}) \times \mathbb{R}$. Consider the test function $z_1 = (T - t, 0)$ which gives $\langle \mathcal{A}'\gamma_k, z_1\rangle = \mu^k([0,T]\times\mathcal{X}) \to \langle a, z_1\rangle < \infty$; since the measures μ^k are non-negative, this implies $\{\mu^k\}$ is bounded. By taking $z_2 = (1, -1)$, we have $\langle \mathcal{A}'\gamma_k, z_2\rangle = \mu_T^k(\mathcal{X}) + \widetilde{\mu}^k([0,T]\times\mathcal{X}_u) + \widehat{\mu}^k([0,T]\times\mathcal{X}) - \mu^k([0,T]\times\mathcal{X}) \to \langle a, z_2\rangle < \infty$; since $\{\mu^k\}$ is bounded, by similar arguments the sequences $\{\widetilde{\mu}^k\}$, $\{\widehat{\mu}^k\}$ and $\{\mu_T^k\}$ are bounded as well.

As a result, $\{\gamma_k\}$ is bounded and we can find a ball B in $\mathbf{M}$ with $\{\gamma_k\} \subset B$. From the weak* compactness of the unit ball (Alaoglu's theorem [217, Section 5.10, Theorem 1]) there is a subsequence $\{\gamma_{k_i}\}$ that weak*-converges to some $\gamma \in \mathcal{K}'$. Notice that $\mathcal{A}'$ is weak*-continuous because $\mathcal{A}z \in \mathbf{C}$ for all $z \in \mathcal{C}^1([0,T] \times \mathcal{X}) \times \mathcal{C}([0,T] \times \mathcal{X})$. So $(a,b) = \lim_{i\to\infty}(\mathcal{A}'\gamma_{k_i}, \langle\gamma_{k_i}, c\rangle) = (\mathcal{A}'\gamma, \langle\gamma, c\rangle) \in P$ by the continuity of $\mathcal{A}'$ and P is closed. □

Proof of Theorem 39. The proof of strong duality follows from standard SDP duality theory. Let $\Delta_\mu = (\widetilde{\mu}, \widehat{\mu}, \mu, \mu_T)$ be the optimal solution to $\mathtt{P}$ and $\Delta_y = (\widetilde{\mathbf{y}}, \widehat{\mathbf{y}}, \mathbf{y}, \mathbf{y_T})$ be their corresponding moment sequences. Any finite truncation of Δ_y gives a feasible solution to $\mathtt{P_r}$. As $\mathcal{X}$ and $\mathcal{X}_u$ have non-empty interior, we have the truncation of Δ_y is strictly feasible for $\mathtt{P_r}$. By Slater's condition [109], there is no duality gap between $\mathtt{P_r}$ and $\mathtt{D_r}$, i.e., $p_r^* = d_r^*$.

The proof of convergence follows from [200, Theorem 3.6]. Since $[0,T]$, $\mathcal{X}$ and $\mathcal{X}_u$ are compact sets, we can assume after appropriate scaling $T = 1$ and $\mathcal{X} \times \mathcal{X}_u \subseteq [-1,1]^{n_\mathcal{X}} \times [-1,1]^{n_{\mathcal{X}_u}}$, which implies that the feasible set of the semidefinite program $\mathtt{P_r}$ is compact. Let $\Delta_r^* = (\widetilde{\mathbf{y}}_r^*, \widehat{\mathbf{y}}_r^*, \mathbf{y}_r^*, \mathbf{y_T}_r^*)$ be the optimal solution of $\mathtt{P_r}$ and complete the finite vectors $(\widetilde{\mathbf{y}}_r^*, \widehat{\mathbf{y}}_r^*, \mathbf{y}_r^*, \mathbf{y_T}_r^*)$ with zeros to make them infinite sequences. By a standard diagonal argument, there is a subsequence $\{r_k\}$ and a tuple of infinite vectors $\Delta^* = (\widetilde{\mathbf{y}}^*, \widehat{\mathbf{y}}^*, \mathbf{y}^*, \mathbf{y_T}^*)$ such that $\Delta_{r_k}^* \to \Delta^*$ as $k \to \infty$, where the convergence is interpreted as elementary-wise. Since the infinite vector $\widetilde{\mathbf{y}}^*$ in Δ^* is the limit point of a subsequence of the optimal solutions $\widetilde{\mathbf{y}}_r^*$ of $\mathtt{P_r}$, $\widetilde{\mathbf{y}}^*$ satisfies all the constraints in $\mathtt{P_r}$ as $r \to \infty$. Then by *Putinar's Positivstellensatz*, $\widetilde{\mathbf{y}}^*$ has a representing measure $\widetilde{\mu}^*$ supported on $[0,T] \times \mathcal{X}_u$. Similarly, $\widehat{\mathbf{y}}^*, \mathbf{y}^*$ and $\mathbf{y_T}^*$ have their representing measures $\widehat{\mu}^*, \mu^*$ and μ_T^* with corresponding supports, respectively.

As problem $\mathtt{P_r}$ is a relaxation of $\mathtt{P}$, $p_r^* \geq p^*$ for each r. Thus we have $\lim_{k\to\infty} \sup \mathtt{P_{r_k}} = \lim_{k\to\infty} L_{\widetilde{\mathbf{y}}_{r_k}^*}(g) = L_{\widetilde{\mathbf{y}}^*}(g) = \int g d\widetilde{\mu}^* \geq p^*$. On the other hand, $\mathcal{A}_r(\Delta^*) = \lim_{k\to\infty} \mathcal{A}_r(\Delta_{r_k}^*) = b_r$ for each $r \in \mathbb{N}$. Let $(\widetilde{\mu}^*, \widehat{\mu}^*, \mu^*, \mu_T^*)$ be the tuple of representing measures of Δ^*. As measures on compact sets are determined by moments, $(\widetilde{\mu}^*, \widehat{\mu}^*, \mu^*, \mu_T^*)$ is a feasible solution to $\mathtt{P}$ which implies $\int g d\widetilde{\mu}^* \leq p^*$. Hence $\int g d\widetilde{\mu}^* = p^*$ and $(\widetilde{\mu}^*, \widehat{\mu}^*, \mu^*, \mu_T^*)$ is an optimal solution of $\mathtt{P}$. For any r we have $p_r^* \geq p_{r+1}^*$ because as r increases, the constraints in $\mathtt{P}_\mathtt{r}$ become more restrict. As a result, $p_{r_k}^* \downarrow p^*$ and furthermore $p_r^* \downarrow p^*$. By strong duality, $d_r^* = p_r^* \downarrow p^* = d^*$. □

Bibliography

[1] S. H. Strogatz. Exploring complex networks. *Nature*, 410(6825):268, 2001.

[2] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[3] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5):175–308, 2006.

[4] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651, 2000.

[5] U. Alon. *An introduction to systems biology: design principles of biological circuits*. Chapman and Hall/CRC, 2006.

[6] U. Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450, 2007.

[7] M. D. Greicius, B. Krasnow, A. L. Reiss, and V. Menon. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences*, 100(1):253–258, 2003.

[8] P. Hagmann, L. Cammoun, X. Gigandet, R. Meuli, C. J. Honey, V. J. Wedeen, and O. Sporns. Mapping the structural core of human cerebral cortex. *PLoS Biology*, 6(7):e159, 2008.

[9] E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186, 2009.

[10] M. Rubinov and O. Sporns. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, 52(3):1059–1069, 2010.

[11] C. L. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239, 1978.

[12] S. Wasserman. *Advances in social network analysis: Research in the social and behavioral sciences*. Sage, 1994.

[13] M. E. Newman, D. J. Watts, and S. H. Strogatz. Random graph models of social networks. *Proceedings of the National Academy of Sciences*, 99(suppl 1):2566–2572, 2002.

[14] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, pages 29–42. ACM, 2007.

[15] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *ACM SIGCOMM Computer Communication Review*, volume 29, pages 251–262. ACM, 1999.

[16] R. Albert, H. Jeong, and A. L. Barabási. Internet: Diameter of the world-wide web. *Nature*, 401(6749):130, 1999.

[17] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Computer Networks*, 33 (1-6):309–320, 2000.

[18] P. Mahadevan, D. Krioukov, M. Fomenkov, X. Dimitropoulos, K. C. Claffy, and A. Vahdat. The internet as-level topology: three data sources and one definitive metric. *ACM SIGCOMM Computer Communication Review*, 36(1):17–26, 2006.

[19] C. Nowzari, V. M. Preciado, and G. J. Pappas. Analysis and control of epidemics: A survey of spreading processes on complex networks. *IEEE Control Systems*, 36 (1):26–46, 2016.

[20] S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca. Network analysis in the social sciences. *Science*, 323(5916):892–895, 2009.

[21] F. Dörfler, M. Chertkov, and F. Bullo. Synchronization in complex oscillator networks and smart grids. *Proceedings of the National Academy of Sciences*, 110 (6):2005–2010, 2013.

[22] J. A. Bondy and U. S. R. Murty. *Graph theory with applications*, volume 290. Citeseer, 1976.

[23] D. B. West. *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River, NJ, 1996.

[24] C. Godsil and G. F. Royle. *Algebraic graph theory*, volume 207. Springer Science & Business Media, 2013.

[25] D. D. Siljak. *Large-Scale Dynamic Systems: Stability and Structure*. Dover Publications, 2007.

[26] R. Barco, L. Díez, V. Wille, and P. Lázaro. Automatic diagnosis of mobile communication networks under imprecise parameters. *Expert Systems with Applications*, 36(1):489–500, 2009.

[27] F. Pasqualetti, S. Zampieri, and F. Bullo. Controllability metrics, limitations and algorithms for complex networks. *IEEE Transactions on Control of Network Systems*, 1(1):40–52, 2014.

[28] Y. Y. Liu, J. J. Slotine, and A. L. Barabási. Controllability of complex networks. *Nature*, 473(7346):167, 2011.

[29] R. E. Kalman. Mathematical description of linear dynamical systems. *Journal of the Society for Industrial and Applied Mathematics, Series A: Control*, 1(2): 152–192, 1963.

[30] K. Murota. *Systems analysis by graphs and matroids: structural solvability and controllability*, volume 3. Springer Science & Business Media, 2012.

[31] C. T. Lin. Structural controllability. *IEEE Transactions on Automatic Control*, 19(3):201–208, 1974.

[32] R. Shields and J. Pearson. Structural controllability of multiinput linear systems. *IEEE Transactions on Automatic Control*, 21(2):203–212, 1976.

[33] K. Glover and L. Silverman. Characterization of structural controllability. *IEEE Transactions on Automatic control*, 21(4):534–537, 1976.

[34] J. M. Dion, C. Commault, and J. Van Der Woude. Generic properties and control of linear structured systems: a survey. *Automatica*, 39(7):1125–1144, 2003.

[35] S. Hosoe and K. Matsumoto. On the irreducibility condition in the structural controllability theorem. *IEEE Transactions on Automatic Control*, 24(6):963–966, 1979.

[36] S. Hosoe. Determination of generic dimensions of controllable subspaces and its application. *IEEE Transactions on Automatic Control*, 25(6):1192–1196, 1980.

[37] S. A. Myers, A. Sharma, P. Gupta, and J. Lin. Information network or social network?: the structure of the twitter follow graph. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 493–498. ACM, 2014.

[38] G. A. Pagani and M. Aiello. The power grid as a complex network: a survey. *Physica A: Statistical Mechanics and its Applications*, 392(11):2688–2700, 2013.

[39] J. Corfmat and A. Morse. Structurally controllable and structurally canonical systems. *IEEE Transactions on Automatic Control*, 21(1):129–131, 1976.

[40] B. D. Anderson and H. M. Hong. Structural controllability and matrix nets. *International Journal of Control*, 35(3):397–416, 1982.

[41] T. Menara, D. S. Bassett, and F. Pasqualetti. Structural controllability of symmetric networks. *IEEE Transactions on Automatic Control*, to be published.

[42] S. S. Mousavi, M. Haeri, and M. Mesbahi. On the structural and strong structural controllability of undirected networks. *IEEE Transactions on Automatic Control*, 2017.

[43] J. Gao, Y. Y. Liu, R. M. D'souza, and A. L. Barabási. Target control of complex networks. *Nature Communications*, 5:5415, 2014.

[44] E. Czeizler, K. C. Wu, C. Gratie, K. Kanhaiya, and I. Petre. Structural target controllability of linear networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2018.

[45] I. Petre. Target controllability of linear networks. In *Computational Methods in Systems Biology: 14th International Conference, CMSB 2016, Cambridge, UK, September 21-23, 2016, Proceedings*, volume 9859, page 67. Springer, 2016.

[46] N. Monshizadeh, K. Camlibel, and H. L. Trentelman. Strong targeted controllability of dynamical networks. In *Proceedings of the IEEE Conference on Decision and Control*, pages 4782–4787. IEEE, 2015.

[47] H. J. van Waarde, M. K. Camlibel, and H. L. Trentelman. A distance-based approach to strong target control of dynamical networks. *IEEE Transactions on Automatic Control*, 62(12):6266–6277, 2017.

[48] K. Murota and S. Poljak. Note on a graph-theoretic criterion for structural output controllability. *IEEE Transactions on Automatic Control*, 35(8):939–942, 1990.

[49] S. Pequito, S. Kar, and A. P. Aguiar. A framework for structural input/output and control configuration selection in large-scale systems. *IEEE Transactions on Automatic Control*, 61(2):303 – 318, 2016.

[50] S. Assadi, S. Khanna, Y. Li, and V. M. Preciado. Complexity of the minimum input selection problem for structural controllability. *IFAC Workshop on Estimation and Control of Networked Systems*, 48(22):70–75, 2015.

[51] A. Olshevsky. Minimum input selection for structural controllability. In *Proceedings of American Control Conference*, pages 2218–2223. IEEE, 2015.

[52] S. Pequito, S. Kar, and A. P. Aguiar. Minimum cost input/output design for large-scale linear structural systems. *Automatica*, 68:384–391, 2016. ISSN 0005-1098. doi: http://dx.doi.org/10.1016/j.automatica.2016.02.005.

[53] S. Pequito, S. Kar, and A. P. Aguiar. On the complexity of the constrained input selection problem for structural linear systems. *Automatica*, 62:193–199, 2015. ISSN 0005-1098. doi: http://dx.doi.org/10.1016/j.automatica.2015.06.022.

[54] C. Commault, J. M. Dion, and J. W. Van der Woude. Characterization of generic properties of linear structured systems for efficient computations. *Kybernetika*, 38 (5):503–520, 2002.

[55] W. Wang, X. Ni, Y. Lai, and C. Grebogi. Optimizing controllability of complex networks by minimum structural perturbations. *Physical Review E*, 85(2):026115, 2012.

[56] J. Ding, Y. Lu, and J. Chu. Recovering the controllability of complex networks. In *Proceedings of the 19th IFAC World Congress*, volume 19, pages 10894–10901, 2014.

[57] C. H. Papadimitriou and K. Steiglitz. *Combinatorial optimization: algorithms and complexity.* Courier Corporation, 1982.

[58] Y. Zhang and T. Zhou. On the edge insertion/deletion and controllability distance of linear structural systems. In *Proceedings of the IEEE Conference on Decision and Control*, pages 2300–2305. IEEE, 2017.

[59] W. Ren, R. W. Beard, and E. M. Atkins. A survey of consensus problems in multi-agent coordination. In *Proceedings of the American Control Conference*, pages 1859–1864. IEEE, 2005.

[60] J. W. Simpson-Porco, F. Dörfler, and F. Bullo. Voltage stabilization in microgrids via quadratic droop control. *IEEE Transactions on Automatic Control*, 62(3): 1239–1253, 2017.

[61] K. Oh, M. Park, and H. Ahn. A survey of multi-agent formation control. *Automatica*, 53:424–440, 2015.

[62] A. Kirkoryan and M. A. Belabbas. Decentralized stabilization with symmetric topologies. In *Proceedings of the IEEE Conference on Decision and Control*, pages 1347–1352. IEEE, 2014.

[63] M. Pajic, R. Mangharam, G. J. Pappas, and S. Sundaram. Topological conditions for in-network stabilization of dynamical systems. *IEEE Journal on Selected Areas in Communications*, 31(4):794–807, 2013.

[64] Z. Pang and G.g Liu. Design and implementation of secure networked predictive control systems under deception attacks. *IEEE Transactions on Control Systems Technology*, 20(5):1334–1342, 2012.

[65] S. Pequito, G. Ramos, S. Kar, A. P. Aguiar, and J. Ramos. The robust minimal controllability problem. *Automatica*, 82:261–268, 2017.

[66] X. Liu, S. Pequito, S. Kar, Y. Mo, B. Sinopoli, and A. P. Aguiar. Minimum robust sensor placement for large scale linear time-invariant systems: a structured systems approach. In *Proceedings of the IFAC Workshop on Distributed Estimation and Control in Networked Systems (NecSys)*, pages 417–424, 2013.

[67] S. Pequito, F. Khorrami, P. Krishnamurthy, and G. J. Pappas. Analysis and design of secured/resilient closed-loop control systems. *IEEE Transactions on Automatic Control*, 2015.

[68] S. Moothedath, R. Gundeti, and P. Chaporkar. Verifying resiliency in closed-loop structured systems. *arXiv preprint arXiv:1712.08407*, 2017.

[69]

[70] M. Zhu and S. Martínez. On the performance analysis of resilient networked control systems under replay attacks. *IEEE Transactions on Automatic Control*, 59(3):804–808, 2014.

[71] C. De Persis and P. Tesi. Input-to-state stabilizing control under denial-of-service. *IEEE Transactions on Automatic Control*, 60(11):2930–2944, 2015.

[72] A. Rai, D. Ward, S. Roy, and S. Warnick. Vulnerable links and secure architectures in the stabilization of networks of controlled dynamical systems. In *Proceedings of the American Control Conference, 2012*, pages 1248–1253. IEEE, 2012.

[73] P. Erdős and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5:17–61, 1960.

[74] D. J. Watts and S. H. Strogatz. Collective dynamics of ‘small-world’ networks. *nature*, 393(6684):440, 1998.

[75] A. L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[76] M. Mihail and C. Papadimitriou. On the eigenvalue power law. In *International Workshop on Randomization and Approximation Techniques in Computer Science*, pages 254–262. Springer, 2002.

[77] D. Chakrabarti, Y. Zhan, and C. Faloutsos. R-mat: A recursive model for graph mining. In *Proceedings of the 2004 SIAM International Conference on Data Mining*, pages 442–446. SIAM, 2004.

[78] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 462–470. ACM, 2008.

[79] L. M. Pecora and T. L. Carroll. Master stability functions for synchronized coupled systems. *Physical Review Letters*, 80(10):2109, 1998.

[80] M. Barahona and L. M. Pecora. Synchronization in small-world systems. *Physical Review Letters*, 89(5):054101, 2002.

[81] J. D. Noh and H. Rieger. Random walks on complex networks. *Physical Review Letters*, 92(11):118701, 2004.

[82] S. Gomez, A. Diaz-Guilera, J. Gomez-Gardenes, C. J. Perez-Vicente, Y. Moreno, and A. Arenas. Diffusion dynamics on multiplex networks. *Physical Review Letters*, 110(2):028701, 2013.

[83] N. E. J. Newman. Spread of epidemic disease on networks. *Physical Review E*, 66 (1):016128, 2002.

[84] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani. Epidemic processes in complex networks. *Reviews of Modern Physics*, 87(3):925, 2015.

[85] A. Barrat, M. Barthelemy, and A. Vespignani. *Dynamical processes on complex networks*. Cambridge university press, 2008.

[86] M. A. Porter and J. P. Gleeson. Dynamical systems on networks. *Frontiers in Applied Dynamical Systems: Reviews and Tutorials*, 4, 2016.

[87] Y. Wang, D. Chakrabarti, C. Wang, and C. Faloutsos. Epidemic spreading in real networks: An eigenvalue viewpoint. In *Reliable Distributed Systems, 2003. Proceedings. 22nd International Symposium on*, pages 25–34. IEEE, 2003.

[88] D. Chakrabarti, Y. Wang, C. Wang, J. Leskovec, and C. Faloutsos. Epidemic thresholds in real networks. *ACM Transactions on Information and System Security (TISSEC)*, 10(4):1, 2008.

[89] V. M. Preciado. *Spectral analysis for stochastic models of large-scale complex dynamical networks*. PhD thesis, Massachusetts Institute of Technology, 2008.

[90] N. Crossley, E. Bellotti, G. Edwards, M. G. Everett, J. Koskinen, and M. Tranmer. *Social network analysis for ego-nets: Social network analysis for actor-centred networks*. Sage, 2015.

[91] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594): 824–827, 2002.

[92] S. Mangan and U. Alon. Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences*, 100(21):11980–11985, 2003.

[93] A. Barabási and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101, 2004.

[94] H. Wolkowicz and G. P. H. Styan. Bounds for eigenvalues using traces. *Linear Algebra and its Applications*, 29:471–506, 1980.

[95] H. Wolkowicz and G. P. H. Styan. More bounds for elgenvalues using traces. *Linear Algebra and its Applications*, 31:1–17, 1980.

[96] L. Kolotilina. Lower bounds for the perron root of a nonnegative matrix. *Linear Algebra and its Applications*, 180:133–151, 1993.

[97] B. G. Horne. Lower bounds for the spectral radius of a matrix. *Linear Algebra and its Applications*, 263:261–273, 1997.

[98] J. K. Merikoski and A. Virtanen. Bounds for eigenvalues using the trace and determinant. *Linear Algebra and its Applications*, 264:101–108, 1997.

[99] O. Rojo, R. Soto, and H. Rojo. Bounds for the spectral radius and the largest singular value. *Computers & Mathematics with Applications*, 36(1):41–50, 1998.

[100] A. Mercer and P. R. Mercer. Cauchy's interlace theorem and lower bounds for the spectral radius. *International Journal of Mathematics and Mathematical Sciences*, 23(8):563–566, 2000.

[101] J. K. Merikoski and A. Virtanen. The best possible lower bound for the perron root using traces. *Linear Algebra and its Applications*, 388:301–313, 2004.

[102] V. Nikiforov. Walks and the spectral radius of graphs. *Linear Algebra and its Applications*, 418(1):257–268, 2006.

[103] L. Wang, M. Xu, and T. Huang. Some lower bounds for the spectral radius of matrices using traces. *Linear Algebra and its Applications*, 432(4):1007–1016, 2010.

[104] V. M. Preciado and A. Jadbabaie. From local measurements to network spectral properties: Beyond degree distributions. In *Proceedings of IEEE Conference on Decision and Control*, pages 2686–2691. IEEE, 2010.

[105] V. M. Preciado and A. Jadbabaie. Moment-based spectral analysis of large-scale networks using local structural information. *IEEE/ACM Transactions on Networking*, 21(2):373–382, 2013.

[106] V. M. Preciado, A. Jadbabaie, and G. C. Verghese. Structural analysis of laplacian spectral properties of large-scale networks. *IEEE Transactions on Automatic Control*, 58(9):2338–2343, 2013.

[107] V. M. Preciado and M. M. Zavlanos. Distributed network design for laplacian eigenvalue placement. *IEEE Transactions on Control of Network Systems*, 4(3): 598–609, 2017.

[108] J. B. Lasserre. *Moments, Positive Polynomials and Their Applications*, volume 1. World Scientific, 2009.

[109] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[110] M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak mathematical journal*, 23(2):298–305, 1973.

[111] R. Olfati-Saber and R. M. Murray. Consensus problems in networks of agents with switching topology and time-delays. *IEEE Transactions on Automatic Control*, 49(9):1520–1533, 2004.

[112] R. Olfati-Saber, J. A. Fax, and R. M. Murray. Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95(1):215–233, 2007.

[113] N. M. M. De Abreu. Old and new results on algebraic connectivity of graphs. *Linear algebra and its applications*, 423(1):53–73, 2007.

[114] A. Ghosh and S. Boyd. Upper bounds on algebraic connectivity via convex optimization. *Linear algebra and its applications*, 418(2-3):693–707, 2006.

[115] M. Franceschelli, A. Gasparri, A. Giua, and C. Seatzu. Decentralized laplacian eigenvalues estimation for networked multi-agent systems. In *Proceedings of the IEEE Conference on Decision and Control*, pages 2717–2722. IEEE, 2009.

[116] P. Yang, R. A. Freeman, G. J. Gordon, K. M. Lynch, S. S. Srinivasa, and R. Sukthankar. Decentralized estimation and control of graph connectivity for mobile sensor networks. *Automatica*, 46(2):390–396, 2010.

[117] Z. Qu, C. Li, and F. Lewis. Cooperative control based on distributed estimation of network connectivity. In *Proceedings of the American Control Conference*, pages 3441–3446. IEEE, 2011.

[118] R. Aragues, G. Shi, D. V Dimarogonas, C. Sagues, and K. H. Johansson. Distributed algebraic connectivity estimation for adaptive event-triggered consensus. In *Proceedings of the American Control Conference*, pages 32–37. IEEE, 2012.

[119] A. Y. Kibangou and C. Commault. Decentralized laplacian eigenvalues estimation and collaborative network topology identification. *IFAC Proceedings Volumes*, 45 (26):7–12, 2012.

[120] T. Tran and A. Y. Kibangou. Distributed estimation of graph laplacian eigenvalues by the alternating direction of multipliers method. *IFAC Proceedings Volumes*, 47 (3):5526–5531, 2014.

[121] S. Leonardos, V.M. Preciado, and K. Daniilidis. Distributed spectral computations: Theory and applications. *Submitted for publication*, 2019.

[122] Z. Wu and V. M. Preciado. Laplacian spectral properties of graphs from random local samples. In *Proceedings of the SIAM International Conference on Data Mining*, pages 343–351. SIAM, 2014.

[123] J. O. Kephart and S. R. White. Directed-graph epidemiological models of computer viruses. In *Proceedings of IEEE Computer Society Symposium on Research in Security and Privacy*, pages 343–359. IEEE, 1991.

[124] Z. Liu, Y. Lai, and N. Ye. Propagation and immunization of infection on general networks with both homogeneous and heterogeneous components. *Physical Review E*, 67(3):031911, 2003.

[125] A. Vespignani. Modelling dynamical processes in complex socio-technical systems. *Nature Physics*, 8(1):32–39, 2012.

[126] J. Ahn and B. Hassibi. Global dynamics of epidemic spread over complex networks. In *Proceedings of IEEE Conference on Decision and Control*, pages 4579–4585. IEEE, 2013.

[127] P. Van Mieghem, J. Omic, and R. Kooij. Virus spread in networks. *IEEE/ACM Transactions on Networking*, 17(1):1–14, 2009.

[128] E. Volz and L. A. Meyers. Susceptible–infected–recovered epidemics in dynamic contact networks. *Proceedings of the Royal Society of London B: Biological Sciences*, 274(1628):2925–2934, 2007.

[129] F. D. Sahneh and C. Scoglio. Epidemic spread in human networks. In *Proceedings of IEEE Conference on Decision and Control and European Control Conference*, pages 3008–3013. IEEE, 2011.

[130] C. Nowzari, V. M. Preciado, and G. J. Pappas. Stability analysis of generalized epidemic models over directed networks. In *Proceedings of IEEE Conference on Decision and Control*, pages 6197–6202. IEEE, 2014.

[131] V. M. Preciado, M. Zargham, C. Enyioha, A. Jadbabaie, and G. J. Pappas. Optimal vaccine allocation to control epidemic outbreaks in arbitrary networks. In *Proceedings of IEEE Conference on Decision and Control*, pages 7486–7491. IEEE, 2013.

[132] V. M. Preciado, M. Zargham, C. Enyioha, A. Jadbabaie, and G. J. Pappas. Optimal resource allocation for network protection against spreading processes. *IEEE Transactions on Control of Network Systems*, 1(1):99–108, 2014.

[133] K. Drakopoulos, A. Ozdaglar, and J. N. Tsitsiklis. An efficient curing policy for epidemics on graphs. *IEEE Transactions on Network Science and Engineering*, 1 (2):67–75, 2014.

[134] Y. Wan, S. Roy, and A. Saberi. Designing spatially heterogeneous strategies for control of virus spread. *IET Systems Biology*, 2(4):184–201, 2008.

[135] E. Cator and P. Van Mieghem. Second-order mean-field susceptible-infected-susceptible epidemic threshold. *Physical Review E*, 85(5):056111, 2012.

[136] A. Lamperski, K. R. Ghusinga, and A. Singh. Analysis and control of stochastic systems using semidefinite programming over moments. *arXiv preprint arXiv:1702.00422*, 2017.

[137] N. J. Watkins, C. Nowzari, and G. J. Pappas. Robust prediction and control of continuous-time epidemic processes. *arXiv preprint arXiv:1707.00742*, 2017.

[138] A. S. Mata and S. C. Ferreira. Pair quenched mean-field theory for the susceptible-infected-susceptible model on complex networks. *Europhysics Letters*, 103(4): 48003, 2013.

[139] X. Chen, M. Ogura, K. R. Ghusinga, A. Singh, and V. M. Preciado. Semidefinite bounds for moment dynamics: Application to epidemics on networks. In *Proceedings of IEEE Conference on Decision and Control*, pages 2448–2454. IEEE, 2017.

[140] F. D. Sahneh, C. Scoglio, and P. Van Mieghem. Generalized epidemic mean-field model for spreading processes over multilayer complex networks. *IEEE/ACM Transactions on Networking (TON)*, 21(5):1609–1620, 2013.

[141] S. Funk and V. A. A. Jansen. Interacting epidemics on overlay networks. *Physical Review E*, 81(3):036118, 2010.

[142] X. Wei, N. Valler, B. A. Prakash, I. Neamtiu, M. Faloutsos, and C. Faloutsos. Competing memes propagation on networks: a case study of composite networks. *ACM SIGCOMM Computer Communication Review*, 42(5):5–12, 2012.

[143] F. D. Sahneh and C. Scoglio. May the best meme win!: New exploration of competitive epidemic spreading over arbitrary multi-layer networks. *arXiv preprint arXiv:1308.4880*, 2013.

[144] B. A. Prakash, L. Adamic, T. Iwashyna, H. Tong, and C. Faloutsos. Fractional immunization in networks. In *Proceedings of the SIAM International Conference on Data Mining*, pages 659–667. SIAM, 2013.

[145] E. Gourdin, J. Omic, and P. Van Mieghem. Optimization of network protection against virus spread. In *Proceedings of the International Workshop on the Design of Reliable Communication Networks (DRCN)*, pages 86–93. IEEE, 2011.

[146] V. M. Preciado, F. D. Sahneh, and C. Scoglio. A convex framework for optimal investment on disease awareness in social networks. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 851–854. IEEE, 2013.

[147] X. Chen and V. M. Preciado. Optimal coinfection control of competitive epidemics in multi-layer networks. In *Proceedings of the IEEE Conference on Decision and Control*, pages 6209–6214. IEEE, 2014.

[148] J. Li, X. Chen, S. Pequito, G. J. Pappas, and V. M. Preciado. Structural target controllability of undirected networks. In *Proceedings of the IEEE Conference on Decision and Control*, pages 6656–6661. IEEE, 2018.

[149] J. Li, X. Chen, S. Pequito, G. J. Pappas, and V. M. Preciado. Resilient structural stabilizability of undirected networks. *arXiv preprint arXiv:1810.00126*, 2018.

[150] X. Chen, S. Pequito, G. J. Pappas, and V. M. Preciado. Minimal edge addition for network controllability. *IEEE Transactions on Control of Network Systems*, 6 (1):312–323, 2019.

[151] X. Chen, S. Chen, and V. M. Preciado. Safety verification of nonlinear autonomous system via occupation measures. *arXiv preprint arXiv:1903.05311*, 2019.

[152] X. Chen, F. Miao, G. J. Pappas, and V. Preciado. Hierarchical data-driven vehicle dispatch and ride-sharing. In *Proceedings of the IEEE Conference on Decision and Control*, pages 4458–4463. IEEE, 2017.

[153] X. Chen, E. Kang, S. Shiraishi, V. M. Preciado, and Z. Jiang. Digital behavioral twins for safe connected cars. In *Proceedings of the 21th ACM/IEEE International Conference on Model Driven Engineering Languages and Systems*, pages 144–153. ACM, 2018.

[154] T. H. Cormen. *Introduction to algorithms*. MIT Press, 2009.

[155] T. Kailath. *Linear systems*, volume 156. Prentice-Hall Englewood Cliffs, NJ, 1980.

[156] H. Federer. *Geometric measure theory*. Springer, 2014.

[157] L.L. Markus and E.B. Lee. On the existence of optimal controls. *ASME. J. Basic Eng.*, 84(1):13–20, 1962.

[158] L. G. Valiant. The complexity of enumeration and reliability problems. *SIAM Journal on Computing*, 8(3):410–421, 1979.

[159] X. Chen. Finding an optimal perturbation configuration. `https://www.mathworks.com/matlabcentral/fileexchange/61536-finding-an-optimal-perturbation-configuration`, 2017.

[160] S. Janson, T. Luczak, and A. Rucinski. *Random graphs*, volume 45. John Wiley & Sons, 2011.

[161] M. L. J. Hautus. (A,B)-invariant and stabilizability subspaces, a frequency domain description. *Automatica*, 16(6):703–707, 1980.

[162] W. A. Wood. Linear multivariable control: A geometric approach. *Springer, New York*, 1986.

[163] B. Awerbuch. A new distributed depth-first-search algorithm. *Information Processing Letters*, 20(3):147–150, 1985.

[164] S. Vinterbo. A note on the hardness of the k-ambiguity problem. Technical report, Technical report, Harvard Medical School, Boston, MA, USA, 2002.

[165] D. S. Hochbaum. Approximating covering and packing problems: set cover, vertex cover, independent set, and related problems. *Approximation Algorithms for NP-Hard Problem*, pages 94–143, 1997.

[166] S. Fujishige. *Submodular functions and optimization*, volume 58. Elsevier, 2005.

[167] E. Chlamtác, M. Dinitz, C. Konrad, G. Kortsarz, and G. Rabanca. The densest k-subhypergraph problem. *SIAM Journal on Discrete Mathematics*, 32(2):1458–1477, 2018.

[168] R. A. Horn and C. R. Johnson. *Matrix analysis.* Cambridge University Press, 1990.

[169] B. Bollobás. *Modern graph theory*, volume 184. Springer Science & Business Media, 2013.

[170] W. Feller. *An introduction to probability theory and its applications*, volume 2. John Wiley & Sons, 2008.

[171] J. Leskovec and A. Krevl. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, 2014.

[172] N. Durak, T. G. Kolda, A. Pinar, and C. Seshadhri. A scalable null model for directed graphs matching all degree distributions: In, out, and reciprocal. In *2013 IEEE 2nd Network Science Workshop (NSW)*, pages 23–30. IEEE, 2013.

[173] F. Chung and F. C. Graham. *Spectral graph theory.* Number 92. American Mathematical Soc., 1997.

[174] C. L. DuBois. UCI network data repository. `http://networkdata.ics.uci.edu`, 2008.

[175] M. Putinar. Positive polynomials on compact semi-algebraic sets. *Indiana University Mathematics Journal*, 42(3):969–984, 1993.

[176] F. B. Hanson. *Applied stochastic processes and control for jump-diffusions: modeling, analysis, and computation*, volume 13. Siam, 2007.

[177] M. W. Hirsch and H. Smith. Monotone dynamical systems. In *Handbook of differential equations: ordinary differential equations*, volume 2, pages 239–357. Elsevier, 2006.

[178] W. W. Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, 33(4):452–473, 1977.

[179] F. Chung, L. Lu, and V. Vu. Eigenvalues of random power law graphs. *Annals of Combinatorics*, 7(1):21–33, 2003.

[180] R. A. Rossi and N. K. Ahmed. The network data repository with interactive graph analytics and visualization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015. URL `http://networkrepository.com`.

[181] J. Hu, M. Prandini, and S. Sastry. Probabilistic safety analysis in three dimensional aircraft flight. In *Proceedings of the IEEE Conference on Decision and Control*, volume 5, pages 5335–5340. IEEE, 2003.

[182] S. Glavaski, A. Papachristodoulou, and K. Ariyur. Safety verification of controlled advanced life support system using barrier certificates. In *International Workshop on Hybrid Systems: Computation and Control*, pages 306–321. Springer, 2005.

[183] J. Ziegler and C. Stiller. Fast collision checking for intelligent vehicle motion planning. In *Intelligent Vehicles Symposium (IV), 2010 IEEE*, pages 518–522. IEEE, 2010.

[184] M. Althoff, O. Stursberg, and M. Buss. Safety assessment of autonomous cars using verification techniques. In *Proceedings of the American Control Conference*, pages 4154–4159. IEEE, 2007.

[185] M. Althoff, O. Stursberg, and M. Buss. Model-based probabilistic collision detection in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 10(2):299–310, 2009.

[186] Alberto Bemporad, Fabio Danilo Torrisi, and Manfred Morari. Optimization-based verification and stability characterization of piecewise affine and hybrid systems. In *International Workshop on Hybrid Systems: Computation and Control*, pages 45–58. Springer, 2000.

[187] Alongkrit Chutinan and Bruce H Krogh. Computational techniques for hybrid system verification. *IEEE Transactions on Automatic Control*, 48(1):64–75, 2003.

[188] B.Bollobás. Volume estimates and rapid mixing. *Flavors of geometry*, 31:151–182, 1997.

[189] M. E. Dyer and A. M. Frieze. On the complexity of computing the volume of a polyhedron. *SIAM Journal on Computing*, 17(5):967–974, 1988.

[190] H. Anai and V. Weispfenning. Reach set computations using real quantifier elimination. In *International Workshop on Hybrid Systems: Computation and Control*, pages 63–76. Springer, 2001.

[191] C. J. Tomlin, I. Mitchell, A. M. Bayen, and M. Oishi. Computational techniques for the verification of hybrid systems. *Proceedings of the IEEE*, 91(7):986–1001, 2003.

[192] E. Asarin, T. Dang, and A. Girard. Reachability analysis of nonlinear systems using conservative approximation. In *International Workshop on Hybrid Systems: Computation and Control*, pages 20–35. Springer, 2003.

[193] S. Prajna and A. Jadbabaie. Safety verification of hybrid systems using barrier certificates. In *International Workshop on Hybrid Systems: Computation and Control*, pages 477–492. Springer, 2004.

[194] S. Prajna. Barrier certificates for nonlinear model validation. *Automatica*, 42(1): 117–126, 2006.

[195] S. Prajna, A. Jadbabaie, and G. J. Pappas. A framework for worst-case and stochastic safety verification using barrier certificates. *IEEE Transactions on Automatic Control*, 52(8):1415–1428, 2007.

[196] C. Sloth, G. J. Pappas, and R. Wisniewski. Compositional safety analysis using barrier certificates. In *Proceedings of the 15th ACM international conference on Hybrid Systems: Computation and Control*, pages 15–24. Citeseer, 2012.

[197] S. Kousik, S. Vaskov, M. Johnson-Roberson, and R. Vasudevan. Safe trajectory synthesis for autonomous driving in unforeseen environments. In *ASME 2017 Dynamic Systems and Control Conference*. American Society of Mechanical Engineers, 2017.

[198] R. Vinter. Convex duality and nonlinear optimal control. *SIAM Journal on Control and Optimization*, 31(2):518–538, 1993.

[199] D. Henrion, J. B. Lasserre, and C. Savorgnan. Approximate volume and integration for basic semialgebraic sets. *SIAM Review*, 51(4):722–743, 2009.

[200] J. B. Lasserre, D. Henrion, C. Prieur, and E. Trélat. Nonlinear optimal control via occupation measures and lmi-relaxations. *SIAM Journal on Control and Optimization*, 47(4):1643–1666, 2008.

[201] D. Henrion and M. Korda. Convex computation of the region of attraction of polynomial control systems. *IEEE Transactions on Automatic Control*, 59(2): 297–312, 2014.

[202] A. Majumdar, R. Vasudevan, M. M. Tobenkin, and R. Tedrake. Convex optimization of nonlinear feedback controllers via occupation measures. *The International Journal of Robotics Research*, 33(9):1209–1230, 2014.

[203] Shizuo Kakutani. Concrete representation of abstract (m)-spaces (a characterization of the space of continuous functions). *Annals of Mathematics*, pages 994–1024, 1941.

[204] Vladimir Igorevich Arnol'd. *Mathematical methods of classical mechanics*, volume 60. Springer Science & Business Media, 2013.

[205] P. Zhao, S. Mohan, and R. Vasudevan. Control synthesis for nonlinear optimal control via convex relaxations. In *Proceedings of the American Control Conference*, pages 2654–2661. IEEE, 2017.

[206] J. B. Lasserre. *An introduction to polynomial and semi-algebraic optimization*, volume 52. Cambridge University Press, 2015.

[207] P. A. Parrilo. *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization.* PhD thesis, California Institute of Technology, 2000.

[208] S. Mohan and R. Vasudevan. Convex computation of the reachable set for hybrid systems with parametric uncertainty. In *Proceedings of the American Control Conference*, pages 5141–5147. IEEE, 2016.

[209] J. Löfberg. Yalmip: A toolbox for modeling and optimization in matlab. In *Proceedings of the CACSD Conference*, volume 3. Taipei, Taiwan, 2004.

[210] ApS Mosek. The mosek optimization toolbox for matlab manual, 2015.

[211] S. Barnett. *Matrices in control theory with applications to linear programming.* Van Nostrand Reinhold, 1971.

[212] M. R. Garey and D. S. Johnson. Computers and intractability: A guide to the theory of NP-completeness (Series of books in the mathematical sciences), ed. *Computers and Intractability*, 340, 1979.

[213] U. Feige. A threshold of ln n for approximating set cover. *Journal of the ACM (JACM)*, 45(4):634–652, 1998.

[214] M. Kirkilionis and S. Walcher. On comparison systems for ordinary differential equations. *Journal of mathematical analysis and applications*, 299(1):157–173, 2004.

[215] C. D. Meyer. *Matrix analysis and applied linear algebra*, volume 2. Siam, 2000.

[216] E. J. Anderson and P. Nash. *Linear programming in infinite-dimensional spaces: theory and applications.* John Wiley & Sons, 1987.

[217] David G Luenberger. *Optimization by vector space methods.* John Wiley & Sons, 1997.