

Universidade Federal de Minas Gerais  
Escola de Engenharia  
Programa de Pós-Graduação em Engenharia Elétrica

Arthur Noronha Montanari

**OBSERVABILITY OF DYNAMICAL NETWORKS**

Belo Horizonte

2021



**Universidade Federal de Minas Gerais**

**Escola de Engenharia**

**Programa de Pós-Graduação em Engenharia Elétrica**

**OBSERVABILITY OF DYNAMICAL NETWORKS**

Arthur Noronha Montanari

Tese de Doutorado submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Engenharia Elétrica da Escola de Engenharia da Universidade Federal de Minas Gerais, como requisito para obtenção do Título de Doutor em Engenharia Elétrica.

Orientador: Prof. Luis Antonio Aguirre

Belo Horizonte - MG

Fevereiro de 2021

M764o

Montanari, Arthur Noronha.  
Observability of dynamical networks [recurso eletrônico] / Arthur  
Noronha Montanari. - 2021.  
1 recurso online (xxi, 127 f. : il., color.) : pdf.

Orientador: Luis Antonio Aguirre.

Tese (doutorado) - Universidade Federal de Minas Gerais,  
Escola de Engenharia.

Apêndices: f. 109-127.  
Bibliografia: f. 95-108.

Exigências do sistema: Adobe Acrobat Reader.

1. Engenharia elétrica - Teses. 2. Controle automático - Teses.  
3. Redes elétricas - Teses. 4. Detectores - Teses. I. Aguirre, Luis Antônio.  
II. Universidade Federal de Minas Gerais. Escola de Engenharia. III.  
Título.

CDU: 621.3(043)



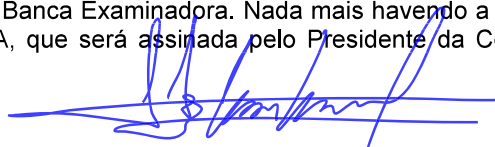
UNIVERSIDADE FEDERAL DE MINAS GERAIS  
ESCOLA DE ENGENHARIA  
*Programa de Pós-Graduação em Engenharia Elétrica*

**ATA DA 355ª DEFESA DE TESE DE DOUTORADO  
DO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA**

ATA DE DEFESA DE TESE DE DOUTORADO do aluno **Arthur Noronha Montanari** - registro de matrícula de número 2018751209. Às 14:30 horas do dia 26 do mês de fevereiro de 2021, reuniu-se de forma virtual a Comissão Examinadora da TESE DE DOUTORADO para julgar, em exame final, o trabalho intitulado "**Observability of Dynamical Networks**" da Área de Concentração em Sinais e Sistemas. O Prof. Luis Antonio Aguirre, orientador do aluno, abriu a sessão apresentando os membros da Comissão e, dando continuidade aos trabalhos, informou aos presentes que, de acordo com o Regulamento do Programa no seu Art. 8.16, será considerado APROVADO na defesa da Tese de Doutorado o candidato que obtiver a aprovação unânime dos membros da Comissão Examinadora. Em seguida deu início à apresentação do trabalho pelo Candidato. Ao final da apresentação seguiu-se a arguição do candidato pelos examinadores. Logo após o término da arguição a Comissão Examinadora se reuniu, sem a presença do Candidato e do público, e elegeu o Prof. Luis Antonio Aguirre para presidir a fase de avaliação do trabalho, constituída de deliberação individual de APROVAÇÃO ou de REPROVAÇÃO e expedição do resultado final. As deliberações individuais de cada membro da Comissão Examinadora foram as seguintes:

Membro da Comissão Examinadora	Instituição de Origem	Deliberação	Assinatura
Prof. Dr. Luis Antonio Aguirre - Orientador	DELT (UFMG)	Aprovado	
Prof. Dr. Elbert Einstein Nehrer Macau	(UNIFESP)	Aprovado	
Prof. Dr. Adilson E. Motter	Department of Physics and Astronomy, (Northwestern University)	Aprovado	
Prof. Dr. Erivelton Geraldo Nepomuceno	Engenharia Elétrica (UFSJ)	Aprovado	
Prof. Dr. Leonardo Antônio Borges Tôres	DELT (UFMG)	Aprovado	

Tendo como base as deliberações dos membros da Comissão Examinadora a Tese de Doutorado foi Aprovada. O resultado final de aprovação foi comunicado publicamente ao Candidato pelo Presidente da Comissão, ressaltando que a obtenção do Grau de Doutor em ENGENHARIA ELÉTRICA fica condicionada à entrega do TEXTO FINAL da Tese de Doutorado. O Candidato terá um prazo máximo de 30 (trinta) dias, a partir desta data, para fazer as CORREÇÕES DE FORMA e entregar o texto final da Tese de Doutorado na secretaria do PPGEE/UFMG. As correções de forma exigidas pelos membros da Comissão Examinadora deverão ser registradas em um exemplar do texto da Tese de Doutorado, cuja verificação ficará sob a responsabilidade do Presidente da Banca Examinadora. Nada mais havendo a tratar o Presidente encerrou a reunião e lavrou a presente ATA, que será assinada pelo Presidente da Comissão Examinadora. Belo Horizonte, 26 de fevereiro de 2021.



ASSINATURA DO PRESIDENTE DA COMISSÃO EXAMINADORA



This is for you, Camila, Mom and Dad.





“You see, but you do not observe. The distinction is clear.”

– *Sherlock Holmes*



## Agradecimentos

Ao meu orientador, Luis Aguirre, pelo acolhimento, ensinamentos e investimento em meu crescimento pessoal durante o desenvolvimento deste trabalho. Obrigado pela confiança depositada, sem a qual toda e qualquer liberdade criativa não seria a mesma. Obrigado também pelo investimento em meu futuro e por me ajudar a trilhar o caminho correto e ético para a minha carreira.

Ao Adilson, por ter me acolhido em seu grupo na Northwestern University, com o qual tive o prazer de trabalhar em projetos tão diversos e empolgantes. Obrigado por ter me aberto a cabeça para novas direções, tendo uma enorme influência na forma como eu seleciono meus problemas científicos e os comunico.

Aos instrutores que me orientaram durante esta jornada acadêmica. Em especial, ao professor Leo Torres, por nossas diversas conversas acadêmicas e apoio constante em minha carreira; e ao meu antigo tutor Everthon, por ter me ajudado a encontrar este caminho no qual todo dia de trabalho é também um dia de viver.

À minha esposa, Camila, meu bem, por tudo. Amor, suporte, encorajamento, e principalmente confiança de que juntos nós vamos sempre mais longe. Obrigado por me abraçar nos dias felizes, me empurrar nos dias longos e me levantar nos dias difíceis. A vida sem seu sorriso do lado não seria a mesma coisa, e sem ela a gente não teria essa perfeição chamada Pollo que nos acompanha em todas as jornadas (deitando mais frequentemente do que eu gostaria em meu teclado durante longos períodos de redação e programação).

Aos meus pais, Mara e Beto, pelo amor e apoio incondicional. Vocês me trouxeram essa vida maravilhosa e graças a vocês cresço de forma mais preparada, forte e consciente. Espero que encontrem em mim o orgulho que procuram. Obrigado por todos os recursos que me forneceram ao longo destes anos para que fosse possível chegar onde cheguei. Agradeço também à (enorme) família que nos acompanha e à minha segunda família que me acolheu já há muitos anos. Obrigado, Rosana, Ernani e Bi.

Aos meus amigos do CPH: Antônio, Petrus, Leo, Leandro, Ercílio, Felipe e João, pelo belo ambiente de trabalho, amizade e reuniões de cafezinho que propiciaram tantas discussões únicas, sejam produtivas ou completamente inúteis.

Aos vários amigos que me acompanharam desde antes de toda essa jornada. Sejam os amigos da época do Colégio Santo Antônio, os amigos da época de faculdade no CEFET, os amigos “da Internet” (Discord), e os amigos que encontrei durante o doutorado-sanduíche. Saibam que minha casa – onde quer que seja e de onde quer que vocês venham – sempre estará aberta para todos vocês.

À UFMG, seus professores e funcionários, por disponibilizarem a infraestrutura e conhecimento necessário para a realização deste trabalho. À CAPES pelo suporte financeiro ao longo de meu programa de Doutorado e Doutorado-Sanduíche. E aos pagadores de impostos, que contribuíram, contribuem e contribuirão para os mecanismos de educação e pesquisa que me subsidiaram nos últimos quatro anos e que tornaram possível minha estadia em Chicago.

## **Acknowledgements**

To the members of Motter’s group, at Northwestern University, in special, Chao and Thomas, with whom I had the pleasure to work together.

## Abstract

A quantitative understanding and precise control of a complex dynamical system, such as natural, social and technological networks, can only be achieved with the ability to observe its internal states either by direct measurement or indirect estimation. For a large-scale dynamical network, however, it is extremely difficult or physically impossible to place enough sensors to make the system fully observable. The problem of determining whether a system is observable has been well addressed by control engineers and, in a high-dimensional context, network scientists in the recent decade. Nevertheless, even if the system is theoretically observable, the high-dimensionality of the network poses fundamental limits on the computational tractability and performance of a full-state observer. To overcome the curse of dimensionality, and noting the fact that often only a small number of state variables in a network are essential for control, intervention, and monitoring purposes, we instead ask the system to be functionally observable, i.e., that only a targeted subset of system states be reconstructable from the available measurements. In this manuscript, we develop a graph-based theory of functional observability, which leads to highly scalable algorithms to determine minimal necessary sensors and to design the corresponding state observer with minimal order. Compared with the full-state observer, the developed functional observer achieves the same estimation quality with much less sensory and computational resources, making it applicable to large-scale networks. We apply the proposed methods to the detection of cyber-attacks in power grids under limited measurement units and the inference of the infected population during a pandemic under limited testing resources. The applications and numerical results show that the functional observer can significantly scale up our ability to explore otherwise hidden dynamical processes on large-scale complex networks.

**Keywords:** Observability, dynamical networks, structural systems, sensor placement, observer design.



## Resumo

A compreensão quantitativa e controle preciso de um sistema dinâmico complexo, como redes naturais, sociais e tecnológicas, podem ser alcançadas apenas com a habilidade de observar seus estados internos, seja por meio de medições diretas ou estimação indireta. No caso de uma rede dinâmica de larga-escala, entretanto, é extremamente difícil ou fisicamente impossível alocar um número suficiente de sensores para tornar um sistema completamente observável. O problema de determinar se um sistema é observável foi intensivamente estudado por engenheiros de controle e, no contexto de alta-dimensionalidade, cientistas de redes na década recente. Não obstante, mesmo se um sistema for teoricamente observável, a alta-dimensionalidade de redes apresenta limites fundamentais à tratabilidade computacional e desempenho de um observador de estados completo. Com o objetivo de superar a maldição da dimensionalidade, e notando o fato que usualmente apenas um pequeno subconjunto das variáveis de estado em uma rede são essenciais para propósitos de controle, intervenção e monitoramento, investiga-se nesta tese as condições para que um sistema seja observável funcional, isto é, que apenas um subconjunto alvo dos estados do sistema sejam reconstrutíveis a partir das medições disponíveis. Neste manuscrito, desenvolve-se uma teoria baseada em grafos da propriedade de observabilidade funcional, que permite o desenvolvimento de algoritmos altamente escaláveis para a determinação do conjunto mínimo necessário de sensores e o projeto de um observador de estados funcional de mínima ordem. Comparado ao observador de estados completo, o observador de estados funcional apresenta a mesma qualidade na estimação de estados com muito menos recursos sensoriais e computacionais, tornando-o adequado à aplicações em redes de larga-escala. Os métodos propostos são aplicados na detecção de ataques cibernéticos em redes de potência sob um limitado número de unidades de medição, e na inferência das populações infectadas durante uma epidemia sob capacidade limitada de testes. As aplicações e resultados numéricos mostram que o observador de estados funcional pode aumentar significativamente nossa habilidade de explorar processos dinâmicos ocultos em redes complexas de larga-escala.

**Palavras-chave:** Observabilidade, redes dinâmicas, sistemas estruturais, alocação de sensores, projeto de observador.





# Table of contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Context and Motivation . . . . .	1
1.2	Contributions . . . . .	3
1.3	List of Publications . . . . .	4
1.4	Outline . . . . .	5
<b>2</b>	<b>Dynamical Networks Modeling</b>	<b>7</b>
2.1	Dynamical Systems Notation . . . . .	7
2.2	Graph Theory . . . . .	8
2.3	Representation of Dynamical Networks . . . . .	11
2.3.1	Graph representation . . . . .	11
2.3.2	State-space representation . . . . .	13
2.3.3	Examples . . . . .	14
<b>3</b>	<b>A Review on Observability of Network Systems</b>	<b>19</b>
3.1	Structural Observability . . . . .	21
3.1.1	Linear dynamical systems . . . . .	21
3.1.2	Nonlinear dynamical systems . . . . .	24
3.2	Dynamical Observability . . . . .	27
3.2.1	Linear dynamical systems . . . . .	27
3.2.2	Nonlinear dynamical systems . . . . .	29
3.3	Topological Observability . . . . .	30
3.3.1	Linear dynamical systems . . . . .	31
3.3.2	Maximum matching algorithm . . . . .	32
3.3.3	Nonlinear dynamical systems . . . . .	35
3.4	Analysis of Related Works . . . . .	38
3.5	Future Research Directions on the Dynamical Observability of Network Systems . . . . .	43

3.6	Application Examples . . . . .	47
3.6.1	Power grids . . . . .	48
3.6.2	Multi-agent consensus . . . . .	52
3.7	Final Considerations . . . . .	57
<b>4</b>	<b>Functional observability and target state estimation in large-scale networks</b>	<b>59</b>
4.1	Background on Functional Observability . . . . .	61
4.2	Structural Functional Observability . . . . .	63
4.3	Methods . . . . .	66
4.3.1	Minimum sensor placement for sets of target nodes . . . . .	66
4.3.2	Minimum order functional observer design . . . . .	68
4.4	Numerical Results in Large-Scale Complex Networks . . . . .	74
4.4.1	Minimum sensor placement . . . . .	75
4.4.2	Minimum order functional observer design . . . . .	77
4.4.3	Performance comparison between observers . . . . .	80
4.5	Applications . . . . .	82
4.5.1	Cyber-attacks detection in power grids . . . . .	82
4.5.2	Estimation of epidemic spreading in target populations . . . . .	86
4.6	Conclusion . . . . .	90
<b>5</b>	<b>Conclusion</b>	<b>93</b>
	<b>References</b>	<b>95</b>
	<b>Appendix A Sensor Placement Algorithm</b>	<b>109</b>
A.1	Minimization of Coefficient of Observability . . . . .	109
A.2	Minimization of Functional Observer Order . . . . .	110
	<b>Appendix B State Observer Design</b>	<b>111</b>
B.1	Luenberger Observer . . . . .	111
B.2	Functional Observer . . . . .	113
B.3	Functional Observer for Nonlinear Systems . . . . .	116
	<b>Appendix C Proof of Structural Functional Observability</b>	<b>119</b>
	<b>Appendix D Related Works on “Target Observability”</b>	<b>125</b>

# List of figures

2.1	Graph examples. . . . .	9
2.2	SCC and root SCC of a digraph. . . . .	10
2.3	Differences between graphs of a nodal dynamical system, a network topology and a full network. . . . .	12
2.4	Nonlinear graph of the Jacobian matrix of the Rössler system. . . . .	16
2.5	Full network examples of Rössler systems, with $m = 3$ , coupled by different variables. . . . .	17
3.1	Graph representation of (3.21). . . . .	33
3.2	Maximum matching of simple networks. . . . .	35
3.3	Root SCC of a Rössler system graph. . . . .	37
3.4	Tank system and corresponding network representation. . . . .	41
3.5	Dynamics of the IEEE power grid benchmark. . . . .	49
3.6	Coefficient of observability $\delta_y$ per number of PMUs. . . . .	51
3.7	Signaling network $W$ of a flock of $m = 150$ agents moving in a 2-dimensional space in two different scenarios. . . . .	55
3.8	Proportion of the minimum number of sensor nodes $ \mathcal{S} /m$ as a function of the signaling network symmetry $\rho_{\text{sym}}(W)$ for the two different scenarios. . . . .	56
4.1	Structural functional observability of dynamical systems. . . . .	64
4.2	Illustrative example of Algorithm 2. . . . .	73
4.3	Minimum sensor placement in large-scale networks. . . . .	77
4.4	Minimum order functional observer design in large-scale networks. . . . .	78
4.5	Performance of functional observers for target state estimation in large-scale networks. . . . .	81
4.6	Target state estimation for cyber-attack detection in power grids. . . . .	83
4.7	Target state estimation in epidemics. . . . .	87



# List of abbreviations

LHS	Left-hand side
ODE	Ordinary differential equation
PMU	Phasor measurement unit
RHS	Right-hand side
RMSE	Root-mean-square error
SCADA	Supervisor control and data acquisition
SCC	Strongly connected components
SIRD	Susceptible-infected-recovered-dead individuals
SF	Scale-free network
SVD	Singular value decomposition
SW	Small-world network



# List of symbols

## Notation

$a$	Scalar.
$\mathbf{a}$	Vector.
$A$	Matrix.
$\mathcal{A}$	Set.
$a(\boldsymbol{\alpha})$	Denotes dependence on $\boldsymbol{\alpha}$ .
$\hat{\mathbf{a}}$	Estimate of $\mathbf{a}$ .
$\mathbb{R}$	Set of real numbers.
$\mathbb{C}$	Set of complex numbers.
$\emptyset$	Empty set.

## Operators

$\odot$	Hadamard product operator (element-wise product).
$\otimes$	Kronecker product operator (direct product).
$\oplus$	Direct sum operator.
$A^\dagger$	Moore-Penrose inverse of $A$ .

## Symbols.

$\mathbf{x}$	State variables. Vector of dimension $\mathbb{R}^n$ .
$\mathbf{y}$	Output variables. Vector of dimension $\mathbb{R}^q$ .
$\mathbf{u}$	Input variables. Vector of dimension $\mathbb{R}^p$ .
$\mathbf{z}$	Target variables desired to be estimated. Vector of dimension $\mathbb{R}^r$ .
$t$	Continuous time instant.
$A, B, C, D$	Dynamic, input, output and feedforward matrix of a dynamical system.
$F$	Functional matrix.
$\mathbf{f}, \mathbf{h}$	Nonlinear functions of the dynamical model.
$\mathcal{G}$	Graph.
$\mathcal{V}, \mathcal{E}$	Set of nodes and edges.
$A_{\text{adj}}, L$	Adjacency and Laplacian matrix.
$\mathcal{X}, \mathcal{S}, \mathcal{D}, \mathcal{T}$	Set of state nodes, sensor nodes, driver nodes and target nodes.
$\lambda_i$	Eigenvalues sorted from smallest to largest values.





# Chapter 1

## Introduction

### 1.1 Context and Motivation

The mathematical modeling of *dynamical systems* is a fundamental framework in engineering that provides a means to analyze aspects of a system, such as its stability, controllability or observability, and thereafter design control laws for practical applications (Chen, 1999; Khalil, 2002). However, being designed for the most part with systems of low-dimensional order in mind, classic control theory methods are not efficient, or even feasible, for large-scale systems, such as interconnected (networked) dynamical systems. This practical limitation has led control theory notions to be adapted, optimized, or even redefined, in the literature for high-dimensional applications (Chen, 2014).

A specific, but recurrent, type of high-dimensional system can be defined as *networks*. A network is a set of nodes interconnected by edges, in which information flows among its elements through pairwise interactions. It can be mathematically modeled by graph structures, which allow a wide range of useful metrics and algorithms of graph theory (Bullo, 2016; Chen et al., 2013; Newman, 2010). For instance, graph theory can be used to assess the robustness to spreading failures in power systems (Schäfer et al., 2018; Zhang et al., 2014) or biological networks (Gilarranz et al., 2017; Schimit and Monteiro, 2009).

Up to the end of the twentieth century, it was believed that real-world interconnected systems, such as neuronal, social, communication, traffic, and energy networks, and even the Internet, were composed of stochastic connections among its nodes. However, works over the last two decades highlighted that most of real-world networks share similar topological characteristics—not being purely random, nor purely regular (Barabási,

1999, 2009; Watts and Strogatz, 1998). *Complex networks*, therefore, are a subclass of mathematical models derived from graph theory, in which topological structures (graphs) show recurrent patterns that are found in the most diverse real networks present in nature and engineering (Barabási and Pósfasi, 2016; Chen et al., 2013). Based on these findings, the last years have been flooded with studies about complex network models, such as *scale-free networks* (Barabási, 1999) and *small-world networks* (Watts and Strogatz, 1998).

The study of complex networks is essential to increase the knowledge about the structural characteristics and recurrent patterns that govern real networks—even when nodal dynamics are disregarded in favor of a higher focus on the graph properties. It is a first step in a long ladder whose final goal is to develop control techniques for *dynamical complex networks* (Wang and Chen, 2003). *Dynamical networks* are defined by a set of dynamical systems that, when analyzed individually, describe relatively simpler behaviours, but, when interconnected, develop interactions that considerably raise the system complexity (Monteiro, 2014). This is a consequence of a twofold interaction between local properties (nodal dynamics) and global properties, such as the network structure or topology (Aguirre et al., 2018).

Many mathematical models were expanded to include complex networks that describe the spatial relations and interactions between their elements. Among the numerous examples are: models of infectious diseases (Moreno et al., 2002; Schimit and Monteiro, 2009), reaction-diffusion systems (Wolfrum, 2012) (e.g. predator-prey models (Nakao and Mikhailov, 2010)), and Boolean systems (Gates and Rocha, 2015). In the field of nonlinear dynamics, the study of synchronization in networks of oscillators stands out (Arenas et al., 2008; Boccaletti et al., 2002; Montanari et al., 2019; Rodrigues et al., 2016), with important applications in power systems (Dorfler et al., 2013; Montanari et al., 2020) and biological networks (Hammond et al., 2007). In this case, each node is composed of an individual dynamical system, a nonlinear oscillator, and its interactions are determined by coupling functions of the state variables of different oscillators. Usually, the main goal is to determine under which conditions the synchronization manifold of a dynamical network of oscillators becomes stable. These conditions can be related to the network structure (Moreno and Pacheco, 2004; Wang and Chen, 2002a), the coupling method (Stankovski et al., 2017), or the nonlinear oscillator model—from the well-studied Kuramoto phase oscillator (Dorfler and Bullo, 2014; Kuramoto, 1975) to chaotic ones (Boccaletti et al., 2002; Eroglu et al., 2017).

It is only natural that as our understanding of complex behaviors, such as synchronization, in network systems increases, the next step is to question how to control such

systems according to our needs. Indeed, “*the ultimate proof of our understanding of natural or technological systems is reflected in our ability to control them*”, as stated by Liu et al. (2011b). One fundamental mechanism that enables the precise control of such dynamical system is *feedback*, which involves sensors, signals and actuators in a closed loop (Wiener, 2019). In large-scale interconnected systems, however, it is not always possible to rely on direct measurements provided by sensors. For instance, it is not always economically viable to place a sophisticated phasor measurement unit (PMU) at every substation in a power grid, as well as it is not physically possible to measure each one of the tens of billions of neurons in our brain. Thus, indirect estimation of unmeasured variables is essential for the control of large-scale dynamical networks.

*Observability* is a key property that determines if the trajectory temporal evolution of the internal states of a dynamical system can be reconstructed based on knowledge of the inputs and outputs, as introduced by Kalman (1959). It can be formulated as condition for the optimal placement of sensors in a network (Haber et al., 2018; Liu et al., 2013; Montanari and Aguirre, 2020) as well as the design of stable state estimators (Luenberger, 1966; Montanari and Aguirre, 2019). However, as network systems grow large, high-dimensionality poses a fundamental obstacle that hampers the direct use of traditional methods developed in control theory (Chen, 2014; Motter, 2015), calling for different approaches to overcome the curse of dimensionality (Montanari and Aguirre, 2020). In the past decade, a different definition of observability, grounded on graph theory, known as *structural observability* (Lin, 1974), has opened a new branch to novel developments and highly intuitive techniques in this field, which although not yet consolidated, have achieved great scalability to high-dimensional systems (Liu and Barabási, 2016; Liu et al., 2011b). This work expands on this field of research by exploring the advantages and disadvantages of recent results and further proposing a novel approach to state estimation in large systems by introducing a generalization of a property from control theory known as *functional observability*.

## 1.2 Contributions

The main goal of this work is to investigate the interplay between the observability property of a dynamical network and its corresponding nodal dynamics, coupling methods and network structure. The contributions of this manuscript are threefold:

Firstly, we thoroughly review the fundamental properties of observability—and by duality controllability—of low-dimensional dynamical systems. We explore the importance of using not only a *crisp* (yes or no) classification of observability, but

also one that gradually quantifies *how good* a system observability really is under a specific set of measures. These notions are extended to a network context, where different metrics need to be developed due to high-dimensionality issues. An extensive review and criticism is developed regarding observability metrics based on the network topology properties as represented by its the adjacency matrix. We conclude the review with some interesting guidelines of research in observability and controllability of network systems.

Secondly, following the developed review on observability in network systems, it becomes clear that the concept of full observability might not be the most suitable approach to dynamical networks due to the high-dimensionality of the problem. To circumvent this, the concept of functional observability is revisited from control theory in a network context, allowing one to focus on a specific subset of nodes desired to be estimated (observed) rather than on the whole set of nodes of a network. We generalize the concept of functional observability to the context of structural functional observability, allowing us to rigorously establish graph-theoretic conditions for the functional observability equivalent to the original rank-based conditions.

Thirdly, based on the proposed theory, we design two highly-scalable algorithms to solve the optimal sensor placement and functional observer design problems in the context of structural networks. The first algorithm determines the minimal set of sensors placed on a dynamical network required to ensure the functional observability of a system with respect to a given set of target nodes, while the second algorithm—after the sensors are placed—designs a minimum-order functional observer whose output converges asymptotically to the target states, thus achieving accurate estimation. Numerical results are shown for both algorithms in the context of large-scale complex networks and applications in power grids and epidemics.

### 1.3 List of Publications

The following manuscripts were published, or are in final stages of editing, during the course of this work:

- **Arthur N. Montanari**, Chao Duan, Luis A. Aguirre, Adilson E. Motter. Functional observability and target state estimation of large-scale networks. *In progress* (2021).
- **Arthur N. Montanari**, Luis A. Aguirre. Observability of Network Systems: A Critical Review of Recent Results. *Journal of Control, Automation and Electrical Systems*, 31:1348-1374 (2020).

- **Arthur N. Montanari**, Ercilio I. Moreira, Luis A. Aguirre. Effects of network heterogeneity and tripping time on the basin stability of power systems. *Communications in Nonlinear Science and Numerical Simulation*, 89:105296 (2020).
- Leonardo L. Portes, **Arthur N. Montanari**, Debora Correa, Michael Small, Luis A. Aguirre. The reliability of recurrence network analysis is influenced by the observability properties of the recorded time series. *Chaos*, 29:083101 (2019).
- **Arthur N. Montanari**, Leandro Freitas, Leonardo A. B. Torres, Luis A. Aguirre. Phase synchronization analysis of bridge oscillators between clustered networks. *Nonlinear Dynamics*, 97:2399-2411 (2019).
- **Arthur N. Montanari**, Luis A. Aguirre. Particle filtering of dynamical networks: Highlighting observability issues. *Chaos*, 29:033118 (2019).

Moreover, the following works were presented in conferences:

- Leonardo L. Portes, **Arthur N. Montanari**, Debora Correa, Michael Small, Luis A. Aguirre. Reliability of Recurrence Network Analysis against the observability properties of the recorded time series. *8th International Symposium on Recurrence Plots*, Zhenjiang, China (2019).
- **Arthur N. Montanari**, Luis A. Aguirre. On functional observability, sensor allocation and dynamical networks. At the *2nd Latin American Conference on Complex Networks*, Cartagena, Colombia (2019).
- **Arthur N. Montanari**, Ercilio I. Moreira, Luis A. Aguirre. Relation of basin stability of perturbed power systems on network heterogeneity and tripping time. At the *2nd Latin American Conference on Complex Networks*, Cartagena, Colombia (2019).

## 1.4 Outline

This manuscript is subdivided in five chapters. Chapter 2 presents the used theoretical foundation and notation of system theory and graph theory applied to network systems modeling, while Chapter 3 thoroughly reviews observability of dynamical and network systems. Following the previous discussion, Chapter 4 presents the main results of this work, generalizing the concept of functional observability to structural networks as well as providing application examples in power grids and multi-group epidemiological models. Finally, Chapter 5 concludes this work with some future work proposals and final considerations.



# Chapter 2

## Dynamical Networks Modeling

A *dynamical network* can be studied at three levels: i) the *node dynamics*, described by a dynamical system; ii) the *network topology*, described by a graph; and iii) the *full network*, a combination of both aforementioned levels (Aguirre et al., 2018). The interconnection among independent and comparatively simpler dynamical systems in a network unravels different kinds of interactions that considerably raises the system complexity (Monteiro, 2014). Thus, to investigate a full network, three components must be considered: i) the graph, which describes the interconnection structure along the network; ii) the coupling method, which describes how these interconnections unfold; and iii) the node dynamics, which describes the behaviour and interactions of a node when isolated from its neighbourhood.

This chapter is organized as follows. Section 2.1 formalizes the adopted notation for dynamical systems representation in this work. Section 2.2 reviews fundamental properties and metrics of graph theory. Section 2.3 mathematically defines a dynamical network based on the three aforementioned levels of definition.

### 2.1 Dynamical Systems Notation

The state-space representation of a linear continuous time-invariant dynamical system is given by

$$\begin{cases} \dot{\mathbf{x}} = A\mathbf{x} + B\mathbf{u}, \\ \mathbf{y} = C\mathbf{x} + D\mathbf{u}, \end{cases} \quad (2.1)$$

where  $\mathbf{x} \in \mathbb{R}^n$  is the state vector,  $\mathbf{u} \in \mathbb{R}^p$  is the input (control) vector,  $\mathbf{y} \in \mathbb{R}^q$  is the output (measurement) vector, and  $(A, B, C, D)$  are matrices of consistent dimensions

known as, respectively, the dynamic matrix, input (control) matrix, output (measurement) matrix and feedforward matrix. Time  $t$  dependence is omitted for compactness of notation, only for the system variables  $\mathbf{x}$ ,  $\mathbf{u}$  and  $\mathbf{y}$ . Vectors are defined as column vectors, denoted by bold lower-case letters, and matrices by upper-case letters.

For an autonomous nonlinear continuous time-invariant dynamical system, the state-space representation is:

$$\begin{cases} \dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}), \\ \mathbf{y} = \mathbf{h}(\mathbf{x}), \end{cases} \quad (2.2)$$

where  $\mathbf{f} : \mathcal{M} \mapsto \mathcal{M}$  and  $\mathbf{h} : \mathcal{M} \mapsto \mathbb{R}^q$  are nonlinear functions, and  $\mathbf{x} \in \mathcal{M} \subseteq \mathbb{R}^n$ . It is assumed that the reader is familiar with linear and nonlinear system theory. For more details, the reader is referred to (Chen, 1999) for linear systems theory, and to (Khalil, 2002; Vidyasagar, 1978) for fundamentals of nonlinear systems.

## 2.2 Graph Theory

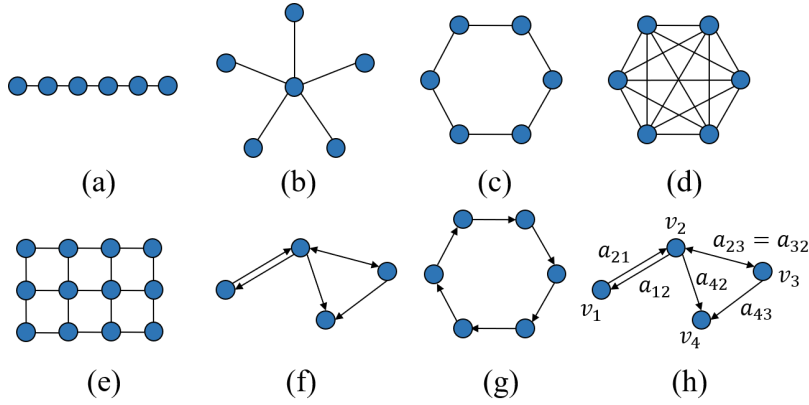
Graph theory provides mathematical definitions, properties and metrics for analysis and design of network systems and even algorithms. This section presents key aspects of graph theory applied throughout the work.

A *graph* is defined as  $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ , where  $\mathcal{V} = \{v_1, v_2, \dots, v_m\}$  and  $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V} = \{e_1, e_2, \dots, e_{\bar{m}}\}$  are finite sets of  $m$  nodes and  $\bar{m}$  edges, respectively. The *cardinality* of a set, denoted by  $|\mathcal{V}|$ , is the number of elements of the set. The *adjacency matrix*  $A_{\text{adj}} = [a_{ij}]$  is a representation that associates elements (edges) of  $\mathcal{E}$  to a pair of elements (nodes) of  $\mathcal{V}$ . In what follows, the notation  $A_{\text{adj}} = [a_{ij}]$  is used to denote that  $a_{ij}$  is an entry of  $A_{\text{adj}}$ .

The following conventions and properties of graph theory are used throughout this work. For more details, we refer the reader to (Bullo, 2016; Chen et al., 2013; Newman, 2010). Figure 2.1 illustrates the following types of graph and exemplifies some usual graph structures in the literature:

- *Undirected and directed graphs.* If  $(v_i, v_j)$  are undirectedly linked, then  $a_{ij} = a_{ji} \neq 0$ , and  $A_{\text{adj}}$  is a symmetric matrix. If  $(v_i, v_j)$  are not linked, then  $a_{ij} = 0$ . This adjacency matrix is denoted as *undirected*. If it is a *directed* graph, or *digraph* for short, then  $a_{ij}$  corresponds to an edge connecting node  $v_j$  to node  $v_i$  (Newman, 2010). If  $a_{ii} \neq 0$ , then node  $i$  has an edge connecting to itself, denoted as *self-edge*.



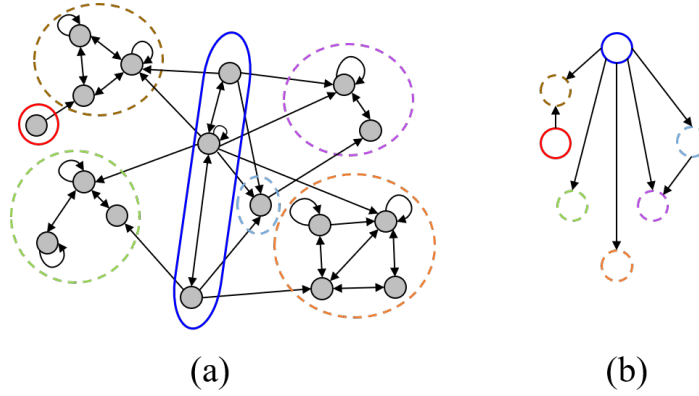


**Figure 2.1:** Graph examples. (a) Chain or path graph. (b) Star graph. (c) Ring or cycle graph. (d) Complete or fully connected graph. (e) Cartesian grid graph. (f) Digraph. (g) Cycle digraph. (h) Weighted graph.

- *Binary and weighted graphs.* If  $A_{\text{adj}} \in \{0, 1\}^{m \times m}$ , then the graph is *binary* or *unweighted*. And the graph is *weighted* if  $A_{\text{adj}} \in (-\infty, \infty)^{m \times m}$ .
- *Paths.* A *path* is an ordered sequence of nodes, interconnected by direct edges (if it is a digraph), between a given pair of nodes. A *simple path* has no repeated node in its sequence, except possibly for the initial and final node.
- *Cycle.* A *cycle* is a simple path where the final node equals the initial one, and it has at least 3 nodes. Otherwise, the graph is *acyclic*.
- *Connected.* A graph is *connected* if there exists a path between any pair of nodes.
- *Subgraph.* A digraph  $\mathcal{G}' = \{\mathcal{V}', \mathcal{E}'\}$  is a subgraph of  $\mathcal{G}$  if  $\mathcal{V}' \subseteq \mathcal{V}$  and  $\mathcal{E}' \subseteq \mathcal{E}$ .
- *Line graph.* A line graph  $L(\mathcal{G})$  is a graph such that each node of  $L(\mathcal{G})$  represents an edge of  $\mathcal{G}$ , and two nodes of  $L(\mathcal{G})$  are adjacent if and only if their corresponding edges are incident in  $\mathcal{G}$  (share a common endpoint). In other words, edges of  $\mathcal{G}$  become nodes of  $L(\mathcal{G})$  and vice-versa.

In a dynamical system context, the Laplacian matrix and the connectivity properties of a graph are very useful in the state-space representation of networks of diffusively coupled oscillators (Bullo, 2016). These concepts are reviewed in what follows:

- $\mathcal{G}$  is *strongly connected* if there exists a directed path between any pair of nodes;
- $\mathcal{G}$  is *weakly connected* if the undirected version of a digraph is connected;
- A *globally reachable node* is a node that can be reached from any node by a direct path; and



**Figure 2.2:** (a) SCC (in dashed circles) and root SCC (in solid circle) of a digraph. (b) Condensation graph of (a). Adapted from (Bullo, 2016).

- A *directed spanning tree* is a subgraph where a node is the root of directed paths to all other nodes.

A particular definition of interest is the *strongly connected components* (SCC). A subgraph  $\mathcal{G}'$  is a SCC if  $\mathcal{G}'$  is strongly connected and any subgraph of  $\mathcal{G}$  strictly containing  $\mathcal{G}'$  is not strongly connected. A *root SCC* is a SCC with no incoming edges. A *condensation digraph*  $C(\mathcal{G})$ , in turn, is defined as a graph whose nodes are a SCC of  $\mathcal{G}$ , and there exists a directed edge from a node formed by  $\mathcal{G}'_1$  to a node formed by  $\mathcal{G}'_2$  if there exists a node from  $\mathcal{G}'_1$  connected to a node from  $\mathcal{G}'_2$ . Figure 2.2 illustrates these concepts.

The Laplacian matrix  $L = [l_{ij}]$  is defined as follows:

$$L = D_{\text{diag}} - A_{\text{adj}}, \quad (2.3)$$

where  $D_{\text{diag}} = \text{diag}(k_{1,\text{in}}, \dots, k_{m,\text{in}})$  is called the degree matrix and  $k_{i,\text{in}} = \sum_{j=1}^m a_{ij}$  is the in-degree of node  $v_i$ . Some useful properties of the Laplacian matrix are (Boccaletti et al., 2006):

- $L$  is always symmetric and positive semidefinite.
- Given that  $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_m$  are eigenvalues of  $L$ , if  $\lambda_2 > 0$  ( $\lambda_1 = 0$ ), then the network is connected.
- The number of connected components in  $\mathcal{G}$  is the dimension of the nullspace of the Laplacian matrix and the algebraic multiplicity of the zero eigenvalue.
- $\text{trace}(L) = 2m$  if the network is unweighted or the weights are normalized such that  $\sum_{j \neq i} a_{ij} = 1$ .

Based on the type of graph under study, several interesting conclusions can be derived from the structure or connectivity properties of a graph, which can be measured by graph and complex network metrics. One metric of interest in Chapter 4 is the global clustering coefficient of a graph  $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ :

$$\text{CC} = \frac{1}{n} \sum_{i=1}^n \frac{2T_i}{k_i(k_i - 1)}, \quad (2.4)$$

where  $k_i = \sum_j a_{ij}$  is the node degree of node  $v_i \in \mathcal{V}$ , and  $T_i$  is the number of closed triangles in  $\mathcal{G}$  including  $v_i$ . The reader is referred to (Chen et al., 2013; Costa et al., 2007) for further details on network metrics.

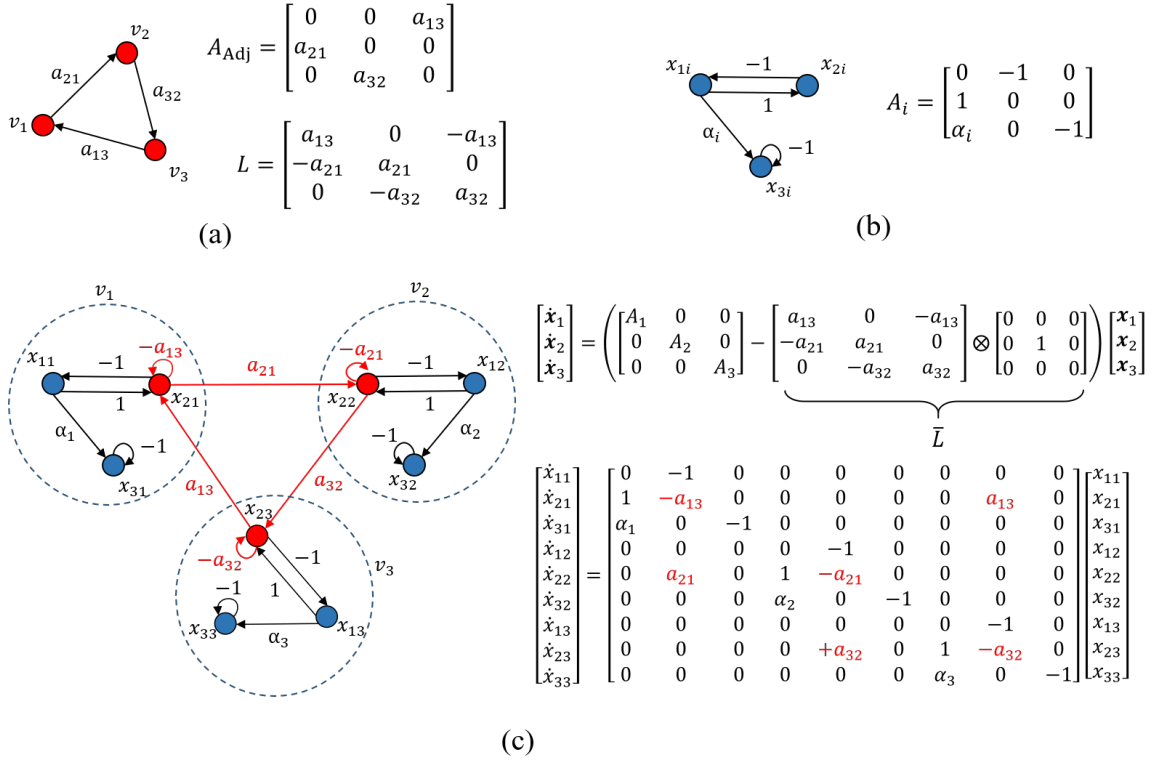
## 2.3 Representation of Dynamical Networks

A *dynamical network* is a set of dynamical systems interconnected according to a network topology described by  $A_{\text{adj}}$ . Although the individual dynamical systems at nodes are relatively simple, the interactions among them considerably raises the network complexity. This interplay is not only governed by the nodal dynamics and the adjacency matrix, but also by the coupling method. In this section, we show how dynamical networks can be mathematically represented from a graph approach and a dynamic systems approach.

### 2.3.1 Graph representation

A network system can be described by a graph  $\mathcal{G}$  which determines the interconnection structure among every element of  $\mathcal{V}$ , that is, the network topology. To  $\mathcal{G}$  we associate  $A_{\text{adj}} \in \mathbb{R}^{m \times m}$ . In the case of a dynamical network composed by linear time-invariant (LTI) systems, each node  $v_i$  is composed of a dynamical system  $(A_i, B_i, C_i, D_i)$  which itself can be represented by a graph  $\mathcal{G}_i = \{\mathcal{X}_i, \mathcal{E}_i\}$ . In this case, the adjacency matrix of  $\mathcal{G}_i$  is the corresponding dynamical matrix  $A_i \in \mathbb{R}^{n_i \times n_i}$ . Hence, every node in  $\mathcal{G}$  is expanded as a subgraph  $\mathcal{G}_i$  and, therefore, the *full network* is represented by a larger and more complex graph  $\mathcal{G}_{\text{full}} = \{\mathcal{V}_{\text{full}}, \mathcal{E}_{\text{full}}\}$  (and corresponding adjacency matrix  $A_{\text{adj}}^{\text{full}} \in \mathbb{R}^{N \times N}$ , where  $N = \sum_{i=1}^m n_i$ ).

We illustrate this representation in Fig. 2.3. Consider a network of  $m = 3$  nodes whose topology is described by a graph  $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$  (Fig. 2.3a). Consider that each node  $v_i \in \mathcal{V}$  represents a 3-dimensional linear dynamical system  $(A_i, 0, 0, 0)$ , with its corresponding graph  $\mathcal{G}_i$  illustrated in Fig. 2.3b. The full network  $\mathcal{G}_{\text{full}}$  (Fig. 2.3c) is,



**Figure 2.3:** (a) Network topology graph  $\mathcal{G}$ , and respective adjacency matrix  $A_{\text{adj}}$  and Laplacian matrix  $L$ . (b) Nodal dynamical system graph  $\mathcal{G}_i$ , for  $i = 1, 2, 3$ , of a 3-dimensional linear oscillator  $\mathcal{X}_i = \{x_{1i}, x_{2i}, x_{3i}\}$ , and respective dynamical matrix  $A_i$ . (c) Full network graph  $\mathcal{G}_{\text{full}}$  of a network topology graph described in (a), where each node is composed of a linear oscillator presented in (b) coupled by the  $x_{2i}$  variable.

therefore, given by  $\mathcal{G}$ , where each element of  $\mathcal{V}$  is expanded as a subgraph  $\mathcal{G}_i$ . In this case, graphs  $\{\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3\}$  are subgraphs of  $\mathcal{G}_{\text{full}}$  (however, this is not always the case, as seen in Example 2.2).

Not only expanding each node as a subgraph is essential since it includes the effects of nodal dynamics in the representation, but it also highlights which is the coupling variable between the interconnected nodes. Albeit for the higher dimensionality ( $|\mathcal{G}_{\text{full}}| = N$ ), it is clear that the full network representation is a more complex and complete model of the dynamical network, which can potentially lead to a more reliable analysis of the system.

This notation is only valid for linear dynamical matrices coupled by a linear graph (i.e., a linear coupling method). For nonlinear dynamical systems (2.2), or nonlinear couplings between the nodes, we use the following representation (Letellier et al., 2018): linear connections are represented by solid lines, while nonlinearities are represented by dashed lines. This is a reminder that nonlinear connections are no longer constant and

might vanish under specific circumstances. The nonlinear graph now faces singularity issues that can have a huge impact for the “information flow” between two nodes, or vertices, interconnected by a nonlinear edge. See Example 2.2 for further details. This representation has great value for symbolic analysis, as presented in an observability context (Bianco-Martinez et al., 2015; Letellier and Aguirre, 2009; Letellier et al., 2018).

### 2.3.2 State-space representation

A dynamical network can also be represented as a larger dynamical system (2.1) of higher-dimensionality  $N = \sum_{i=1}^m n_i$ , where  $n_i$  is the dimension of the  $i$ -th nodal dynamical system and  $m$  is the cardinality of the network topology  $\mathcal{G}$ . This high-dimensional representation, however, is usually detrimental to classical methods from system analysis and control design (Chen, 2014). Nevertheless, a dynamical network is a special case of a high-dimensional system. Since many applications in network systems have a rather sparse network topology and similar nodal dynamics (same set of ODEs but with parametric differences, yielding  $N = mn$ , where  $n_i = n, \forall i$ ), it is in the best interest of analysis and control methods of dynamical networks to take advantage of these properties.

In this sense, a compact state-space representation of a *full (dynamical) network* is presented in what follows, adapted from the work of Pecora and Carroll (1998):

$$\begin{bmatrix} \dot{\mathbf{x}}_1 \\ \dot{\mathbf{x}}_2 \\ \vdots \\ \dot{\mathbf{x}}_m \end{bmatrix} = \left( \begin{bmatrix} A_1 & 0 & \dots & 0 \\ 0 & A_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A_m \end{bmatrix} - \bar{L} \right) \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_m \end{bmatrix}, \quad (2.5)$$

where  $\mathbf{x} = [\mathbf{x}_1^\top \dots \mathbf{x}_m^\top]^\top \in \mathbb{R}^N$  and  $\mathbf{x}_i = [x_{1i} \dots x_{ni}]^\top \in \mathbb{R}^n$ , for  $i = 1, \dots, m$ . Thus,  $x_{ji}$  is the  $j$ -th state variable of the dynamical system at node  $v_i$ , and  $\mathbf{x}_i$  is the corresponding state vector.  $\bar{L}$  is the Laplacian matrix of the full network  $\mathcal{G}_{\text{full}}$ , describing the connection among all the state variables. The negative sign before  $\bar{L}$  implies that the state variables are diffusively coupled<sup>1</sup>.

The Laplacian matrix  $\bar{L}$  of  $\mathcal{G}_{\text{full}}$  is related to the Laplacian matrix  $L$  of the network topology graph  $\mathcal{G}$  as follows:

$$\bar{L} = L \otimes M, \quad (2.6)$$

<sup>1</sup>Two variables  $(\mathbf{x}_i, \mathbf{x}_j)$  are diffusively coupled by a coupling function  $\mathbf{g}(\mathbf{x}_i, \mathbf{x}_j)$  if  $\mathbf{g}(\mathbf{x}_i, \mathbf{x}_j) = -\mathbf{g}(\mathbf{x}_j, \mathbf{x}_i)$ .

where  $\otimes$  denotes the Kronecker product, and  $M = [m_{ij}] \in \{0, 1\}^{n \times n}$  is the “coupling matrix” that defines how the state variables are interconnected among themselves. That is, if  $m_{\bar{ij}} = 1$ , then an edge connecting node  $v_j$  to  $v_i$  in  $\mathcal{G}$  is actually coupling the state variable  $x_{\bar{jj}}$  to  $x_{\bar{ii}}$  in  $\mathcal{G}_{\text{full}}$ .

For instance, note that the “full graph” in Fig. 2.3c can be represented as in (2.5). Furthermore, if we assume that  $\alpha_i = \alpha$  (yielding  $A_i = A$ ), for  $i = 1, 2, 3$ , then (2.5) can be represented in a compact form as:

$$\dot{\mathbf{x}} = (I_3 \otimes A - L \otimes M) \mathbf{x}, \quad (2.7)$$

where  $I_3$  is the identity matrix of order 3. Indeed, this compact notation highlights the several levels of interplay in a dynamical network, including: (i) the nodal dynamics  $A$ , (ii) the network topology  $L$ , (iii) and the coupled state variables between interconnected nodes, represented by the coupling matrix  $M$ .

### 2.3.3 Examples

In this work, two specific oscillators are taken as benchmark examples for theoretical and numerical analysis in dynamical networks: the Kuramoto phase oscillator (Kuramoto, 1975) and the Rössler attractor (Rössler, 1976). The former for its rich dynamical behaviour with the added advantage of being described by rather simple equations. The latter for its chaotic behaviour, wide knowledge in the literature and interesting observability properties. The state space representations of both oscillators are presented in the following examples.

#### Example 2.1. Network of Kuramoto phase oscillators.

The Kuramoto phase oscillator is a 1-dimensional linear dynamical system defined as (Kuramoto, 1975):

$$\dot{x} = \omega, \quad (2.8)$$

where  $x \in \mathbb{R}$  is the oscillator phase angle, and  $\omega$  is the natural frequency.

Although described by a rather simple equation, the Kuramoto oscillator shows a rich dynamical behaviour when interconnected in a network by a sinusoidal coupling (Dorfler and Bullo, 2014). Consider a network  $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$  of Kuramoto oscillators where each oscillator (node)  $v_i \in \mathcal{V}$  is represented by the phase angle  $x_i \in \mathbb{R}$  (i.e.  $v_i := x_i$ ). This dynamical network can be represented by the following continuous state-space model (Chen et al., 2013; Dorfler and Bullo, 2014; Moreno and Pacheco,

2004; Wang and Chen, 2002a):

$$\dot{x}_i = \omega_i + \sum_{j=1}^N a_{ij} \sin(x_j - x_i), \quad i = 1, 2, \dots, m, \quad (2.9)$$

where  $\mathbf{x} = [x_1 \ x_2 \ \dots \ x_N]^\top \in \mathbb{R}^N$ ,  $N = m$  is the network size, and  $A_{\text{adj}} = [a_{ij}]$ . Note that the coupling among the oscillators is additive, diffusive and proportional to the coupling strength  $a_{ij}$ . Alternatively, (2.9) can be rewritten as a function of the Laplacian matrix:

$$\dot{\mathbf{x}} = \boldsymbol{\omega} + \sum_{j=1}^N A_{\text{adj}} \odot \sin(\mathbf{1}\mathbf{x}^\top - \mathbf{x}\mathbf{1}^\top), \quad (2.10)$$

where  $\boldsymbol{\omega} = [\omega_1 \ \dots \ \omega_N]^\top$  and  $\mathbf{1} \in \{1\}^N$  ( $N$ -dimensional column vector of “ones”), and  $\odot$  denotes the Hadamard product (element-wise product). If  $(x_j - x_i)$  is bounded in a small region around the equilibrium point, it is possible to linearize (2.10) with a reasonable accuracy, yielding the following linear representation

$$\dot{\mathbf{x}} = \boldsymbol{\omega} - L\mathbf{x}. \quad (2.11)$$

### Example 2.2. Network of Rössler systems.

The well-known Rössler system is given by the following set of ODEs (Rössler, 1976):

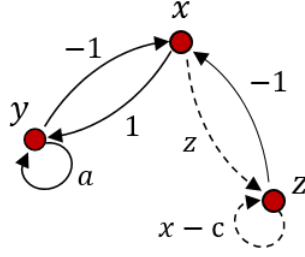
$$\begin{cases} \dot{x} &= -y - z \\ \dot{y} &= x + ay \\ \dot{z} &= b + z(x - c) \end{cases} \quad (2.12)$$

where this system settles to a chaotic attractor for  $(a, b, c) = (0.398, 2, 4)$ .

Figure 2.4 illustrates how the Rössler system can be represented as a “nonlinear graph”, proposed by Letellier et al. (2018), using the Jacobian matrix  $D\mathbf{f}$  of (2.12):

$$D\mathbf{f} = \begin{bmatrix} 0 & -1 & -1 \\ 1 & a & 0 \\ z & 0 & x - c \end{bmatrix}. \quad (2.13)$$

Consider now a network of  $m$  Rössler oscillators linearly coupled by means of the variable  $y$  (Boccaletti et al., 2002; Pecora and Carroll, 1990). Hence, at each node  $v_i \in \mathcal{V}$  there is a system with state variables  $\mathbf{x}_i = [x_i \ y_i \ z_i]^\top$ . The dynamical network



**Figure 2.4:** Nonlinear graph of the Jacobian matrix of the Rössler system. Linear and nonlinear edges are represented by solid and dashed lines, respectively. This convention is a reminder that nonlinear connections are no longer constant and might vanish under specific circumstances. For instance edges  $a_{33} = x - c$  and  $a_{31} = z$  vanish at  $x(t) = c$  and  $z(t) = 0$ , respectively. Such singularities might have a huge impact on the “information flow” between two nodes interconnected by a nonlinear edge.

can be represented by the following state-space model:

$$\begin{cases} \dot{x}_i &= -y_i - z_i \\ \dot{y}_i &= x_i + a_i y_i + \sum_{j=1}^m a_{ij} (y_j - y_i), \\ \dot{z}_i &= b_i + z_i (x_i - c_i) \end{cases} \quad (2.14)$$

for  $i = 1, \dots, m$ ,  $(a_i, b_i, c_i)$  are the parameters of the  $i$ th Rössler system, and  $A_{\text{adj}} = [a_{ij}]$ . This is a state-space model of dimension  $N = 3m$ . Note that, differently from Example 2.1, the diffusive coupling is linear.

Coupling the Rössler oscillators directly from  $y$  to  $x$ , yields

$$\begin{cases} \dot{x}_i &= -y_i - z_i + \sum_{j=1}^m a_{ij} (y_j - x_i) \\ \dot{y}_i &= x_i + a_i y_i \\ \dot{z}_i &= b_i + z_i (x_i - c_i) \end{cases} \quad (2.15)$$

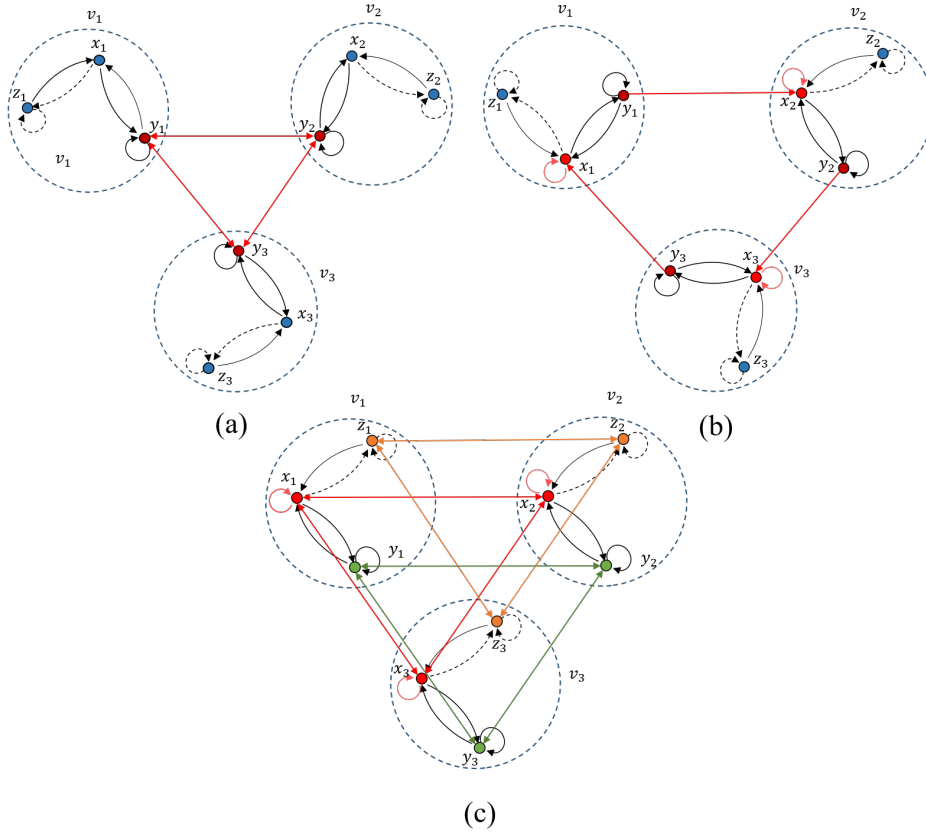
or undirectedly coupled by all variables, also known as “network of networks” (Chapman et al., 2014):

$$\begin{cases} \dot{x}_i &= -y_i - z_i + \sum_{j=1}^m a_{ij} (x_j - x_i) \\ \dot{y}_i &= x_i + a_i y_i + \sum_{j=1}^m a_{ij} (y_j - y_i) \\ \dot{z}_i &= b_i + z_i (x_i - c_i) + \sum_{j=1}^m a_{ij} (z_j - z_i) \end{cases} \quad (2.16)$$

for  $i = 1, \dots, m$ .

Dynamical networks (2.14), (2.15) and (2.16) are illustrated as full networks in Fig. 2.5. Equations (2.14)–(2.16) can be represented similarly to (2.7), highlighting the





**Figure 2.5:** Full network  $\mathcal{G}_{\text{full}}$  of Rössler systems coupled (a) undirectedly by  $y$  variable (diffusive coupling), (b) directly from variable  $y$  to  $x$ , and (c) undirectedly between all respective variables. Network topology  $\mathcal{G}$  is a cycle graph with  $m = 3$ . Self-edges are included (or have the weight modified, if already present) to the coupled vertices because of the diffusive coupling. Note that, unlike to Fig. 2.3, in example (b),  $\mathcal{G}$  is not a subgraph of  $\mathcal{G}_{\text{full}}$ .

network structure and coupling variable, as follows:

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) - (L \otimes M) \mathbf{x}, \quad (2.17)$$

where  $\mathbf{x} = [x_1 \ y_1 \ z_1 \ \dots \ x_m \ y_m \ z_m]^\top$ ,  $\mathbf{f} = [\mathbf{f}_1^\top(\mathbf{x}_1) \ \dots \ \mathbf{f}_m^\top(\mathbf{x}_m)]^\top : \mathbb{R}^N \mapsto \mathbb{R}^N$ ,  $\mathbf{f}_i(\mathbf{x}_i) : \mathbb{R}^3 \mapsto \mathbb{R}^3$  is a nonlinear function that describes the nodal dynamics of  $\mathbf{x}_i$  according to (2.12), and the coupling matrix  $M$  is given by

$$M = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \text{or} \quad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad (2.18)$$

for (2.14), (2.15) and (2.16), respectively.  $\triangle$



## Chapter 3

# A Review on Observability of Network Systems

Observability is a property that determines if the trajectory temporal evolution of the internal states of a dynamical system can be reconstructed based on knowledge of the inputs and outputs, as introduced by [Kalman \(1959\)](#). This classic concept of observability, addressed here as *structural observability*, is based on a *crisp* definition, i.e. the system is or is not observable ([Chen, 1999](#)). This crisp classification might be misleading since, in some *ill-conditioned* cases, a small change in the parameter space of an unobservable dynamical system might make it observable, and vice-versa ([Friedland, 1975](#)). Thus, although observability is a sufficient and necessary condition for the design of a state observer ([Luenberger, 1966](#)), a more important question for practical purposes might be whether a system is *almost unobservable* or not. This structural definition can be extended by metrics that quantify observability in a gradual or continuous manner, i.e. measuring *how well* the system trajectory can be reconstructed ([Aguirre, 1995](#); [Friedland, 1975](#)), which we address as *dynamical observability*. Dynamical observability not only allows one to identify if a given dynamical system is observable from a practical point-of-view, but it also allows one to quantify and rank the degree of observability conveyed by different sets of output measures—and therefore choose the best option ([Letellier et al., 1998](#)). Sections [3.1](#) and [3.2](#) review both approaches to quantify the observability of a linear and nonlinear dynamical system.

In the context of network systems, it is a reasonable assumption that not all nodes are available for measurement. For instance, not every single neuron of the one hundred billion neurons present in the brain are physically accessible for direct measurement. Likewise, it might not be economically viable to place a phasor measurement unit

in every single electrical substation of a power system. Thus, two important goals in the field of network systems are: i) to determine if a given (minimum) set of sensor nodes<sup>1</sup> renders the network observable, a problem that can be assessed by *structural observability* metrics; and ii) to determine the *best* set of sensor nodes from different configurations with the same cardinality, a problem that can only be solved by *dynamical observability* metrics. Indeed, the optimal sensor placement problem can be approached as an observability problem.

However, classical observability metrics from control theory are unfeasible for high-dimensional (network) systems. To circumvent this issue, Liu et al. (2011b) presented a pioneer work on the controllability of network systems<sup>2</sup>, where the notion of network controllability is revisited under a different definition grounded on graph theory (Lin, 1974), which we address here as *topological controllability and observability*, reviewed in Section 3.3. This opened a recent and new branch for research in the control of (complex) network systems (Liu and Barabási, 2016), which has gathered several—but not yet consolidated—results in the literature.

Undoubtedly many works in the literature embraced the graph-theoretical approach to the controllability and observability of network systems (Gao et al., 2014; Jia et al., 2013; Leitold et al., 2017; Nacher and Akutsu, 2013; Pósfai et al., 2013; Yan et al., 2015), leading to some major developments in this field, including applications in the control of neuronal networks (Gu et al., 2015; Su et al., 2017). This preference for the graph-theoretical perspective of observability, developed by Lin (1974), is mainly due to the intuitive representation of network systems by graph models and the high scalability of graph metrics. However, it must be noted that a great number of results in this field totally disregard the effects of nodal dynamics. This led to recurrent criticism on the true applicability of their work (Cowan et al., 2012; Gates and Rocha, 2015; Leitold et al., 2017; Wang et al., 2017), as upcoming works showed that the topological observability severely underestimates the required number of sensor nodes for practical purposes such as state estimation (Haber et al., 2018; Montanari and Aguirre, 2019). Not only that, but the conclusions might be drastically different when the effects of nonlinearity are properly taken into account (Jiang and Ying-Cheng Lai, 2019; Letellier et al., 2018; Motter, 2015). Indeed, the interplay between observability of a network

---

<sup>1</sup>A common jargon in literature is to refer to nodes available for measurement (that issue output signals) as *sensor nodes*, and control nodes (that have input signals) as *driver nodes*.

<sup>2</sup>The study of controllability and observability of network systems was born in a controllability context. Thus, some of the discussion in this work might be focused on controllability rather than observability, although we can rely on the duality between both concepts (in a linear context) to understand and compare the available results.

system, the graph structure and the nodal dynamics, therefore, remains an open subject to study in network systems.

The main contribution of this chapter is not only to survey the literature of observability of network systems—which has been done before by [Liu and Barabási \(2016\)](#)—but mainly to review, with a critical mindset, recent advances on the observability of network systems from a control theory perspective. The pros and cons of a topological approach to the observability of network systems are exposed in [Section 3.4](#), while [Section 3.5](#) provides some guidelines and future research directions in this field of work. To confront the concepts of structural, dynamical and topological observability at a “practical” level, we provide two application examples of optimal sensor placement in the context of power systems and multi-agent system consensus in [Section 3.6](#).

We refer to [\(Aguirre et al., 2018\)](#) for the adopted nomenclature and classification of observability metrics in this chapter. The contents of this chapter were published in [\(Montanari and Aguirre, 2020\)](#).

## 3.1 Structural Observability

We hereby address as *structural observability* to any crisp definition of observability based on a “yes-no” condition, i.e. a definition that classifies a system as either observable or not.

### 3.1.1 Linear dynamical systems

The classic concept of observability for linear systems was introduced by [Kalman \(1959\)](#). The following definition and theorem is further discussed and proven in many textbooks in linear systems theory, including the work by [Chen \(1999\)](#).

**Definition 3.1.** ([Chen, 1999, Definition 6.01](#)) *The linear system (2.1) or the pair  $(A, C)$  is said to be observable if for any unknown initial state  $\mathbf{x}(0)$ , there exists a finite time  $t_1 > 0$  such that the knowledge of the input  $\mathbf{u}$  and the output  $\mathbf{y}$  over  $t \in [0, t_1]$  suffices to uniquely determine  $\mathbf{x}(0)$ . Otherwise, (2.1) is said to be unobservable.*

**Theorem 3.1.** ([Chen, 1999, Theorem 6.01](#)) *The following statements are equivalent.*

1. *The  $n$ -dimensional pair  $(A, C)$  is observable.*
2. *The matrix  $W_o(t) \in \mathbb{R}^{n \times n}$*

$$W_o(t) = \int_0^t e^{A^\top \tau} C^\top C e^{A\tau} d\tau \quad (3.1)$$

is nonsingular for any  $t > t_0$ .

3. The observability matrix  $\mathcal{O} \in \mathbb{R}^{nq \times n}$

$$\mathcal{O} = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{n-1} \end{bmatrix} \quad (3.2)$$

has full column rank, i.e.  $\text{rank}(\mathcal{O}) = n$ .

4. The matrix  $\begin{bmatrix} (A - \lambda_i I)^\top & C^\top \end{bmatrix}^\top$  has full column rank at every eigenvalue  $\lambda_i$  of  $A$ , for  $i = 1, \dots, n$ .

5. If  $\text{Re}\{\lambda_i\} < 0$ , for all  $i = 1, \dots, n$ , then the unique solution of

$$A^\top W_o + W_o A = -C^\top C \quad (3.3)$$

is positive definite, is called observability Gramian and can be expressed as

$$W_o = \int_0^\infty e^{A^\top \tau} C^\top C e^{A\tau} d\tau. \quad (3.4)$$

*Proof.* Equivalence between statements (1)-(5) is proven in (Chen, 1999). We state an alternative proof on the equivalence between statements (1) and (3), found in (O'Reilly, 1983). Consider the linear system (2.1), where the temporal evolution of  $\mathbf{y}(t)$  is given by

$$\mathbf{y}(t) = C e^{At} \mathbf{x}(t_0) + C \int_{t_0}^t e^{A(t-\tau)} B \mathbf{u}(\tau) d\tau + D \mathbf{u}(t). \quad (3.5)$$

Following Definition 3.1, since we suppose that  $\mathbf{u}(t)$  and  $\mathbf{y}(t)$  are known over  $t \in [t_0, t_1]$ , (3.5) can be rewritten as

$$\bar{\mathbf{y}}(t) = C e^{At} \mathbf{x}(t_0). \quad (3.6)$$

where  $\bar{\mathbf{y}}(t) := \mathbf{y}(t) - C \int_{t_0}^t e^{A(t-\tau)} B \mathbf{u}(\tau) d\tau + D \mathbf{u}(t)$  is a known variable. From Definition 3.1, system (2.1) is observable if  $\mathbf{x}(t_0)$  can be solved from (3.6). Differentiating (3.6)

successively around  $t = t_0$  yields

$$\begin{bmatrix} \bar{\mathbf{y}}(t_0) \\ \bar{\mathbf{y}}^{(1)}(t_0) \\ \vdots \\ \bar{\mathbf{y}}^{(n-1)}(t_0) \end{bmatrix} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} \mathbf{x}(t_0) \quad (3.7)$$

$$\tilde{\mathbf{y}}(t_0) := \mathcal{O}\mathbf{x}(t_0).$$

Equation (3.7) is a set of linear algebraic equations. Since  $\tilde{\mathbf{y}}(t_0)$  is known, the initial state  $\mathbf{x}(t_0)$  can be uniquely determined if  $\tilde{\mathbf{y}}$  lies in the column space of  $\mathcal{O}$ . We prove the sufficiency and the necessity of equivalence (1)-(3) from (3.7).

*Sufficiency.* If the system is unobservable, then there is an  $\mathbf{x} \neq 0$  such that  $\bar{\mathbf{y}} = Ce^{A(t-t_0)}\mathbf{x} = 0, \forall t \geq t_0$ . Then, from (3.6), we have  $C\mathbf{x} = 0, CA\mathbf{x} = 0, CA^2\mathbf{x} = 0, \dots$ . Therefore, the system is unobservable if there is a  $\mathbf{x} \neq 0$  that is orthogonal to every element of  $\mathcal{O}$ , which is only possible if  $\text{rank}(\mathcal{O}) < n$ .

*Necessity.* Suppose  $\text{rank}(\mathcal{O}) < n$ . Then there is an  $\mathbf{x} \neq 0$  in  $\mathbb{R}^n$  such that  $C\mathbf{x} = 0, CA\mathbf{x} = 0, CA^2\mathbf{x} = 0, \dots, CA^{n-1}\mathbf{x} = 0$ . Because, from Cayley-Hamilton Theorem (Chen, 1999, Theorem 3.4),  $CA^n$  is a linear combination of its lower degree terms  $\{C, CA, \dots, CA^{n-1}\}$ , the rank of  $\mathcal{O}$  stops growing if terms  $CA^i$ , for  $i \geq n$ , were added to it—which implies that the dimension of the column space of  $\mathcal{O}$  also stops growing. Thus, if  $\text{rank}(\mathcal{O}) < n$ , then there is an  $\mathbf{x} \neq 0$  lying in the column space of  $\mathcal{O}$  which is not reconstructible (yielding an unobservable system).  $\square$

From the proof above, it is clear that, if a pair  $(A, C)$  is observable, then a solution for (3.7) exists, given by

$$\mathbf{x}(0) = \mathcal{O}^\dagger \tilde{\mathbf{y}}(0), \quad (3.8)$$

where  $\mathcal{O}^\dagger$  denotes the Moore-Penrose inverse (pseudoinverse). Since  $\text{rank}(\mathcal{O}) = n$ , then  $\mathcal{O}^\dagger := (\mathcal{O}^\top \mathcal{O})^{-1} \mathcal{O}^\top$  and, therefore, solution  $\mathbf{x}(0)$  is unique.

**Remark 3.1.** Computation of (3.8) is not feasible for practical applications since it requires differentiation of  $\bar{\mathbf{y}}(t)$ , which amplifies high-frequency noise in measurements. Nevertheless, proving that  $\text{rank}(\mathcal{O}) = n$  guarantees that  $\mathbf{x}(t_0)$  can be uniquely determined and is also a sufficient and necessary condition for existence of stable state observers. Moreover, the observability matrix is also related to subspace identification methods (Haber and Verhaegen, 2014; Overschee and De Moor, 1996).

Note that Definition 3.1 only classifies the pair  $(A, C)$  as observable or unobservable. Thus, we refer to Theorem 3.1 as a *structural observability* property. This *crisp*

classification is due to the observability property being based on a rank condition of (3.2). Consequently, the observability property is a discontinuous function of the system parameters. Hence a small change in the parameter space of (2.1) can move a dynamical system from unobservable to observable, and vice-versa.

Suppose that matrix  $\mathcal{O}^\top \mathcal{O}$  is nonsingular, but *ill-conditioned*. For practical purposes, this means that the computation of its inverse is prone to large numerical errors, leading to large errors in the solution of (3.8). One might argue that in this case, a pair  $(A, C)$  is *almost unobservable*—or rather unobservable for practical purposes—and, therefore, the structural observability property of Theorem 3.1 is not suitable for certain applications due to its sensitivity to an ill-conditioned matrix (Friedland, 1975). This problem is further explored in Section 3.2.

### 3.1.2 Nonlinear dynamical systems

Several generalizations to observability of nonlinear systems have been proposed in the literature (Hermann and Krener, 1977; Letellier and Aguirre, 2009; Mesbahi et al., 2019; Sontag, 1991; Zabczyk, 1995; Zhirabok and Shumsky, 2012). This work follows the definition of *local weak observability*<sup>3</sup>, grounded on differential geometry, established by Hermann and Krener (1977).

Consider the nonlinear system (2.2), or the pair  $\{\mathbf{f}, \mathbf{h}\}$ . Let the flow map  $\Phi_t(\mathbf{x}(t_0)) : \mathcal{M} \mapsto \mathcal{M}$  be the solution of (2.2), which defines the trajectory from a initial state  $\mathbf{x}(t_0)$  to a final state  $\mathbf{x}(t_0 + t)$ , given by

$$\Phi_t(\mathbf{x}(t_0)) := \mathbf{x}(t_0 + t) = \mathbf{x}(t_0) + \int_{t_0}^{t_0+t} \mathbf{f}(\mathbf{x}(\tau)) d\tau. \quad (3.9)$$

From Definition 3.1, the concept of observability can be generalized to nonlinear systems (2.2) by determining whether an initial state  $\mathbf{x}(t_0)$  can be uniquely reconstructed from the image of the composition map  $\mathbf{h} \circ \Phi_t$ . This is formally defined as follows (Hermann and Krener, 1977; Mesbahi et al., 2019).

**Definition 3.2.** *The nonlinear system (2.2), or the pair  $\{\mathbf{f}, \mathbf{h}\}$ , is*

- **locally observable at  $\mathbf{x}_0$**  *if there exists a neighbourhood  $\mathcal{U} \subseteq \mathcal{M}$  of  $\mathbf{x}_0$  such that for every state  $\mathbf{x}_0 \neq \mathbf{x}_1 \in \mathcal{U}$ ,  $\mathbf{h} \circ \Phi_t(\mathbf{x}_0) \neq \mathbf{h} \circ \Phi_t(\mathbf{x}_1)$  for some finite  $t > t_0$ ;*
- **locally observable** *if it is locally observable at every  $\mathbf{x}_0 \in \mathcal{M}$ ;*

<sup>3</sup>Note that the definitions of “weak” or “local weak” observability found in Hermann and Krener (1977) are omitted since they are equivalent to the “local” observability definition (defined below) when considering autonomous systems (2.2).



- and **observable** if its locally observable and the neighbourhood  $\mathcal{U}$  can be taken as  $\mathcal{M}$ .

Otherwise,  $\{\mathbf{f}, \mathbf{h}\}$  is said to be locally unobservable at  $\mathbf{x}_0$ .

An advantage of the local observability definition is that it can be verified through a simple algebraic test as follows.

**Theorem 3.2.** *Let the nonlinear system (2.2), or the pair  $\{\mathbf{f}, \mathbf{h}\}$  be of class  $C^s$ ,  $s \geq 1$ . The pair  $\{\mathbf{f}, \mathbf{h}\}$  is locally observable at  $\mathbf{x}_0$  if the observability matrix*

$$\mathcal{O}(\mathbf{x}_0) = \frac{\partial}{\partial \mathbf{x}} \begin{bmatrix} \mathcal{L}_{\mathbf{f}}^0 \mathbf{h}(\mathbf{x}) \\ \vdots \\ \mathcal{L}_{\mathbf{f}}^{s-1} \mathbf{h}(\mathbf{x}) \end{bmatrix}_{\mathbf{x}=\mathbf{x}_0}, \quad (3.10)$$

is full rank, i.e.  $\text{rank}(\mathcal{O}(\mathbf{x}_0)) = n$ , where  $\mathcal{L}_{\mathbf{f}}^j \mathbf{h}(\mathbf{x}_0) := \nabla \mathbf{h} \cdot \mathbf{f}$  is the  $j$ -th Lie derivative of  $\mathbf{h}$  along the vector field  $\mathbf{f}$  at  $\mathbf{x} = \mathbf{x}_0$ .

*Proof.* (Klaus and Reinschke, 1999) From Definition 3.2,  $\{\mathbf{f}, \mathbf{h}\}$  is observable if the map

$$\Psi_{\mathbf{h}}(\mathbf{x}) : \mathbf{x} \mapsto \left[ \mathbf{y}^{\top} \left[ \mathbf{y}^{(1)} \right]^{\top} \dots \left[ \mathbf{y}^{(s-1)} \right]^{\top} \right]^{\top} \quad (3.11)$$

is invertible (injective) for a given  $s \geq 1$ —in other words, if it is possible to uniquely determine  $\mathbf{x}$  from  $\mathbf{y}$  (and its successive derivatives). The inverse function theorem provides a *sufficient condition* for local invertibility of general nonlinear maps:  $\Psi_{\mathbf{h}}(\mathbf{x})$  is locally invertible at  $\mathbf{x}_0$  if its Jacobian matrix has full rank, i.e.

$$\text{rank} \left( \frac{\partial \Psi_{\mathbf{h}}(\mathbf{x})}{\partial \mathbf{x}} \right) \Big|_{\mathbf{x}=\mathbf{x}_0} = n. \quad (3.12)$$

Substituting (2.2) in (3.11), yields the nonlinear map

$$\Psi_{\mathbf{h}}(\mathbf{x}) := \begin{bmatrix} \mathbf{y} \\ \mathbf{y}^{(1)} \\ \vdots \\ \mathbf{y}^{(s-1)} \end{bmatrix} = \begin{bmatrix} \mathbf{h}(\mathbf{x}) \\ \frac{d\mathbf{h}(\mathbf{x})}{dt} \\ \vdots \\ \frac{d^{s-1}\mathbf{h}(\mathbf{x})}{dt^{s-1}} \end{bmatrix} = \begin{bmatrix} \mathcal{L}_{\mathbf{f}}^0 \mathbf{h}(\mathbf{x}) \\ \mathcal{L}_{\mathbf{f}}^1 \mathbf{h}(\mathbf{x}) \\ \vdots \\ \mathcal{L}_{\mathbf{f}}^{s-1} \mathbf{h}(\mathbf{x}) \end{bmatrix}, \quad (3.13)$$

where higher-order Lie derivatives are defined as

$$\mathcal{L}_{\mathbf{f}}^j \mathbf{h}(\mathbf{x}) := \frac{\partial \mathcal{L}_{\mathbf{f}}^{j-1} \mathbf{h}(\mathbf{x})}{\partial \mathbf{x}} \mathbf{f}(\mathbf{x}), \quad (3.14)$$

for  $\mathcal{L}_f^0 \mathbf{h}(\mathbf{x}) := \mathbf{h}(\mathbf{x})$ .

It follows that the Jacobian matrix of  $\Psi_h(\mathbf{x})$  is equivalent to  $\mathcal{O}(\mathbf{x})$  and hence condition (3.12) implies local invertibility (and observability) at  $\mathbf{x}_0$ .  $\square$

**Remark 3.2.** It follows that (3.10) reduces to (3.2) if  $\mathbf{f}$  and  $\mathbf{h}$  are linear functions.

**Remark 3.3.** In the context of single measurement ( $q = 1$ ), the nonlinear observability matrix (3.10) is full rank only if  $s \geq n$ . Theoretically, however, the necessary number  $s$  of Lie derivatives so that the nonlinear system is observable depends on  $\{\mathbf{f}, \mathbf{h}\}$  and can even tend to infinity (Mesbahi et al., 2019; Zabczyk, 1995).

The discussion in Section 3.1.1 regarding this crisp classification of observability and the effects of ill-conditioning of  $\mathcal{O}$  also holds for the nonlinear case. Naturally, computational burden is aggravated for the nonlinear case, since computation of (3.10) is more intensive than (3.2).

**Observability and embedding theory.** A relation between observability and embedding theory follows naturally from the proof of Theorem 3.2 (Letellier et al., 2005). If  $h : \mathbb{R}^n \rightarrow \mathbb{R}^1$  and  $y \in \mathbb{R}^1$  (single output), then the pair  $\{\mathbf{f}, h\}$  is locally observable if the Jacobian matrix of the map  $\Psi_h$  is locally nonsingular. In this case, (3.10) is the Jacobian matrix of the map  $\Psi_h$  between the original state-space and the  $n$ -dimensional differential embedding space (Letellier et al., 2005)<sup>4</sup>. If  $\Psi_h$  is nonsingular for all  $\mathbf{x}$ , then there is a global diffeomorphism and the pair  $\{\mathbf{f}, h\}$  is fully and globally observable.

**Example 3.1. Structural observability.**

Consider Rössler system (2.12). If  $h(\mathbf{x}) = y$  (the recorded variable is the  $y$  variable of the Rössler system), then the observability matrix  $\mathcal{O}_y(\mathbf{x})$  (or the Jacobian matrix of the map  $\Psi_y^3 : \mathbf{x} \mapsto [y \ \dot{y} \ \ddot{y}]$ ) is

$$\frac{\partial \Psi_y^3}{\partial \mathbf{x}} \equiv \mathcal{O}_y(\mathbf{x}) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & a & 0 \\ a & a^2 - 1 & -1 \end{bmatrix}. \quad (3.15)$$

Since  $\mathcal{O}_y(\mathbf{x})$  is constant and nonsingular (implying full rank) for all  $\mathbf{x} \in \mathbb{R}^n$ , the Rössler system is fully (globally) observable from the  $y$  variable.

<sup>4</sup>The relation between the observability matrix and the Jacobian matrix of map  $\Psi_h$  was investigated for multivariate embedding ( $\mathbf{h} : \mathbb{R}^n \mapsto \mathbb{R}^q$ ) in (Aguirre and Letellier, 2005).

If  $h(\mathbf{x}) = z$ , then the observability matrix  $\mathcal{O}_z(\mathbf{x})$  (or the Jacobian matrix of the map  $\Psi_z^3 : \mathbf{x} \mapsto [z \ \dot{z} \ \ddot{z}]$ ) is

$$\frac{\partial \Psi_z^3}{\partial \mathbf{x}} \equiv \mathcal{O}_z(\mathbf{x}) = \begin{bmatrix} 0 & 0 & 1 \\ z & 0 & x - c \\ b + 2z(x - c) & -z & (x - c)^2 - y - 2z \end{bmatrix}. \quad (3.16)$$

Since  $\mathcal{O}_z(\mathbf{x})$  is not constant, we refer to Theorem 3.2. Note that  $\mathcal{O}_z(\mathbf{x})$  is singular for  $z = 0$ , since  $\det(\mathcal{O}_z(\mathbf{x})) = -z^2$ . Thus, considering the definition of structural observability, the Rössler system is unobservable for  $z = 0$  and observable for  $z \in \mathbb{R} \setminus \{0\}$ .

This raises the following question: how good is the trajectory reconstruction from measurements on  $z$  in the vicinity of  $z = 0$ ? This question is further explored with the definition of dynamical observability.  $\triangle$

## 3.2 Dynamical Observability

As detailed in Section 3.1, the structural classification of observability faces several practical problems, such as the feasibility of reconstructing a dynamical system trajectory from a set of output signals whose observability matrix is ill-conditioned (sensitive to small changes). If one has access to different sets of output signals for a given dynamical system, it is relevant, for practical purposes, to classify a system not only as observable or not, but also to establish a continuous quantification of observability levels conveyed by each available set of output signals. In this way, one can distinct observable systems between conditions of “poor” and “rich” observability, which directly affect the reconstruction quality of the dynamical system trajectory. Metrics that quantify observability in a continuous manner, rather than discrete, are referred in this work as *dynamical observability* coefficients (Aguirre et al., 2018).

### 3.2.1 Linear dynamical systems

As the problem of investigating dynamical observability seems to be related with the conditioning of matrix  $\mathcal{O}$ , Friedland (1975) proposed the *condition number* as an alternative:

$$\kappa(A) = \|A^{-1}\| \|A\| = \frac{\sigma_1(A)}{\sigma_n(A)}, \quad (3.17)$$

where we have used the  $\ell_2$ -norm, and  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n$  are the singular values of a square matrix  $A$ . As  $\kappa(A)$  increases, the numerical condition of  $A$  degrades.

Friedland (1975) adapted the conditioning number for a more intuitive quantification of observability as follows.

**Definition 3.3.** The “coefficient of observability” of a pair  $(A, C)$  is defined as

$$\delta := \left| \frac{\lambda_{\min}(F)}{\lambda_{\max}(F)} \right|. \quad (3.18)$$

where  $0 \leq \delta \leq 1$ ,  $F = \mathcal{O}^\top \mathcal{O}$  or  $F = W_o$ , and  $\lambda_{\min}$  and  $\lambda_{\max}$  refer to the minimum and maximum eigenvalues<sup>5</sup> of  $F$ . If  $\lambda_{\max}(F) = 0$ , then necessarily  $\lambda_{\min}(F) = 0$  and the dimension of the observable space is zero by definition.

Larger values of  $\delta$  indicate that  $(A, C)$  is more observable. Thus, even when  $\mathcal{O}$  is full rank, a small  $\delta$  points to *poor observability*. If  $\delta = 0$ , then  $(A, C)$  is not observable. The use of observability coefficients allows one to decide if a given set of output signals (e.g.  $\mathbf{y}_1 = C_1 \mathbf{x}$ ) conveys more or less observability to the dynamical system (2.1) than another set of output signals (e.g.  $\mathbf{y}_2 = C_2 \mathbf{x}$ ). Thus, if both pairs  $(A, C_1)$  and  $(A, C_2)$  are observable according to Theorem 3.1, one can use Definition 3.3 to distinguish between conditions of “poor” and “rich” observability, which directly affects the reconstruction quality of this dynamical system trajectory (Montanari and Aguirre, 2019).

**Remark 3.4.** Although the structural observability definition is invariant under similarity transformations (Chen, 1999, Theorem 6.02), the coefficient of dynamical observability  $\delta$  is sensitive to similarity transformations (Aguirre, 1995).

**Remark 3.5.** Coefficient  $\delta$  usually increases with the dimension  $q$ : typically more outputs increase the quality of observability. However this is not always true for some networks of oscillators (Montanari and Aguirre, 2019).

Definition 3.3 is found in several works in the literature based on the conditioning number of the observability matrix (Aguirre and Letellier, 2005; Aguirre et al., 2018; Friedland, 1975; Luan and Tsvetkov, 2019; Montanari and Aguirre, 2019; Wang et al., 2017; Whalen et al., 2015). Nevertheless, it is worth mentioning that other coefficients of dynamical observability have been proposed in literature based on the observability Gramian (Johnson, 1969; Pasqualetti et al., 2013b; Summers et al., 2016):

- the trace of the Gramian  $\text{tr}(W_o)$ , related to the average observation energy in all directions of the observable subspace;

---

<sup>5</sup>Since  $F$  is symmetric, its singular values are equal to the absolute values of its eigenvalues.

- the determinant of the Gramian  $\det(W_o)$ , a volumetric measure of the set of spaces which can be observed within one unit of energy;
- and the smallest eigenvalue of the Gramian  $\lambda_{\min}(W_o)$ , a worst-case metric related to the amount of energy required to observe the most difficult state.

It is not yet clear in the literature the distinction between coefficients of observability based on the observability Gramian and the observability matrix<sup>6</sup>. For instance, [Sun and Motter \(2013\)](#) show that a given continuous-time dynamical system might have an ill-conditioned observability Gramian and a well-conditioned observability matrix.

### 3.2.2 Nonlinear dynamical systems

Definition 3.3 was extended to nonlinear dynamical systems by [Letellier and Aguirre \(2002\)](#); [Letellier et al. \(1998\)](#).

**Definition 3.4.** The “local coefficient of observability” at  $\mathbf{x}_0$  of a pair  $\{\mathbf{f}, \mathbf{h}\}$  is defined as

$$\delta(\mathbf{x}_0) := \left| \frac{\lambda_{\min}(\mathcal{O}^\top(\mathbf{x}_0) \cdot \mathcal{O}(\mathbf{x}_0))}{\lambda_{\max}(\mathcal{O}^\top(\mathbf{x}_0) \cdot \mathcal{O}(\mathbf{x}_0))} \right|, \quad (3.19)$$

where  $0 \leq \delta(\mathbf{x}_0) \leq 1$ . The “global coefficient of observability” of  $\{\mathbf{f}, \mathbf{h}\}$  is the average along a trajectory  $\Phi_{t_1}(\mathbf{x}(t_0))$ , for  $t \in [t_0, t_1]$ :

$$\delta = \frac{1}{t_1 - t_0} \int_{t_0}^{t_1} \delta_y(\mathbf{x}(t)) dt. \quad (3.20)$$

The coefficient in (3.20) is not normalized. Hence, it is not meaningful to compare the coefficient of observability of variables (or sets of output signals) for *different* systems. It must be noted that the comparison is relative ([Aguirre et al., 2008](#)).

#### Example 3.2. Dynamical observability.

Consider Rössler system (2.12) and the observability matrix  $\mathcal{O}_y(\mathbf{x})$  built considering variable  $y$  as the measured (recorded) signal (see (3.15)). Since  $\mathcal{O}_y$  is constant over the entire state space there is a global diffeomorphism between original and reconstructed spaces. Computing (3.19) yields  $\delta_y = 0.133$ .

On the other hand, consider  $\mathcal{O}_z(\mathbf{x})$  in (3.16). As explained in Example 3.1, since  $\det(\mathcal{O}_z(\mathbf{x})) = -z^2$ , the system is not observable at  $z = 0$  and, from a dynamical observability point of view, it is poorly observable in the vicinity of the plane  $z = 0$ .

<sup>6</sup>Especially in the context of continuous-time systems. In the discrete-time case, the definitions of the observability matrix and Gramian are equivalent.

For instance, for  $z = 0.3$ , using (3.19) yields  $\delta_z(z = 0.3) = 1.32 \cdot 10^{-6}$ . However, equation (3.19) only provides a *local* quantification of the variable  $z$  observability. For a more *global* quantification, we compute the average (3.20) over the entire chaotic attractor, which yields  $\delta_z = 0.006$ —indicating, nonetheless, the poor observability of  $z$  when compared to  $y$ .

For the Rössler system with  $(a, b, c) = (0.398, 2, 4)$ , the following values were computed using (3.20):  $\delta_x = 0.022, \delta_y = 0.133, \delta_z = 0.006$ . Since  $\delta_y > \delta_x > \delta_z$ , it is stated that the *observability rank* of the recorded variables is  $y \triangleright x \triangleright z$  (Letellier and Aguirre, 2002).  $\triangle$

Examples 3.1 and 3.2 show that the loss of *local observability* is intrinsically related to the local singularities that  $\Psi_h$  may have, which are a consequence of nonlinearities. The following remark holds generally, though.

**Remark 3.6.** According to Takens' theorem (Takens, 1981), assuming that  $\{\mathbf{f}, h\}$  is structurally observable, if the dimension of the reconstructed space is increased, that is  $\Psi_h : \mathbf{x} \mapsto [y \ y^{(1)} \ \dots \ y^{(s-1)}]^\top$ , with  $s > n$  usually, then any singularities of  $\Psi_h$  may vanish and the pair  $\{\mathbf{f}, h\}$  gradually becomes more dynamically observable (Letellier et al., 2005).

### 3.3 Topological Observability

Sections 3.1 and 3.2 approached the study of observability from a system theory point of view. However, the developed methods are not particularly efficient for high-dimensional dynamical systems such as networks. Indeed, even if the full network dynamics were known, to find a set of sensor nodes that render a full network observable would require a brute force computation of  $\mathcal{O}$  over  $(2^N - 1)$  distinct combinations (Liu et al., 2011b)—which is not feasible at all. This process would be even more demanding if computation of eigenvalues in (3.18) or Lie derivatives in (3.10) were involved.

Faced with these challenges, a possible strategy to study the observability of a network system is to investigate it from a graph approach, which we refer to as *topological observability*. In this case, the topological observability is usually assessed solely from the network topology graph, although some works argue that the results are more representative when the topological observability is assessed from the full network graph (Aguirre et al., 2018; Cowan et al., 2012; Leitold et al., 2017), as discussed in Section 2.3.

Most studies developed over this idea follow the pioneer line of work of Liu et al. (2011b), grounded on the structural observability definition of Lin (1974).

### 3.3.1 Linear dynamical systems

Lin (1974) proposed a novel concept of structural controllability for linear systems, which was later extended to observability (Willems, 1986).

**Definition 3.5.** (Li et al., 2019, Definition 1) A matrix  $\bar{A} \in \{0, \star\}^{n \times n}$  is called a structured matrix if  $A = [a_{ij}]$  is either a fixed zero entry or an independent free parameter, denoted by a  $\star$ . A matrix  $\tilde{A}$  is a numerical realization of  $A$  if a real number is assigned to all free parameters of  $A$ .

**Definition 3.6.** The structured pair  $(A, C)$  is structurally observable if and only if there exists some numerical realization  $(\tilde{A}, \tilde{C})$  that is observable.

**Remark 3.7.** Note that  $\text{rank}(\tilde{A}) \leq \text{rank}(A)$ . This upper bound is also known as structural or generic rank.

A possible interpretation to Definition 3.5 based on graph theory is that two pairs  $(A_0, C_0)$  and  $(A_1, C_1)$  are of the same structure if their corresponding graphs share the same structure, i.e. the same set of nodes  $\mathcal{V}$  and edges  $\mathcal{E}$ , although the edge weights do not need to share the same values—provided that they are different from zero.

**Remark 3.8.** Definition 3.5 is grounded on the assumption that, in real applications, the true entries of  $(A, B, C, D)$  are usually uncertain, while zero entries are somewhat guaranteed (Lin, 1974). Thus, if a system is structurally observable, then it is observable for a wide range of parameters except for a *proper algebraic variety* in the parameter space which renders it unobservable (Liu et al., 2011b).

**Remark 3.9.** Note that Lin’s definition of structural observability is structural in two senses: (i) it is a crisp definition, as detailed in Section 3.1; and (ii) it is independent of the specific entries of  $(A, B, C, D)$ —relying only on the fact that zero entries are specifically known. “Structural observability (controllability) in Lin’s sense” (Definition 3.6) is only a *necessary* condition for “observability (controllability) in Kalman’s sense” (Definition 3.1).

The structural observability definition proposed by Lin (1974) has a strong graph-theoretical interpretation. In order to understand Theorem 3.3 (and later Theorem 4.3), we present some required definitions.

The *corresponding graph of a dynamical system*  $(A, B, C)$  is denoted by  $\mathcal{G}(A, B, C) = \{\mathcal{V}, \mathcal{E}\}$ , where  $\mathcal{V} = \mathcal{X} \cup \mathcal{U} \cup \mathcal{S}$  and  $\mathcal{E} = \mathcal{E}_{\mathcal{X}} \cup \mathcal{E}_{\mathcal{U}} \cup \mathcal{E}_{\mathcal{S}}$ . Nodes sets are the set of state variables  $\mathcal{X} = \{x_1, \dots, x_n\}$ , set of input variables  $\mathcal{U} = \{u_1, \dots, u_p\}$ , and set of output

variables  $\mathcal{S} = \{y_1, \dots, y_q\}$ . An edge  $(x_i, x_j)$  (directed arrow from  $x_j$  to  $x_i$ ) is an element of  $\mathcal{E}_X$  if  $A_{ij}$  is a free parameter entry of structured matrix  $A$ , an edge  $(x_i, u_j)$  is an element of  $\mathcal{E}_U$  if  $B_{ij}$  is a free parameter entry of structured matrix  $B$ , and an edge  $(y_i, x_j)$  is an element of  $\mathcal{E}_S$  if  $C_{ij}$  is a free parameter entry of structured matrix  $C$ . For brevity sake, when studying the observability property of a pair  $(A, C)$  we refer to the corresponding graph simply as  $\mathcal{G}(A, C) = \{\mathcal{X} \cup \mathcal{S}, \mathcal{E}_X \cup \mathcal{E}_S\}$ . Likewise for the pair  $(A, B)$  and  $\mathcal{G}(A, B) = \{\mathcal{X} \cup \mathcal{U}, \mathcal{E}_X \cup \mathcal{E}_U\}$ .

**Definition 3.7.** *A subset of nodes  $\mathcal{V}' \subseteq \mathcal{X}$  has a dilation<sup>7</sup> in a corresponding graph  $\mathcal{G}(A, C)$  if and only if  $|T(\mathcal{V}')| < |\mathcal{V}'|$ , where  $T(\mathcal{V}')$  is the set of all nodes  $v_i \in \mathcal{X} \cup \mathcal{S}$  with the property that there is a direct edge from a node in  $\mathcal{V}'$  to  $v_i$ .*

The following theorem derives from this graph approach to structural observability.

**Theorem 3.3.** *(Lin, 1974, See equivalent theorem and proof for controllability) The pair  $(A, C)$  is structurally observable if and only if the corresponding graph  $\mathcal{G}(A, C)$  satisfies both of the following conditions:*

1. every state  $x_i \in \mathcal{X}$  has a path to some output  $y_i \in \mathcal{S}$ ;
2.  $\mathcal{G}(A, C)$  has no dilations.

**Example 3.3. Structural observability in Lin's sense.**

Let a linear system (2.1), or a pair  $(A, C)$ , be expressed as structured matrices

$$A = \begin{bmatrix} \star & \star & 0 \\ \star & \star & 0 \\ \star & \star & \star \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 0 & \star \end{bmatrix}. \quad (3.21)$$

Following the definitions above, a corresponding graph of (3.21) can be drawn as shown in Fig. 3.1. Since every node has a self-edge,  $\mathcal{G}(A, C)$  has no dilations. Moreover, it is possible to reach  $y_1$  from all nodes  $\mathcal{X} = \{x_1, x_2, x_3\}$ . Thus, we note that the graph in Fig. 3.1 is structurally observable according to Lin's definition.  $\triangle$

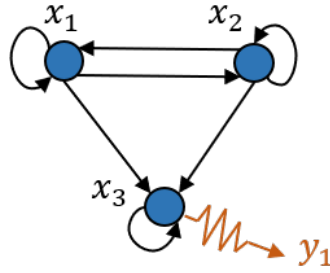
### 3.3.2 Maximum matching algorithm

Grounded on the structural and topological definition of controllability proposed by Lin (1974), Liu et al. (2011b) presented a pioneering work for controllability of *complex networks*<sup>8</sup>. The main goal is twofold: (i) to identify the *minimum set of driver nodes*

<sup>7</sup>Note that the definition of dilation here is dual to Lin's definition Lin (1974), since here we focus on the structural observability property rather than controllability.

<sup>8</sup>Although the main focus of this work is on observability, the review of Liu et al. (2011b)'s proposal for controllability is still relevant due to duality.





**Figure 3.1:** Graph representation of (3.21).

$\mathcal{D} = \{u_1, \dots, u_p\}$ , i.e. input signals  $\mathbf{u}$ , which can steer a (linear) network system entire state; and (ii) to understand the relations between controllability and the complex network (topological) properties.

Liu et al. (2011b) argued that other pioneering works on controllability of network systems, e.g. (Lombardi and Hörnquist, 2007; Rahmani et al., 2009; Tanner, 2004), are based on a weak assumption that, in a network system, the topology and nodal dynamics are entirely known. This assumption allowed previous works to explore the spectral graph properties of a network, such as the spectrum of the Laplacian matrix (Rahmani et al., 2009). However, even in the face of recent developments in modeling of complex networks, the accurate estimation of edge weights is not quite realistic yet. Indeed, in the case of biological or social networks, not even the nodal dynamics are fully known.

Thus, in order to study controllability of complex networks, Liu et al. (2011b) turned to the topological and structural controllability definition of Lin (1974) since: (i) it has a convenient interpretation grounded on a theoretical graph approach, which is very useful when the network topology can be established; and (ii) following Remark 3.8, Lin's structural controllability is not sensitive to parameter fluctuations, also a convenient feature since parameter estimation is often unreliable in large network systems.

Indeed, Liu et al. (2011b) show that Lin's structural controllability problem maps into an equivalent graph problem where one can gain full control over a directed network<sup>9</sup>  $\mathcal{G}(A, B)$  if and only if each unmatched node is directly connected to a driver node, and there are direct paths from any input signal to all matched nodes. A *matching* is formally defined as:

<sup>9</sup>Definition of  $\mathcal{G}(A, B)$  is analogous to that of  $\mathcal{G}(A, C)$ , where  $\mathcal{G}(A, B)$  is the corresponding graph of pair  $(A, B)$ .

**Definition 3.8.** (*Liu et al., 2011b, Definition 8 of Supplementary Information*) An edge subset  $\mathcal{E}_M$  is a matching if no two edges of  $\mathcal{E}_M$  share a common starting node or a common ending node. A node is matched if it is an ending node of an edge in the matching. Otherwise, it is unmatched.

This leads to the following theorem.

**Theorem 3.4.** (*Liu et al., 2011b, Theorem 2 of Supplementary Information*) The minimum number of driver nodes  $n_D$  needed to render  $\mathcal{G}(A, B)$ , or the pair  $(A, B)$ , controllable, is defined by

$$n_D = \max\{m - |\mathcal{E}_M|, 1\} \quad (3.22)$$

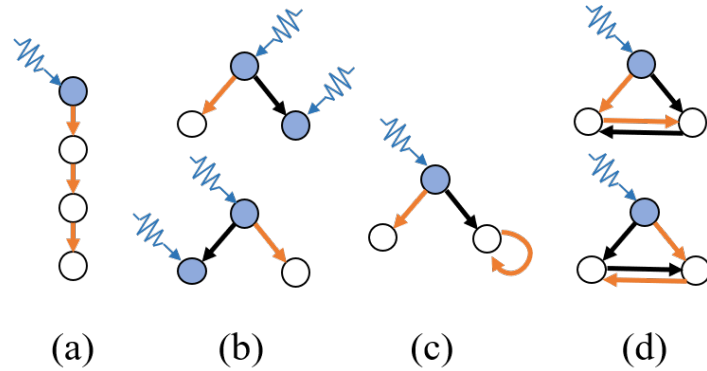
where  $m$  is the number of nodes in  $\mathcal{G}(A, B)$ . If there is a perfect matching in  $\mathcal{G}(A, B)$ ,  $n_D = 1$  (i.e.  $|\mathcal{D}| = 1$ ). Otherwise,  $n_D$  equals the number of unmatched nodes with respect to any maximum matchings (i.e.  $\mathcal{D}$  is composed by the unmatched nodes).

**Remark 3.10.** Adding more edges to  $\mathcal{G}(A, B)$  will never weaken a system structural controllability by Definition 3.6, which is not necessarily true for Definition 3.1. This feature makes Theorem 3.4 meaningful in dealing with missing links in network topology modeling (Liu et al., 2011b).

**Remark 3.11.** Differently from a brute force search for a minimum  $\mathcal{D}$ , which is of order  $\mathcal{O}(2^N)$ , the maximum matching algorithm allows  $\mathcal{D}$  to be identified with at most  $\mathcal{O}(\sqrt{m}|\mathcal{E}|)$  steps (Liu et al., 2011b)—a highly scalable algorithm to solve the minimum driver (sensor) placement problem.

From these results, Liu et al. (2011b) reached several conclusions on the controllability of complex networks. The most interesting ones are: (i) controlling heterogeneous and sparse networks is harder than controlling homogeneous and dense ones; and (ii) the counter-intuitive notion that driver nodes tend to avoid high-degree nodes. Some other conclusions, such as the correlation between the  $n_D$  and the network degree distribution are arguable considering the applied methodology and assumptions (Cowan et al., 2012). Section 3.4 criticizes and discusses the main results derived from this framework.

Figure 3.2 illustrates, with some simple network examples, the available choices of  $\mathcal{D}$  that render a network structurally controllable in Lin’s sense via the maximum matching algorithm. This technique is readily available for use in the MATLAB-based “NOCAD” toolbox (Leitold et al., 2019).



**Figure 3.2:** Examples of maximum matching in simple networks. Set  $\mathcal{E}_M$  is composed of orange edges. Unmatched and matched nodes are represented, respectively, in blue and white colors. To render the graph structurally controllable, all unmatched nodes must receive an input signal. (a) In a direct path, a network is controllable from the top node. (b) In this directed star, there is two sets  $\mathcal{E}_M$  with the same minimum cardinality, thus the network is controllable for two different configurations of driver placement. (c) The addition of a self-edge removes the necessity for an additional driver node in the same network topology of (b). (d) The addition of a bidirectional edge has the same effect of (c).

### 3.3.3 Nonlinear dynamical systems

In a later work, Liu et al. (2013) proposed a means to determine the topological observability of complex networks in the context of nonlinear polynomial networks (e.g., chemical reactions). The work is also grounded on Lin’s definition of observability and motivation is similar to the one stated in Section 3.3.2, whereas the main goal of Liu et al. (2013) is to determine the *minimum set of sensor nodes*  $\mathcal{S} = \{y_1, \dots, y_q\}$  which can render the system topologically observable.

Entitled *graph approach* (GA), Liu et al. (2013) proposed a procedure to guarantee the *topological observability* of a network:

1. draw an “inference diagram”, a nonlinear graph  $\mathcal{G}$ , from (2.2) according to Section 2.3<sup>10</sup>;
2. transpose the adjacency matrix of  $\mathcal{G}$  (i.e. invert the edge directions);
3. decompose  $\mathcal{G}$  in *strongly connected components* (SCCs) (see Section 2.2);
4. determine the *root SCCs* (i.e. a SCC with no incoming edges);
5. place a sensor node  $y_i$  in at least one node of each root SCC.

<sup>10</sup>Note that this inference diagram  $\mathcal{G}$  drawn from (2.2) does not distinguish between linear (solid) and nonlinear (dashed) interconnections. The problem of singularities and vanishing nonlinear interconnections is explored by Letellier et al. (2018), and further illustrated in Example 3.4.

If no nodes of a root SCC are observed, the network is unobservable, because one or more columns of  $\mathcal{O}(\mathbf{x})$  in (3.10) are composed of only zero entries and  $\text{rank}(\mathcal{O}(\mathbf{x})) < n$ . A more physical interpretation is that if there is no path from a given node to a sensor node (which always happens if no node of a root SCC is a sensor), then the information from this state cannot be inferred from any existing sensor nodes (Liu et al., 2013).

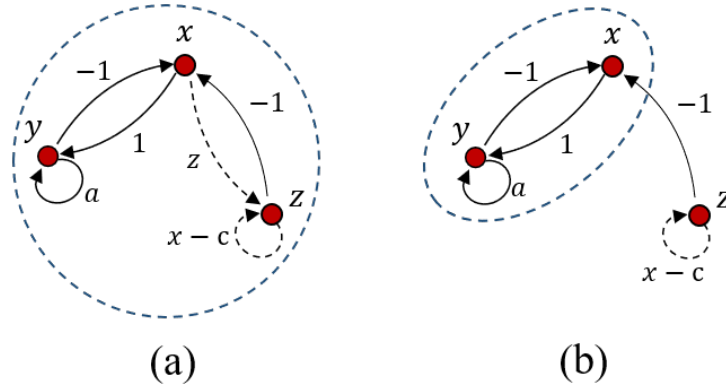
**Remark 3.12.** A graph is observable from a single sensor node if and only if the graph  $\mathcal{G}$  has only one SCC. A sufficient and necessary condition for this is that the corresponding adjacency matrix  $A_{\text{adj}} \in \mathbb{R}^{N \times N}$  is irreducible. A real non-negative  $A_{\text{adj}} \in \mathbb{R}^{N \times N}$  is irreducible if and only if  $(I + A_{\text{adj}})^{N-1} > 0$ .

**Remark 3.13.** The procedure is only a *necessary condition* for observability of nonlinear polynomial systems. If all sensor nodes selected from GA are measured, then  $\mathcal{O}(\mathbf{x})$  has no zero columns (Liu et al., 2013). Nevertheless, there is no guarantee that  $\mathcal{O}(\mathbf{x})$  columns are linearly independent, i.e. that  $\mathcal{O}(\mathbf{x})$  is full rank in the sense of Theorem 3.2.

Following Remark 3.13, based on an empirical analysis of multiple randomly generated chemical reactions, Liu et al. (2013) argue that the probability of correlation among the columns of  $\mathcal{O}(\mathbf{x})$  is rather small, if not zero, due to the “complicated polynomials” entries of  $\mathcal{O}(\mathbf{x})$ . This is analogous to Remark 3.8 on topological observability, where it is said that the proper algebraic variety of the parameter space that renders the system unobservable is comparatively small. However, the presence of symmetries in a dynamical network can render the system unobservable (Whalen et al., 2015), leading GA to underestimate  $\mathcal{S}$ . This is addressed by Liu et al. (2013) when comparing the lower bound of  $\mathcal{S}$  for topological observability provided by GA to the one provided by the maximum matching algorithm.

It must be mentioned that, although the probability that two columns of  $\mathcal{O}(\mathbf{x})$  are (exactly) linearly dependent is rather small, the columns of  $\mathcal{O}(\mathbf{x})$  might be *almost linearly dependent* nonetheless—leading to a rather small coefficient of (dynamical) observability  $\delta$ . Indeed, since the GA method is based on a structural classification of observability, it cannot identify the *best* sensor placement inside a root SCC, only which sets of sensor nodes are able to render the network topologically observable.

As a concluding remark, we highlight that GA is a methodology developed exclusively for nonlinear systems described by polynomial functions and, therefore, is



**Figure 3.3:** Root SCC (dashed circle) of a Rössler system graph. (a) All edges are nonzero. (b) For  $z = 0$ , the nonlinear edge vanishes.

not directly<sup>11</sup> applicable to other nonlinear systems, such as networks of Kuramoto oscillators.

**Example 3.4. Topological observability.** We illustrate a potential failure of the GA method to classify the topological observability of a polynomial nonlinear system, the Rössler system (Letellier et al., 2018).

As detailed in Section 2.3, let the Rössler system be represented by a nonlinear graph (see Fig. 2.4). Clearly, if  $z$  tends to 0 in (2.12), then the nonlinear edge connecting  $x$  to  $z$  vanishes. Figure 3.3 represents the root SCC of a Rössler system graph when the nonlinear edge is still present and after vanishing. Following GA method, in Fig. 3.3a, any state ( $x$ ,  $y$  or  $z$ ) could be chosen as a sensor node in order to render the network topologically observable. However, if  $z = 0$  one edge vanishes and  $z$  is no longer part of the root SCC. Hence only the  $x$  and  $y$  variables could be chosen as sensor nodes in order to render the network topologically observable.

This counter-example shows the importance of considering the singularity effects of nonlinear systems when determining the potential root SCC, especially when the system operates in the vicinity of this point of singularity. Indeed, the “vanishing” of the nonlinear edge connecting  $x$  to  $z$  is the cause of the poor observability of the  $z$  variable of Rössler system, as discussed in previous examples.  $\triangle$

<sup>11</sup>There are “universal” representations for nonlinear systems as polynomial systems (i.e. polynomial vector fields) at the expense of augmenting the number of states and considering only predefined initial conditions (“consistent” initial conditions) (Kerner, 1981; Ohtsuka, 2005).

### 3.4 Analysis of Related Works

This section reviews the extensions and criticism in the literature regarding the aforementioned methods of *topological observability*. Most earlier discussions in the study of controllability and observability of network systems focused mainly on controllability rather than observability. However, due to the duality of these properties, there is no harm or loss of generality in comparing and discussing methods designed for controllability or observability of linear dynamical systems.

In this review, we are more interested in a broader class of works that propose metrics designed for *generalized networks*—that is, networks with no specific class of network topology or nodal dynamics (only linear and nonlinear distinctions). Nevertheless, several works in the literature explore the relation of observability and the graph properties of certain types of network topologies, including Cartesian grid graph (Notarstefano and Parlange, 2013), chain and cycle graph (Parlange and Notarstefano, 2012), clustered networks (Ruths and Ruths, 2014), and specific complex network models, such as the scale-free network (Fu et al., 2016). The discussion of observability has also been directed to networks of specific dynamics and applications, such as Boolean networks (Chen and Qi, 2009; Gates and Rocha, 2015; Laschov et al., 2013), chemical reactions (Liu et al., 2013), traffic networks (Castillo et al., 2008), biological systems (e.g. neuronal networks (Gu et al., 2015; Su et al., 2017)) and power systems (Baldwin et al., 1993; Monticelli and Wu, 1985).

**Lin’s topological definition of controllability.** Lin (1974)’s definition of structural controllability (see Sec. 3.3.1) allows an intuitive analysis of a given *linear* dynamical system from its corresponding graph representation. This approach, which we refer to as topological controllability (observability), is not concerned with the specific entries of system matrices  $(A, B, C, D)$  such as the Kalman rank condition in Theorem 3.1, but rather if those matrices present a structure that *might* allow controllability under a correct and arbitrary choice of parameters. Lin argues that, in a mathematical model of a real process, the parameters estimations are contaminated by uncertainties whereas “zero” entries are practically guaranteed. Thus, a first step towards deciding if a system is controllable is to establish if it is structurally and topologically controllable.

**Liu and coworker’s controllability of complex networks.** Liu et al. (2011b), motivated by the fact that complex networks often have a reliable topological (graph) representation but an unreliable estimation of edge weights, took advantage of Lin’s topological approach to investigate controllability in complex networks. Based on the maximum matching of the corresponding graph, Liu et al. (2011b) identify the

“minimum” set of driver nodes  $\mathcal{D}$  to render a complex network structurally controllable. However, Lin’s definition of structural controllability is a *crisp* definition, so how can one assure that this  $\mathcal{D}$  provided by the maximum matching is really the *best* set? This is specially true since the maximum matching set of a graph is not necessarily unique. It is shown that Liu and coworkers’ controllability (Liu et al., 2011b) and observability (Liu et al., 2013) methods underestimate the required set of driver (Gates and Rocha, 2015; Leitold et al., 2017; Wang et al., 2017) and sensor (Haber et al., 2018; Montanari and Aguirre, 2019) nodes—mainly because they do not consider the specific entries of  $(A, B, C, D)$ . Perhaps a more relevant question rather than if a network is structurally controllable is if it is almost uncontrollable (Cowan et al., 2012; Friedland, 1975).

**Control via “control hubs”.** A fundamental result of Liu et al. (2011b) is that heterogeneous and sparse networks are harder to control than homogeneous and dense ones. This result is based on an analysis of correlation between the number of driver nodes  $n_{\mathcal{D}}$  required for controllability and the network degree distribution, which was further discussed by Pósfai et al. (2013). This led to a counter-intuitive notion that high-degree nodes, also called *hubs*, are less desirable to be driver nodes (Liu et al., 2011b). As discussed by Nepusz and Vicsek (2012); Slotine and Liu (2012), this is a consequence of the fact that, since network models in (Liu et al., 2011b) do not consider nodal dynamics, the control signal injected by driver nodes spread homogeneously among its neighbouring nodes, raising symmetries that restrict the state-space exploration.

Nepusz and Vicsek (2012) show that control by nodes of high-degree is also possible if a different paradigm is taken: to change the analysis from nodal dynamics to *edge dynamics* (see also (Pang et al., 2017)). The argumentation follows that by choosing a hub node as a driver, if one can control its edge dynamics individually instead of its nodal dynamics, then the spread of control signals no longer suffers from symmetry issues. In case of controlling edge dynamics, homogeneous and dense networks become harder to control than heterogeneous and sparse ones. Moreover, one can benefit from controllability metrics (and other network metrics) designed for nodal analysis by performing a transformation from nodal dynamics representation to edge dynamics (that is, drawing a *line graph* from the original graph). This approach reduces the number of driver nodes in exchange for a higher control energy cost per driver node.

**Influence of nodal dynamics.** Contradicting some claims in (Liu et al., 2011b; Nepusz and Vicsek, 2012), Cowan et al. (2012) affirm that the minimum number of driver nodes is not mostly dependent on node degree distributions (at least in linear networks), but rather on the underlying nodal dynamics of the studied system. Indeed, in the network modelling of Liu et al. (2011b), the individual nodes of the studied

benchmarks show no independent behaviour, acting as pure integrators if the network were fully disconnected. This is a consequence of Liu et al. (2011b) not including the presence of *self-edges*, which are a fundamental part of a network dynamics. Consider a linear dynamical network (2.1), rewritten as

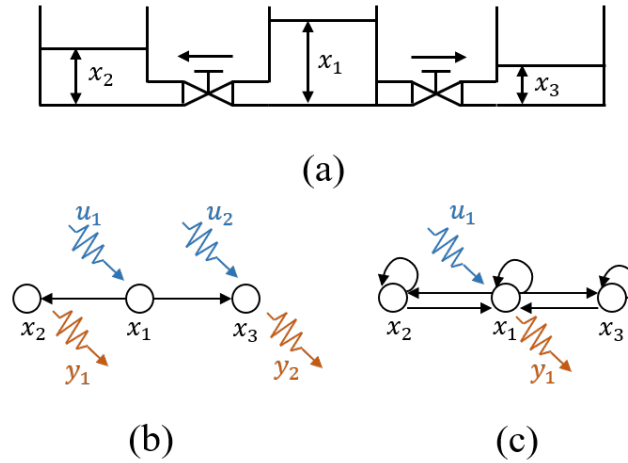
$$\dot{x}_i = \lambda_i x_i + \sum_{j=1}^m a_{ij} x_j + \sum_{j=1}^p b_{ij} u_j, \quad (3.23)$$

for  $i = 1, \dots, N = m$ , where  $x_i$  is the state at node  $v_i$  (only 1-dimensional systems are considered at each node). A simplified form of (2.5), equation (3.23) highlights that, in a dynamical network, each node has its dynamics described not only by its neighbouring interactions (which includes a potential self-edge  $a_{ii}$  related to the network topology and usually translated in the Laplacian matrix, if  $a_{ij} \neq 0$ ), but also by its own independent dynamical behaviour—given by an eigenvalue  $\lambda_i$  that determines its individual time constant (when absent of external influences).

If all nodes of a dynamical network include self-edges, then all nodes are matched according to Definition 3.8. The network is referred as perfectly matched and, according to Theorem 3.4, the number of driver nodes required to render it structurally controllable is *one*—as long as this unique driver node is attached to *all nodes* (Cowan et al., 2012). Once more, hub nodes are shown to be fundamental for network control, despite demanding higher control energy costs. Although a single control input might be sufficient to (theoretically) render a network structurally controllable, one should not expect it to be feasible from a practical point-of-view as the system dimension increases. This is related to the problem of dynamical controllability (observability).

**Benchmarks studied in (Liu et al., 2011b).** A common topic of criticism to Liu et al. (2011b)'s work is that the networks used as benchmarks for the proposed maximum matching method (see (Liu et al., 2011b, Table 1)) were analyzed on an exclusively topological level while the internal states that describe the nodal dynamics were disregarded. Consider the tank system example in Fig. 3.4, as originally discussed in Leitold et al. (2017). Fig. 3.4b shows a minimum set of driver nodes given by a maximum matching search if a network model is purely represented on a topological level, as Liu et al. (2011b) did in most of their benchmark models. Fig. 3.4c, on the other hand, shows this same outcome if a network model is built from the dynamical matrix that describes the underlying process, according to Aguirre et al. (2018); Cowan et al. (2012); Leitold et al. (2017). Note that when including the network dynamics, only one driver and sensor is (theoretically) required, a result in line with the argument raised by Cowan et al. (2012).





**Figure 3.4:** Adapted from (Leitold et al., 2017). (a) Tank system. (b) Liu and coworker’s representation based exclusively on the network topology, which is given by the water flow in the tank system. (c) Dynamical network representation where both the network topology and dynamics are taken into account.

Results pointing to a correlation between node degree distributions and (structural) controllability, therefore, can be misleading since the benchmark studied on (Liu et al., 2011b, Table 1) *do not* represent dynamical processes at all. In fact, Fig. 3.4 is only a simple example that, if the network dynamics were considered when dealing with systems such as food webs, regulatory networks, power grids, electronic circuits, neuronal networks and metabolic systems, it is expected that the minimum set of driver nodes would be significantly different from the results found by Liu et al. (2011b). Moreover, specially for strongly connected and diffusively coupled networks, a maximum matching search can lead to a single sensor node being capable of rendering the *whole* dynamical network structurally observable, even if it has hundreds or thousands of nodes. Once again, this kind of result questions the feasibility of structural controllability and observability approaches to the control and state estimation of such types of real-world networks. A more in-depth discussion is presented in Section 3.6

**Further works related to topological observability.** Despite the criticism regarding the limitations of using graph approaches to determine the controllability or observability of a network system, a lot of effort has been devoted to this front for three main reasons: i) the plainness from which the graph approach proposed by Lin (1974) can be extended to the context of network systems, ii) the high scalability of such approach to large-scale networks with thousands of nodes, as seen in Liu et al. (2011b), and iii) the fact that the recent and relevant studies established several conclusions on the effects of network structure to the controllability and observability of a system.

Numerous extensions to Liu and colleagues' work led to results on target control (Gao et al., 2014; Gutiérrez et al., 2012; Jia et al., 2013; Nacher and Akutsu, 2013), study of correlations between controllability and network properties (Li et al., 2017; Pósfai et al., 2013), analysis of the relationship between control energy and the chosen set of driver nodes (Yan et al., 2015), novel controllability methods based on graph properties (Leitold et al., 2017), robustness analyzes to cascading failures (Pu et al., 2012), and recent developments in network control of neuronal networks (Gu et al., 2015; Su et al., 2017).

**Lack of validation.** A common problem in studies involving network systems and novel controllability and observability proposals is the lack of an “independent” validation. Many algorithms and techniques are applied to network databases and compared to other metrics, leading to conclusions regarding which method provides the smaller set of driver (sensor) nodes (Liu et al., 2011b, 2013; Nepusz and Vicsek, 2012; Yuan et al., 2013). However, the relevance of the provided set of driver nodes is not usually questioned. Is it the smaller set of driver nodes really *better* than the larger one? In which sense should “better” be understood? This question is a matter resolved by dynamical observability metrics, although most, if not all, are only applicable to systems of lower dimensions.

Using a first-order (linear) electronic circuit interconnected by a chain graph as a validation benchmark, Wang et al. (2017) gave attention to this topic when studying the practical feasibility of two methods to determine the minimum set of driver nodes: the maximum matching algorithm proposed by Liu et al. (2011b), and the “exact” controllability method proposed by Yuan et al. (2013). The authors noticed that, when applying a single control signal to one of the chain extremities, the longer the “control chain”<sup>12</sup>, the closer to being singular was the *controllability Gramian* (dual to (3.4)). This is an expected result since, as mentioned in Section 3.2.1, the smallest and largest eigenvalues of the controllability Gramian are related to the maximum and minimum energy costs (also known as control energy) of driving a system state through the state-space. Indeed, in order to compare the practical (physical) capability of a set of driver nodes to control a dynamical system, Wang et al. (2017) investigated the conditioning number of the controllability Gramian (Definition 3.3) conveyed by each set of driver nodes. This is, essentially, a dynamical observability approach to this problem. One interesting result is that it is possible to raise the coefficient of

<sup>12</sup>The “control chain” here refers to the path of nodes between the driver node (the input signal) and the target node.

controllability by a slight addition of driver nodes along the network chain, “breaking” the long control chain into smaller sections.

Indeed, when validating the feasibility of a minimum set of driver to actually control the system state, a frequent conclusion is that controllability metrics based on graph-theoretical (topological) approaches, such as (Lin, 1974; Liu et al., 2011b), usually underestimates the minimum set of driver nodes (Gates and Rocha, 2015; Leitold et al., 2017; Wang et al., 2017). This conclusion is also present in the context of observability, where a minimum set of sensor nodes determined by the GA method (Liu et al., 2013) was shown to be insufficient to provide a reliable estimation of the system states when using Bayesian filtering techniques (Haber et al., 2018; Montanari and Aguirre, 2019).

An interesting approach to validate controllability metrics is to use boolean networks, since they highlight the interaction between the network topology and nonlinear dynamics involving simple binary variables (Gates and Rocha, 2015). Moreover, Gates and Rocha (2015) show that the controllability predicted by the maximum matching method might fail even for (linearized) small nonlinear examples. Likewise, Aguirre et al. (2018) show that the structural observability defined by GA is susceptible to failures if the procedures do not take into account the nonlinearity of the edges (Aguirre et al., 2018; Letellier et al., 2018)—as discussed in Example 3.4.

## 3.5 Future Research Directions on the Dynamical Observability of Network Systems

Most *topological* observability methods are mainly concerned with distinguishing which set of sensor nodes renders a network system observable. Another important goal is to quantify if a given set of sensor nodes renders a network more or less observable than another set—noting that the network is observable from both sets—and how this quantification is related to practical purposes (trajectory reconstruction, state estimation, differential embedding, and so on). This is a matter of *dynamical* observability, which still is an open problem in the literature.

A natural approach, for instance, to measure the degree of observability of a given set of sensor nodes is to extend or simply apply the dynamical observability metrics discussed in Section 3.2 to network systems. Indeed, this approach was pursued in some works in the literature. Based on the smallest eigenvalue of the Gramian (see Section 3.2.1), Yan et al. (2012) focused on the study of scaling laws for the control energy of network systems as a function of the control horizon, while Pasqualetti et al.

(2013b)—supported by the findings of Sun and Motter (2013)—studied the trade-offs between control energy and the number of control nodes. Gu et al. (2015) directed their study to a neuronal network application, and Bof et al. (2017) demonstrated a relation between the controllability of a network and its eigenvector centrality<sup>13</sup>, showing that it is harder to control a network whose nodes have a similar centrality degree. Another interesting result is the report of a trade-off between the controllability of complex networks and its resilience to perturbations and failures (Pasqualetti et al., 2018; Zhao and Pasqualetti, 2019). Following Definition 3.4, Whalen et al. (2015) investigated the presence of symmetries in small motifs ( $|\mathcal{V}| = 3$ ) and its restrictions on the “state-space exploration”.

However, one of the reasons that led graph-inspired (topological) techniques discussed in Section 3.3 to dominate this field of work in place of matrix-theoretical ones is the high scalability of graph tools compared to those developed in control theory. Moreover, when dealing with driver (sensor) node selection, most developed solutions suffer from dimensionality issues, since they are either based on combinatorial or non-scalable optimization techniques, or heuristic approaches that are limited to the specific studied systems and show no guarantees of control (Pasqualetti et al., 2013b).

In light of this open problem, in this section, we investigate some interesting paradigms to quantify observability (controllability) of network systems in a computationally feasible way. In what follows, we briefly discuss five interesting alternatives to quantify observability (controllability) in network systems.

**(i) Network partitioning.** Formally, network partitioning consists of dividing the set of nodes  $\mathcal{V}$  of a given graph  $\mathcal{G} = \{\mathcal{X}, \mathcal{E}\}$  into  $P$  disjoint sets  $\mathcal{P} = \{\mathcal{X}_1, \dots, \mathcal{X}_P\}$ , where  $\mathcal{G}_i = \{\mathcal{X}_i, \mathcal{E}_i\}$  is the  $i$ th subgraph of  $\mathcal{G}$ , for  $i = 1, \dots, P$ .

Clearly, network partitioning methods (Fortunato, 2010) are a viable alternative in network systems to subdivide a high-order systems into several and, if possible, non-intersected “clusters” of lower dimension. Ideally, the low-order subgraphs could be assessed by traditional methods from control theory. In practice, however, to subdivide a network into independent systems, or even to uncover its remaining dynamical interdependences, might be a challenge. For instance, Liu et al. (2013) proposed a sensor selection method based on a network partitioning into SCC (Section 3.3.3).

---

<sup>13</sup>It must be noted, however, that this relation was established based on an assumption that the dynamic matrix  $A$  is non-negative, which is not generally the case when the nodal dynamics are taken into account. When studying combustion and biological networks, Haber et al. (2018) detected no clear correlation between the optimal selection of sensor nodes and the corresponding node centrality measures.

Desynchronized (partitioned) control has also been implemented by [Su et al. \(2017\)](#) to validate its experiments.

For instance, following the GA method, [Pasqualetti et al. \(2013b\)](#) proposed an elegant solution to actuator placement by choosing all nodes at each SCC boundaries<sup>14</sup> as driver nodes. Hence, the authors developed a control law strategy that *decouples* the SCC dynamical interdependences in such a manner that its selected “internal” driver nodes are solely responsible for the SCC steering from the origin to a target state. In fact, by “forcing” all the SCC to behave independently, the high-order network problem is reduced to independent low-order dynamical systems that can be controlled by local control centers. The authors argue that their method is scalable since it depends on the number of partitions rather than the network cardinality. By breaking the network in partitions with a sufficiently small cardinality, actuators (sensors) can be optimally placed in each partition such that a given dynamical controllability (observability) measure is maximized by brute-force search.

Although the aforementioned method, as well as other network partitioning methods, are tempting, highly centralized networks are not suitable for decomposition. In such cases, even if SCC can still be identified, their subgraphs might still be very high-order systems for traditional techniques of control theory.

**(ii) Set function optimization.** [Summers et al. \(2016\)](#) formulated the sensor<sup>15</sup> placement problems as a *set function* optimization problem as follows:

$$\max_{\mathcal{S} \subseteq \mathcal{X}, |\mathcal{S}|=q} J(\mathcal{S}), \quad (3.24)$$

where given a  $\mathcal{X} = \{x_1, \dots, x_n\}$ , the problem is to select a  $q$ -element subset  $\mathcal{S}$  of  $\mathcal{X}$  that maximizes an objective set function  $J(\mathcal{S}) : 2^n \rightarrow \mathbb{R}^1$ —i.e. a function that assigns a real number to each subset  $\mathcal{S}$ . The objective function desired to be maximized is chosen so that it describes the trade-off between the number of required sensor nodes and the related estimation energy costs. Possible functions are the dynamical observability metrics discussed in Section 3.2.1. As one might note, (3.24) is a combinatorial optimization problem that could only be solved by brute force search if the network dimension were of lower order.

The greatest contribution of [Summers et al. \(2016\)](#) is to show that most dynamical observability metrics based on the observability Gramian are *submodular* functions. Thus, although solving the optimization problem through brute-force search is compu-

---

<sup>14</sup>See ([Pasqualetti et al., 2013b](#)) for a mathematical definition.

<sup>15</sup>It was actually formulated in the context of controllability and optimal actuator placement.

tationally hard, if the objective function is submodular, then the optimization problem can be solved by a greedy algorithm with guaranteed performance (Nemhauser et al., 1978). Via numerical simulations, the authors show that, considering  $|\mathcal{V}| = 25$  and  $|\mathcal{S}| = 7$ , the greedy optimization displays a result better than 99.93% of all other combinations. The method is further validated on a power system model and scalability of the greedy algorithm is discussed in detail. Haber et al. (2018) apply, for comparison purposes, Summers and coworkers' approach to nonlinear networks and show that their method performs well, although they do not provide any mathematical proofs for the nonlinear case.

**(iii) Symbolic observability.** To deal with larger systems with more complicated dynamics, Bianco-Martinez et al. (2015); Letellier and Aguirre (2009) provided a dynamical symbolic observability metric that does not depend on the specific parameter entries but rather on the presence and quantity of linear, nonlinear and rational couplings within the dynamical system. The observability metric is normalized in a  $[0, 1]$  range, allowing one to compare different dynamical systems—which is not possible following Definition 3.4. This approach seems to be promising for dynamical networks as long as they are “not very large” (Letellier et al., 2018).

Some other interesting symbolic approaches to quantify observability have been presented in a power system analysis context (Bretas and London, 1998; Slutsker and Scudder, 1987). Indeed, the problem of sensor and actuator placement in power systems is a very relevant (and old) one due to emerging and expensive technologies designed to monitor or control the system states (Baldwin et al., 1993; Monticelli and Wu, 1985). An application example to power systems is given in Section 3.6.

**(iv) Indirect measures of “reconstruction quality”.** In an observability context, the degree of observability can be indirectly assessed through the “estimation quality” of the state. This is a more empirical approach. For instance, in a Bayesian filtering context, this is motivated by the fact that if the measured signals do not provide relevant information to the filter, i.e. they convey poor observability, then the update stage of the filter is impaired and, consequently, the estimates show poor performance. This is based on an assumption that the filters or algorithms are well-tuned (Montanari and Aguirre, 2019).

Examples found in the literature are the estimation error of a moving horizon estimation technique (Haber et al., 2018) or a particle filtering framework (Montanari and Aguirre, 2019), as well as the fitting error from training a Gauss-Newton algorithm (Guan et al., 2018) or a reservoir computer (Carroll, 2018). Due to the scalability issues of Bayesian filtering methods and similar techniques, this approach is not very

useful for networks with dimensionality  $N \gg 100$ . It is an interesting approach to further understand the interplay between observability, network topology and nodal dynamics (Guan et al., 2018; Montanari and Aguirre, 2019).

Note that all discussed observability metrics throughout the text depend on knowledge of the system equations. When such are not available, Aguirre and Letellier (2011) provided a data-based procedure to infer the degree of observability from a recorded time series. This approach relies on measuring the reconstruction quality from an embedding of the available time series as a function of specific properties associated with poor coefficients of observability, such as sharp folds and strong squeezing of trajectories as consequence of singularity issues.

**(v) Observability of a subset of nodes of interest.** As discussed, a recurring goal in the literature is the search of an optimal set of sensor nodes that renders a dynamical network fully observable. Motter (2015) argues that this goal is not quite realistic, though, since, from a practical point-of-view, to reconstruct (steer) the dynamical system trajectory of a high-dimensional system requires direct measurement (control) of most nodes in a network—which is not always feasible. An alternative, therefore, is to focus on a particular subset of nodes of interest and determine what is the required set of sensor nodes in order to render this subset observable. Basically, this approach revolves around reducing the dimensionality of the problem. This problem has been explored by Judice et al. (2019) in the context of node observability (controllability), i.e. focusing on a particular node observability, as well as by Gao et al. (2014), grounded on the definition of output controllability. In Chapter 4, we expand on the concept of functional observability, from control theory (Hieu and Tyrone, 2012; Jennings et al., 2011), to determine the graph-theoretical conditions under which it is possible to reconstruct the state of a particular subset of nodes with a reduced-order observer, thereby reducing computational costs in the state reconstruction of a large-scale system.

## 3.6 Application Examples

Throughout this chapter, we showed how the concepts of structural, dynamical and topological observability can be applied to a (low-dimensional) nonlinear dynamical system, the Rössler system. In the following, we provide an example of how these concepts are applied in the high-dimensional context of power grids and multi-agent coordination. Moreover, we also take this opportunity to show how the studied dynamical systems can be modelled in a network context.

It should be mentioned, however, that since the numerical computation of the observability matrix (3.10) of a nonlinear model is quite unfeasible in a high-dimensional setting, we focus on a linearized model of the nonlinear network dynamics around the operation point. In what follows, to determine the “minimum set of sensor nodes  $\mathcal{S}$ ” using the maximum matching search (Section 3.3.2), we use its MATLAB-based implementation found in the NOCAD toolbox (Leitold et al., 2019).

### 3.6.1 Power grids

**Model dynamics.** This section provides an application example in the IEEE power grid benchmark with 50 generators and 145 buses (Nishikawa and Motter, 2015; Vittal, 1992). Figure 3.5a illustrates the benchmark model as network system. In this context, the power grid dynamics are modelled as a network of interconnected Kuramoto oscillators (Arenas et al., 2008; Dorfler et al., 2013; Montanari et al., 2020), given by

$$\frac{2H_i}{\omega_R} \ddot{\phi}_i + \frac{D_i}{\omega_R} \dot{\phi}_i = P_i + \sum_{j=1, j \neq i}^m K_{ij} \sin(\phi_j - \phi_i + \beta_{ij}), \quad (3.25)$$

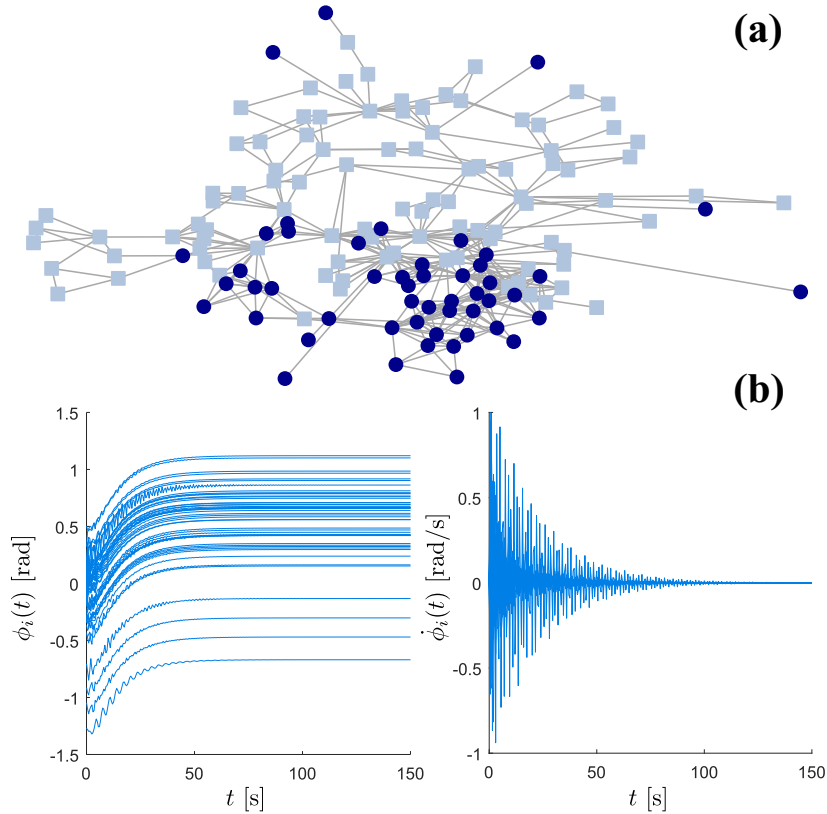
for  $i = 1, \dots, m$ , where  $m$  is the number of nodes,  $\phi_i(t)$  is the phase angle of oscillator  $v_i$  at time  $t$  relative to a frame that rotates at the reference frequency  $\omega_R$  rad/s,  $H_i$  and  $D_i$  are inertia and damping constants respectively,  $P_i$  is related to the power supply of generator at node  $v_i$ ,  $K_{ij}$  is the coupling weight related to the maximum power transfer capacity in the respective transmission line interconnection two nodes  $(v_i, v_j)$ , and  $\beta_{ij}$  is the corresponding phase shift.

All parameters are estimated from the power grid benchmark dataset provided by Vittal (1992) following the “effective network model” paradigm<sup>16</sup> described by Nishikawa and Motter (2015). While  $(H_i, D_i)$  are estimated from the generator constructive parameters,  $(P_i, K_{ij}, \beta_{ij})$  are inferred from the power grid steady-state distribution of power flow. Although the power grid benchmark has 145 buses, the power grid model in (3.25) is reduced, via Kron reduction (Sauer et al., 2017), to an effective network<sup>17</sup> with  $m = 50$  nodes, where each node  $v_i$  represents a generator whose dynamics are described by a second-order Kuramoto oscillator. The power grid dynamics are shown in Fig. 3.5b. If  $\dot{\phi}_i(t)$  converges to zero for all generators, then

<sup>16</sup>Different models of power grids based on networks of coupled oscillators are discussed in (Moreira and Aguirre, 2019; Nishikawa and Motter, 2015).

<sup>17</sup>Differently from the power grid structure shown in Fig. 3.5, due to the Kron reduction, the effective network model is fully connected (i.e. each generator is directly connected to every other generator in the network, with a coupling strength  $K_{ij}$  related to their corresponding power flow).





**Figure 3.5:** (a) IEEE power grid benchmark, where generator and load buses are represented by circle and square nodes, respectively, and transmission lines by edges. (b) Phase and instantaneous frequency time evolution. Linearization is performed around the equilibrium point at  $t = 150$  s.

the power grid is said to be fully synchronized. A MATLAB implementation of this framework is provided by [Nishikawa and Motter \(2015\)](#).

For the following computations, model (3.25) is linearized around its equilibrium point (computed through numerical integration after  $t = 150$  s, as in Fig. 3.5b). The linearized model is represented in its canonical form:

$$\begin{bmatrix} \dot{\phi} \\ \ddot{\phi} \end{bmatrix} = \begin{bmatrix} \mathbf{0} & I_m \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} \phi \\ \dot{\phi} \end{bmatrix}, \quad (3.26)$$

where  $\phi = [\phi_1 \dots \phi_m] \in \mathbb{R}^m$ ,  $\mathbf{0} \in \{0\}^{m \times m}$ ,  $A_{21} \in \mathbb{R}^{m \times m}$  is a fully dense matrix, and  $A_{22} \in \mathbb{R}^{m \times m}$  is a diagonal matrix.

**PMU placement.** One specific problem related to observability in the context of power grids is that of optimal placement of phasor-measurement units (PMU) ([Phadke and Thorp, 2008](#); [Yang et al., 2012](#)). A PMU placed on a generator bus (node) allows

real-time measurement of its voltage and line currents. Moreover, depending on the power grid structure, voltages and/or line currents of neighboring buses (nodes) can be determined from Kirchhoff's law. Indeed, in the scientific community of power systems (Cruz and Rocha, 2017; Peng et al., 2006; Tran and Zhang, 2018), the search for an optimal PMU placement such that the voltages of all buses can be determined is known as a “topological observability” problem<sup>18</sup>.

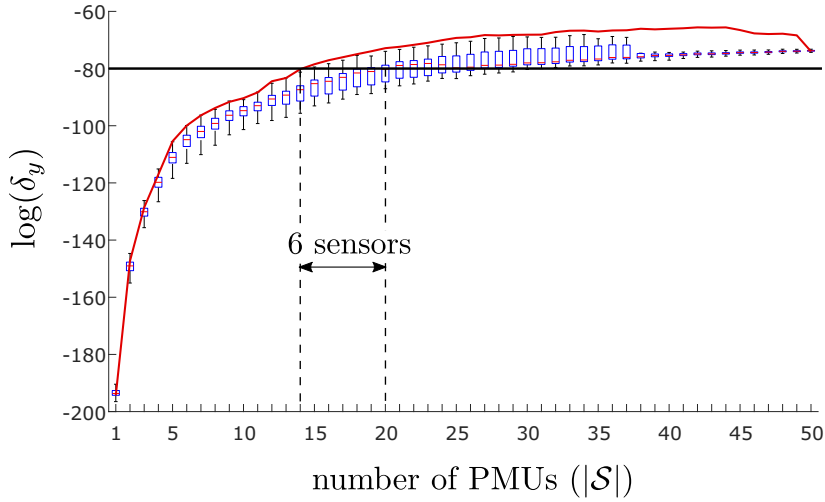
Two main reasons motivate this problem. On the one hand, a PMU is a very expensive equipment and it might not be economically viable to install one in every node of a power grid (Rocha et al., 2018). On the other hand, communication between two (geographically apart) areas might have been interrupted by some failure or malicious attack and one might need to estimate, from its accessible set of measurements, what is the state of some other generator connected to the power grid. Indeed, a reliable communication structure (which includes both direct and estimated measurements) is a very important concern for the implementation of Wide Area Control techniques, especially with the current transition of power grids to smart grids (Malik, 2013).

In the following, assume that if a PMU is placed in a generator node (bus)  $v_i$ , then one has access sufficient information to infer the respective generator states  $\{\phi_i, \dot{\phi}_i\}$ . In this sense, the optimal PMU placement can be framed as a (traditional) observability problem in the sense that one wants to estimate, from knowledge of the generator states where a given set of PMU is placed, the generator states of all other nodes that do not have a PMU. We use this example to counterpoise the different types of observability discussed in this manuscript, stating pros and cons in each case.

**Structural and topological observability.** Consider the linearized model (3.26) and that, if a PMU is placed on generator  $v_i$ , then node  $v_i$  is said to be a sensor node of  $\mathcal{S}$  and hence  $\{\phi_i, \dot{\phi}_i\} \in \mathcal{S}$ . Since  $A_{21}$  is a fully dense matrix whose individual entries  $a_{ij}$  are related to the power system parameters  $(K_{ij}, \beta_{ij}, H_i)$ , the probability that the columns of  $\mathcal{O}$  are linearly dependent is practically zero (as pointed out by Remark 3.8). This is specially true if we consider that the generator parameters and power supply levels are not homogeneous along the power grid (which is true in reality). Thus, the dynamical network (3.26) is topologically observable (in Lin's sense) from

---

<sup>18</sup>Although this problem shares the same nomenclature discussed throughout this paper, it has a different meaning. “Topological observability of power systems” is a solvability problem where one wants to determine, based on Kirchhoff's law, the whole vector of voltages and line currents from knowledge of the admittance matrix and some measurements available by a given set of PMUs. “Topological observability of dynamical systems”, on the other hand, is graph-theoretical way to certify if the basic conditions for the design of a stable linear observer can be satisfied.



**Figure 3.6:** Coefficient of observability  $\delta_y$  per number of PMUs (sensor set cardinality  $|\mathcal{S}|$ ), considering an optimal sensor placement (solid red line) and 1,000 Monte Carlo runs with random placements (boxplot).

any generator node (implying that  $|\mathcal{S}| = 1$  is a sufficient and necessary condition). Likewise, it is structurally observable (in Kalman’s sense) from any node.

This illustrates the discussion in Section 3.4 that, when nodal dynamics are taken into account, self-edges are included in the graph representation, leading to a trivial solution where all nodes are matched and thus only one sensor node (placed anywhere) is required to render the network observable. In this case, self-edges are only represented in states  $\ddot{\phi}_i$  (due to the diagonal block  $A_{22}$ ). Nevertheless, since there is a bidirectional connection between state variables  $\dot{\phi}_i$  and  $\ddot{\phi}_i$ , for all  $i$ , then the conclusion remains: every node is matched. This is in line with the examples in Fig. 3.2.

**Dynamical observability.** Although, in the linear case, any node is sufficient as a sensor node to render the power grid topologically or structurally observable, one might argue that, for practical purposes,  $\mathcal{S}$  is almost unobservable. Indeed, if  $|\mathcal{S}| = 1$ , then  $\delta_y \approx 10^{-85}$  (regardless of where the PMU is placed). Hence topological and structural observability approaches do not provide any relevant insight into the problem of optimal sensor placement *in this case*.

Assessing the dynamical observability of a power system, on the other hand, might not only be useful to determine the minimum number of sensor nodes required to render the network observable from a practical point-of-view, but also to assess the optimal placement. Figure 3.6 shows how the coefficient of observability  $\delta_y$  increases

with the number of PMUs considering an optimal<sup>19</sup> and multiple random placements. The optimal PMU placement is based on the set function optimization problem (3.24), which is revisited in Appendix A.1. We also provide a MATLAB implementation of a greedy algorithm to solve (3.24) at <https://doi.org/10.13140/RG.2.2.22524.28803/1>, as proposed by Summers et al. (2016).

It is interesting to see that the network dynamical observability  $\delta_y$  greatly benefits from the first few sensor nodes (approximately 20% of the network cardinality), before reaching a stationary value (around half the network cardinality) where further sensor nodes do not increase  $\delta_y$ . An interesting interpretation, discussed by Guan et al. (2018), is that as the number of sensor nodes increases, useful and relevant data about the system dynamics is acquired by the output measurements until a turning point where additional sensors provide redundant information about the system dynamics. This sudden increase of the network observability with the first few choices of sensor nodes is in line with results in (Guan et al., 2018; Qi et al., 2015; Summers et al., 2016).

Not only the network dynamical observability improves with the addition of sensor nodes, but it can also benefit from an optimal placement. Figure 3.6 shows the efficacy of the framework put forward by Summers et al. (2016), where the greedy algorithm approach to sensor placement provides an optimal performance clearly above random placements of the same sensors. For instance, to reach a given degree of observability (e.g.,  $\log(\delta_y) \approx -80$ ), 14 PMUs are needed if they are *optimally* placed over the power grid. On the other hand, a *random* placement requires 20 PMUs to reach (with a  $\sim 50\%$  chance) the desired degree of observability.

This is especially important since a sensor placement that conveys a higher coefficient of observability  $\delta_y$  is shown to be related to a better state estimation (Montanari and Aguirre, 2019; Singh and Hahn, 2005). Indeed, analogous to the results exposed in Fig. 3.6 and also in the context of PMU placement and power grids, Qi et al. (2015) showed that a PMU placement that conveys higher coefficient of observability (albeit based on the determinant of an empirical observability Gramian) to the power grid leads to a smaller state estimation error (based on an unscented Kalman filter framework).

### 3.6.2 Multi-agent consensus

**Model dynamics.** This section provides an application example in the context of collective behavior of locally interacting adaptive and identical individuals, hereby called agents. Examples of these agents and collective dynamics include fishes in schools, birds

<sup>19</sup>Note that it is not feasible to find the optimal placement via brute-force with a system of dimensionality  $N \gtrsim 100$ .

in flocks and robots in artificial swarms. The following multi-agent system describes the general flocking behavior of a group of moving agents in a 2-dimensional space (Arenas et al., 2008; Bouffanais, 2016; Vicsek et al., 1995):

$$\dot{x}_i = \frac{1}{k_{i,\text{in}}} \sum_{j=1}^m W_{ij}(x_j - x_i), \quad (3.27)$$

for  $i = 1, \dots, m$ , where  $x_i(t)$  is the velocity direction (angle) of agent  $i$  at time  $t$ ,  $k_{i,\text{in}} = \sum_j W_{ij}$  is the node in-degree of agent  $i$ ,  $m$  is the number of agents, and  $W \in \{0, 1\}^{m \times m}$  is an adjacency matrix that describes the signaling network between agents, where  $W_{ij} = 1$  if agent  $i$  perceives or senses agent  $j$  in its neighborhood, and  $W_{ij} = 0$  otherwise.

This local consensus protocol represents the decentralized information flow throughout the swarm as a behavioral response to changes in the leading agents states<sup>20</sup>. The Laplacian matrix  $L$  associated with the adjacency matrix  $W$  plays an important role in this analysis since (3.27) can be represented as

$$\dot{\mathbf{x}} = -D_{\text{diag}}^{-1} L \mathbf{x}, \quad (3.28)$$

where  $D$  is the degree matrix defined in Section 2.2.

Albeit simple, model (3.27) allows us to illustrate our discussion without straying too far from the main focus of this work. It is expected that a more specific application would also have to take into account a more detailed model.

**Sensor networks.** Compared to the application example in Section 3.6.1, the collective motion of agents adds an extra layer to our analysis: the signaling network  $W$  between agents is not fixed, but rather time-dependent. The communication between agents depends on physiological (technological) limitations of living (artificial) agents, which often constrain the sensory range that a single agent can both perceive and process in its neighborhood (Bouffanais, 2016; Pearce et al., 2014). For example, the presence of obstacles forces the swarm of agents to engage in different maneuvers that essentially change the neighborhood of interactions of each single agent (Olfati-Saber, 2006), re-configuring the signaling network. This is a feature that is particularly connected to the design problem of wireless sensor networks (Derakhshan and Yousefi, 2019).

---

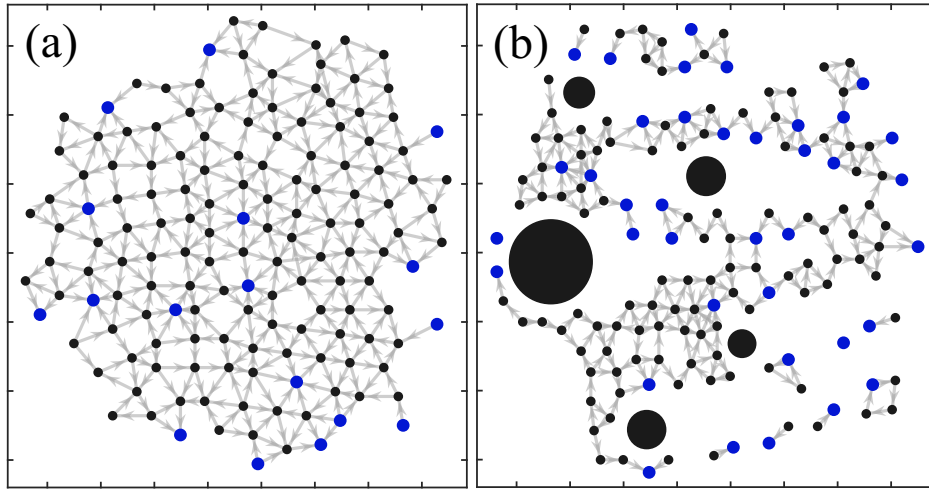
<sup>20</sup>Leading agents can represent agents with privileged information about external factors, such as predators and obstacles, or agents that receive external input from a control center in artificial swarms.

The optimal sensor placement in multi-agent systems, therefore, has a different motivation than in Section 3.6.1. In the present case, each agent usually has its own sensory system that assess neighboring information for decision-making in a decentralized manner. From the point-of-view of a single agent, access to the whole state of a system is not needed and, therefore, there is no observability problem. On the other hand, access to the whole system state might be required for monitoring and control purposes in applications that rely on a command center, such as for unmanned aerial vehicle coordination (Olfati-Saber, 2006) or cyber-attack detection in self-driving vehicles coordination (Vivek et al., 2019). Although the system state can be monitored, in principle, by transmitting the measured states of all agents to the command center, this is not energetically efficient. Battery life is one of the major limitations in artificial swarms applications, and relying on each single agent to transmit its current state to a command center has a high overall energy consumption. A more efficient alternative is to find the minimum set of agents (sensor nodes) that need to transmit their current state to a command center such that the state of the remaining agents can be estimated/reconstructed. This is an observability problem.

**Sensor placement.** In multi-agent systems, the swarm of agents is continuously self-organizing in motion, changing the signaling network  $W$  as a function of the agents position over time (similarly to a switching network). For instance, after a split maneuver, the system might become unobservable from a set of sensor nodes chosen before the start of the maneuver, requiring a new set of sensor nodes to be assigned for the new configuration.

Figure 3.7 illustrates the chosen minimum set of sensor nodes required to convey observability to the system in two different scenarios: the first when the group of agents moves in consensus as a flock, and the second when the group of agents splits apart to engage in a evasive maneuver against obstacles in its path. In this example, the minimum set of sensor nodes is determined by using the maximum matching algorithm (Section 3.3.2) on the graph  $\mathcal{G}(A)$  associated with the dynamical matrix  $A = -D_{\text{diag}}^{-1}L$  given by model (3.28).

Since the signaling network re-configures in real-time, it is important that the minimum sensor placement algorithm is sufficiently fast for such application. For  $m = 150$  agents, the maximum matching algorithm solves the minimum sensor placement problem in less than a second. Indeed, considering the present state-of-the-art, a



**Figure 3.7:** Signaling network  $W$  of a flock of  $m = 150$  agents moving in a 2-dimensional space in two different scenarios: (a) the agents are moving in consensus, and (b) a snapshot of the agents engaged in a split maneuver to dodge obstacles (large circles) along the way. Agents are represented by black nodes (in  $xy$  coordinates), and the minimum set of sensor nodes to render the system structurally observable (according to the maximum matching algorithm) is represented by blue nodes. The signaling networks were adapted from (Olfati-Saber, 2006).

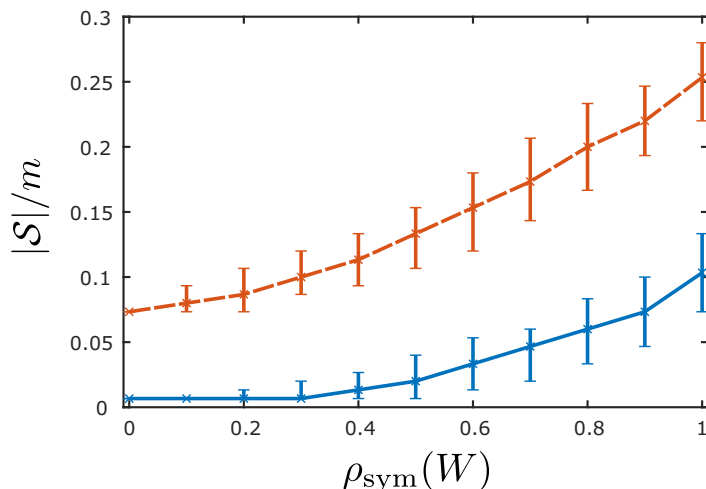
topological observability approach is the only feasible alternative<sup>21</sup> to solve the minimum set of sensor nodes in real-time applications involving swarms with up to thousands of agents. This is further supported by the low complexity order  $O(\sqrt{m}|\mathcal{E}|)$  predicted in a worst-case scenario.

**Observability and network topology.** We now investigate the relation between the minimum set of sensor nodes and the network topology  $\mathcal{G}(A)$ . It is clear that, after the split maneuver, the loss of connectivity in  $W$  increases the size of the minimum set of sensor nodes, specially because additional sensor nodes must be included to track (observe) the state of isolated agents and disconnected subgraphs. On the other hand, Fig. 3.8 shows that the minimum number of sensor nodes is related not only to the number of subgraphs in  $\mathcal{G}(A)$ , but also to the degree of symmetry  $\rho_{\text{sym}}(W)$  of the binary matrix  $W$ , defined as<sup>22</sup>:

$$\rho_{\text{sym}}(W) = \frac{\text{number of unidirectional edges in } \mathcal{G}(W)}{\text{number of edges in } \mathcal{G}(W)}. \quad (3.29)$$

<sup>21</sup>For example, solving the minimum sensor placement problem based on the rank condition (3.2) for observability was shown to be a NP-hard problem (Olshevsky, 2014).

<sup>22</sup>A symmetric adjacency matrix corresponds to an undirected graph, which only has bidirectional edges. Contrariwise, an asymmetric adjacency matrix is related to a directed graph, which has unidirectional (and possibly bidirectional) edges.



**Figure 3.8:** Proportion of the minimum number of sensor nodes  $|\mathcal{S}|/m$  as a function of the signaling network symmetry  $\rho_{\text{sym}}(W)$  for the two different scenarios: consensus (solid line) and split maneuver (dashed line). Simulations are presented for 1,000 Monte Carlo runs where edges in the graph associated with  $W$  are randomly assigned a single direction according to  $\rho_{\text{sym}}(W)$  in each run. Lines show the median values, while error bars show the 5th and 95th percentile.

The lower the symmetry of the signaling network, the larger the minimum sensor set. This is a straight consequence of the changes in the graph topology: a lesser symmetry reduces the number of paths between nodes, and increases the number of dilations in the graph, therefore requiring a higher number of sensor nodes to guarantee structural observability in Lin’s sense (Theorem 3.3). Likewise, reducing the adjacency matrix symmetry (changing bidirectional edges to unidirectional edges) also increases the number of unmatched nodes (as seen in Fig. 3.2c,d), increasing the number of sensor nodes determined by the maximum matching algorithm.

If  $W$  is symmetric (and  $\mathcal{G}(W)$  is undirected), as discussed in Section 3.6.1, any agent can be chosen as a sensor node, regardless of the network size and topology. This raises once again the question of whether a structural or topological approach to observability is sufficient to determine the minimum set of sensor nodes for practical purposes. Nevertheless, Fig. 3.8 suggests that, as  $W$  becomes more asymmetric, determining the minimum set of sensor nodes becomes a more complicated problem which a topological approach helps to shed some light on it. For instance, Fig. 3.8a shows that the studied multi-agent system in consensus can be rendered observable by assigning less than 10% of the total number of agents as sensors regardless of the level of symmetry. Although the study of controllability of multi-agent systems has found several advances in recent years (Guan and Wang, 2018; Komareji and Bouffanais, 2014; Rahmani et al., 2009),



we emphasize that studying the topological observability of multi-agent systems, such as the work of Lu et al. (2017), can lead to novel developments in different fields, including swarm monitoring applications, development of cyber-defense protocols, and optimal configuration of sensor networks.

## 3.7 Final Considerations

In this chapter, we reviewed some definitions of *observability* of dynamical systems, as far as network systems are concerned. Firstly, we presented the traditional concepts of observability proposed by Kalman (and its following extension to nonlinear systems), where a system is classified simply as observable or unobservable. We classify this approach as *structural observability*. However, due to numerical issues, a set of outputs that renders a system observable, but badly conditioned, might not provide a satisfactory state reconstruction under a practical context. Thus, a more relevant question arises: whether an observable system is *almost unobservable*. Several indices were proposed in the literature to quantify the quality of observability of a linear and nonlinear system. We refer to this continuous quantification of observability as *dynamical observability*.

In a network context, traditional control methods usually fail to be applied due to high-dimensionality issues, especially when nonlinear systems are considered. This is also true in an observability context, leading to the need for novel methods that circumvent this problem. The intuitive modelling of network systems via graph representations led to a new set of observability methods that benefits from the network topology properties to derive the minimum set of sensor nodes under which a network is rendered observable. We classify this approach as *topological observability*.

Although intuitive and suitable for high-dimensional network systems, these topological observability metrics have several limitations under a practical context, especially when the nodal dynamics are considered. Indeed, we provide a critical review of the recent progress in the study of observability (and controllability) of network systems, emphasizing the main advantages (e.g., its high scalability and interpretability) and disadvantages (e.g., the underestimation of the necessary set of sensor nodes under a practical context) of topological observability. To circumvent the main disadvantages in the study of observability of network systems, we briefly review some interesting approaches in the literature in order to provide future research directions of interest in this field.

Finally, we show and discuss in two application examples how the concepts of structural, dynamical and topological observability can aid in the problem of sensor placement. In power grids, since all nodes are diffusively coupled according to a full network and self-edges are represented, topological observability (and structural) is easily achieved with any single choice of sensor node. Thus, a dynamical observability is necessary not only to distinguish which choice of sensor node is the best, but also to determine what is the smallest number of sensor nodes so that the degree of observability of the power system reaches a satisfactory value (where it stops growing). On the other hand, a topological approach provides the only highly scalable solution to cope, nowadays, with real-time applications in multi-agent consensus.

As a general conclusion, we argue that, contrariwise to some recent results in the literature, depending on the characteristics of the network system, a topological or structural approach to observability might not “tell the full story”, requiring a further investigation based on dynamical observability. It is suggested that, the less directed a network is and the higher the number of self-edges in its topology, the more a topological observability underestimates the set of sensor nodes.

## Chapter 4

# Functional observability and target state estimation in large-scale networks

Understanding the properties and control principles of large-scale dynamical networks, such as power grids, neuronal networks and food webs, allows at least in principle the development of intervention strategies that shape the behavior of these systems to achieve the desired functionality. As formalized by [Wiener \(2019\)](#), the fundamental mechanism enabling precise control of a dynamical system is *feedback*, which involves sensors, signals, and actuators in a closed loop. A sensor provides immediate measurements of a particular state variable. As the dynamical network grows large, it becomes prohibitive or even impossible to implement a sensor for each system state variable, be it for economical reasons or physical limitations. Therefore, the indirect estimation of the unmeasured variables is essential for the control of large-scale dynamical networks. Observability, as reviewed in [Chapter 3](#), is a key property for the optimal sensor placement and design of state estimators in large-scale networks. Despite the success of state observers in many engineering applications, high-dimensionality is still an obstacle that hampers the direct use of these methods to large-scale dynamical networks ([Chen, 2014](#); [Motter, 2015](#)), calling for different approaches and novel techniques ([Cornelius et al., 2013](#); [Fiedler et al., 2013](#); [Liu et al., 2011b, 2013](#); [Wang and Chen, 2002b](#); [Zañudo et al., 2017](#)) to overcome the curse of dimensionality.

For many real-world problems, estimating the *entire* state vector of high-dimensional systems does not seem necessary at all ([Motter, 2015](#)). It is often sufficient to focus on a particular subset of nodes of interest. For instance, in decentralized control strategies

applied to network systems, each controller only requires feedback signals from a few particular subset of nodes in the neighborhood of its controlled area (Olfati-Saber and Murray, 2004; Xue and Chakraborty, 2018). This is also true for fault detection and monitoring against unforeseen failures or cyber-attacks, which finds several applications in supply networks (Pasqualetti et al., 2013a), power grids (Singh and Pal, 2014; Zhang and Vittal, 2013) and autonomous vehicle coordination (Vivek et al., 2019). In biological networks, a small number of “target nodes”, also known as biomarkers, are known to be of interest for control (intervention) or estimation (medical diagnostics) purposes (Barabási et al., 2011), such as the specific nodes associated with cancer in regulatory networks (Cornelius et al., 2013), or clusters of synchronized neurons associated with Parkinson’s disease (Hammond et al., 2007) and epilepsy (Lehnertz et al., 2009).

These practical problems motivate the concept of *functional observability* (Fernando et al., 2010b; Jennings et al., 2011) which enables the existence of a *functional observer* capable of reconstructing a targeted subset of the state variables of a dynamical system from the inputs and measurements. Though conceptually attractive, previous works on functional observability and the design of functional observers (Darouach, 2000; Fernando et al., 2010a,b; Hieu and Tyrone, 2012; Jennings et al., 2011) were based on numeric rank-based conditions without explicitly taking advantage of the network topology and thus do not lead to scalable algorithms applicable to large-scale dynamical networks.

In this chapter, we develop a graph-theoretic characterization of functional observability and the associated sensor placement and observer design algorithms, making it possible to accurately estimate a desired subset of state variables—hereby known as target variables—of a large-scale dynamical network using minimal sensory and computational resources. The contributions are threefold: Firstly, in Section 4.2, we propose a new concept, i.e., “structural functional observability”, which can be seen as a generalization of Lin’s structural observability (Lin, 1974). It allows us to rigorously establish graph-theoretic conditions for functional observability equivalent to the original rank-based conditions (Jennings et al., 2011). Secondly, based on the proposed theory, two highly-scalable algorithms are developed to solve the sensor placement and observer design problems in Section 4.3. The first algorithm determines the minimal set of sensors placed on a dynamical network to ensure the functional observability with respect to a given set of target nodes. After the placement is decided, the second algorithm designs a minimum-order functional observer whose output converge asymptotically to the target states thus achieving accurate estimation. Numerical results in

large-scale complex networks are shown for both algorithms in Section 4.4. Thirdly, we demonstrate the advantages of the proposed methods with two concrete applications to cyber-security of power grids and surveillance for epidemics in Section 4.5. For power-grid cyber-security, we show that the proposed functional observers can be implemented as active monitors for cyber-attacks in power grids, effectively providing state estimates that allow for cross-validation among different information sources and detection of fake measurement data in real-time. For the epidemic surveillance, we demonstrate that, during a pandemic like COVID-19, the proposed functional observer can infer the infected population at places where testing is inadequate from the data collected at other places where sufficient testing has been done, and our algorithms can also guide the optimal allocation of limited testing resources.

## 4.1 Background on Functional Observability

Generically speaking, a dynamical system is completely observable if it is possible to reconstruct the initial state  $\mathbf{x}(0)$  from knowledge of the input  $\mathbf{u}(t)$  and measurement  $\mathbf{y}(t)$  over a finite time interval. If the rank condition  $\text{rank}(\mathcal{O}) = n$  is true (see Theorem 3.1), there exists straightforward methods to design a full-state observer, i.e., an auxiliary dynamical system whose states converge asymptotically to those of the original system (2.1) when taking  $\mathbf{y}$  and  $\mathbf{u}$  as inputs, providing an estimation of the state vector  $\mathbf{x}$ . Since the direct measurement  $\mathbf{y}$  already contains  $q$  linear combinations of state  $\mathbf{x}$ , only  $(n - q)$  state variables are required to be estimated, which can be accomplished by a reduced-order state observer, also known as Luenberger observer (Luenberger, 1966) (see Appendix B for details).

In practice, it is often unnecessary to estimate the entire state vector  $\mathbf{x}$ . Instead, only a lower-dimensional function

$$\mathbf{z} = F\mathbf{x}, \quad (4.1)$$

where  $\mathbf{z} \in \mathbb{R}^r$ , is usually of interest (e.g., for feedback control or monitoring purposes). Given the desirable  $F \in \mathbb{R}^{r \times n}$ , functional observability is defined as follows.

**Definition 4.1.** *The linear system (2.1) and (4.1), or the triple  $(A, C, F)$ , is said to be functionally observable if for any unknown initial state  $\mathbf{x}(0)$ , there exists a finite time  $t_1 > 0$  such that the knowledge of  $\mathbf{u}$  and  $\mathbf{y}$  over  $t \in [0, t_1]$  suffices to determine uniquely  $\mathbf{z}(0) = F\mathbf{x}(0)$ . Otherwise,  $(A, C, F)$  is said to be functional unobservable.*

**Theorem 4.1.** (*Jennings et al., 2011; Rotella and Zambettakis, 2016b*) *The triple  $(A, C, F)$  is functionally observable if and only if*

$$\text{rank}(\mathcal{O}) = \text{rank} \begin{bmatrix} \mathcal{O} \\ F \end{bmatrix}. \quad (4.2)$$

■

**Remark 4.1.** If condition (4.2) is satisfied, then  $F$  is some linear combination of the rows of  $\mathcal{O}$ , i.e.,  $F = \sum_{i=0}^{n-1} L_i C A^i$ , where  $L_i \in \mathbb{R}^{r \times nq}$ . Substituting it in (4.1) yields

$$\begin{aligned} \mathbf{z}(t) &= F \mathbf{x}(t) = \sum_{i=0}^{n-1} L_i C A^i \mathbf{x}(t) = \sum_{i=0}^{n-1} L_i \bar{\mathbf{y}}^{(i)}(t), \\ &= \begin{bmatrix} L_0 & \dots & L_{n-1} \end{bmatrix} \cdot \begin{bmatrix} \bar{\mathbf{y}}^{(0)}(t) \\ \vdots \\ \bar{\mathbf{y}}^{(n-1)}(t) \end{bmatrix}, \\ &:= L \tilde{\mathbf{y}}, \end{aligned} \quad (4.3)$$

where it is known from (3.7) that  $\bar{\mathbf{y}}^{(i)}(t) = C A^i \mathbf{x}(t)$ . Therefore, Theorem 4.1 guarantees the existence of a linear map  $L$  that determines  $\mathbf{z}$  uniquely from  $\tilde{\mathbf{y}}$ . Nevertheless, even if  $L$  is known, determining  $\mathbf{z}$  from (4.3) has the same shortcomings described in Remark 3.1.

**Remark 4.2.** From a geometrical point-of-view (*Jennings et al., 2011*), condition (4.2) is equivalent to  $\text{row}(F) \subseteq \text{row}(\mathcal{O})$ , where  $\text{row}(\cdot)$  is the row space. In other words, the subspace desired to be estimated (the image of  $F$ ) must be contained inside the observable subspace from  $\mathbf{y}$  (the image of  $\mathcal{O}$ ). Clearly, complete observability (Theorem 3.1) is a special case of functional observability (Theorem 4.1) for  $F = I$ .

Contrariwise to the observability property, which is the sole condition for the straightforward design of a Luenberger observer (see Appendix B.1), condition (4.2) only guarantees the theoretical existence of a functional observer (*Jennings et al., 2011*), and does not readily lead to an algorithm to design a functional observer (*Fernando et al., 2010b; Rotella and Zambettakis, 2016b*). To this end, two additional conditions must be satisfied as depicted in the theorem below (*Darouach, 2000*).

**Theorem 4.2.** (*Darouach, 2000, Theorem 2*) *The necessary and sufficient conditions for the existence and stability of a functional observer of form (B.8) for a triple*

$(A, C, F_0)$  is:

$$\text{rank} \begin{bmatrix} C \\ CA \\ F_0 \\ F_0A \end{bmatrix} = \text{rank} \begin{bmatrix} C \\ CA \\ F_0 \end{bmatrix}, \quad (4.4)$$

$$\text{rank} \begin{bmatrix} \lambda F_0 - F_0A \\ CA \\ C \end{bmatrix} = \text{rank} \begin{bmatrix} CA \\ C \\ F_0 \end{bmatrix}, \quad (4.5)$$

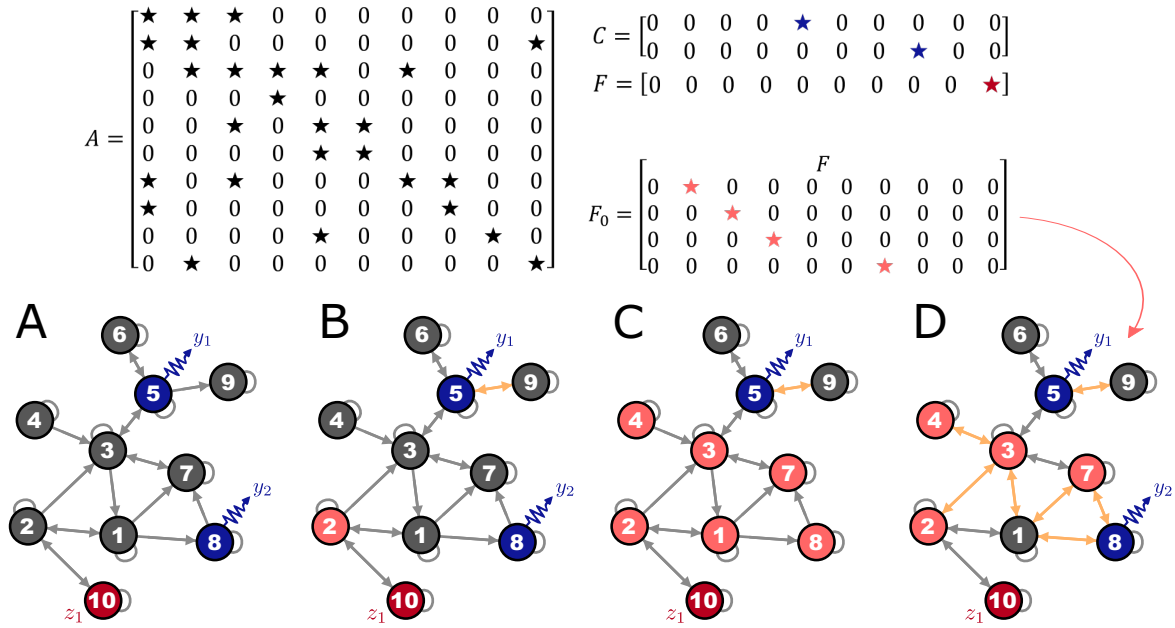
for every eigenvalue  $\lambda$  of  $A$ .

If  $(A, C, F)$  satisfies condition (4.2) but not conditions (4.4)–(4.5), then there exists some matrix  $F_0 \in \mathbb{R}^{r_0 \times n}$  whose row space contains that of  $F$ , i.e.,  $\text{row}(F_0) \supseteq \text{row}(F)$ , that satisfies conditions (4.2)–(4.5) for a triple  $(A, C, F_0)$ . If such an  $F_0$  can be determined, then a functional observer of size  $r_0 \geq r$  can be methodically designed following Algorithm 5 (see Appendix B.2 for details).

## 4.2 Structural Functional Observability

The rank-based conditions (4.2) and (4.4)–(4.5) are not numerically stable and computationally efficient for the design of functional observers on large-scale systems. Here, we adopt a graphical approach that explicitly takes advantages of the network structures of the dynamical systems. As discussed in Section 3.3.1, the system matrix  $A$  can be structurally represented as a corresponding graph  $\mathcal{G}(A)$  whose nodes represent the internal state variables  $\mathcal{X} = \{x_1, \dots, x_n\}$ . The edges of  $\mathcal{G}(A)$  capture the interaction pattern among state variables, i.e., there is an edge from  $x_j$  to  $x_i$  on graph  $\mathcal{G}(A)$  if  $A_{ij}$  is non-zero. We denote a node  $x_j$  on graph  $\mathcal{G}(A)$  as a sensor node if  $C_{ij} \neq 0$  for some  $i$ , and a node  $x_k$  as a target node if  $F_{ik} \neq 0$  for some  $i$ . We assume that each sensor or target is only related to one internal state variable, meaning that each row of  $C$  or  $F$  has only one non-zero entry. The sets of all sensor and target nodes are denoted  $\mathcal{S}$  and  $\mathcal{T}$ , respectively. Figure 4.1a illustrates the construction of graph  $\mathcal{G}(A)$ .

We now provide a generalization of the concept of structural observability to structural functional observability. Following Definition 3.5, we shall say the triple  $(A, C, F)$  has the same structure as another triple  $(\tilde{A}, \tilde{C}, \tilde{F})$ , if by the construction described above, they share the same structure graph  $\mathcal{G}(A)$ , sensor node set  $\mathcal{S}$ , and



**Figure 4.1:** Structural functional observability of dynamical systems. (a) Dynamical system matrices  $(A, C, F)$  (top row) and corresponding graph  $\mathcal{G}(A)$ , with  $n = 10$  state nodes  $\mathcal{X} = \{x_1, \dots, x_{10}\}$ ,  $q = 2$  sensor nodes (blue) in  $\mathcal{S} = \{x_5, x_8\}$ , and  $r = 1$  target node (red) in  $\mathcal{T} = \{x_{10}\}$ . In this choice of sensor and target nodes, the system is unobservable ( $\text{rank}(\mathcal{O}) = 9 < n$ ), but is structurally functionally observable ( $\text{rank}[\mathcal{O}^\top F^\top]^\top = 9$ ). (b) The triple  $(A, C, F)$  is observable (and thereby functionally observable). While a Luenberger observer has to estimate the states of every unmeasured node (order  $n - q = 8$ ), a functional observer only needs to estimate, along with the target node  $x_{10}$ , the pink node  $x_2$  (order  $r_0 = 2$ ). Note that  $(A, C, F_0)$  satisfies conditions (4.2)–(4.5). (c) Sensor node  $x_5$  is a minimum sensor set (among other options) required for the structural functional observability of target  $x_{10}$ . However, the absence of sensor node  $x_8$  increases the functional observer order to  $r_0 = 7$ . (d) The network is strongly connected. This stronger connectivity slightly compromises the target state estimation by increasing the functional observer order to  $r_0 = 5$  (more pink nodes), albeit it is still smaller than  $(n - q) = 8$ .

target node set  $\mathcal{T}$ . For the following definitions and theorems, consider that the following assumption holds.

**Assumption 4.1.** Let  $\text{rank}[C^\top F^\top]^\top = q + r$ , where  $\text{rank}(C) = q$  and  $\text{rank}(F) = r$ .

**Remark 4.3.** Assumption 4.1 provides no loss of generality since any linearly dependent output, or target state, can be determined from a linear combination of outputs and target states, hence not requiring estimation.

**Definition 4.2.** *The structured triple  $(A, C, F)$  is structurally functionally observable if and only if there exists some numerical realization  $(\tilde{A}, \tilde{C}, \tilde{F})$  that is functionally observable.*



According to Definition 4.2, structural functional observability is purely determined by the state variable interaction structure, encoded by graph  $\mathcal{G}(A)$ , sensor node set  $\mathcal{S}$ , and target node set  $\mathcal{T}$ , which is independent of the specific numerical realization of  $(A, C, F)$ . In fact, if a triple  $(A, C, F)$  is structurally functionally observable, a system that shares the same structure as  $(A, C, F)$  is functionally observable for a wide range of parameters except for a proper algebraic variety with null Lebesgue dimension (see Remark 3.8). This structural approach allows us to establish a graph-theoretic characterization of functional observability:

**Definition 4.3.** Let  $\mathcal{D}_k$  be a minimal dilation set of  $\mathcal{G}(A, C)$ , that is, a set with the property that, for all subset  $\mathcal{D}'_k \subset \mathcal{D}_k$ , the subset  $\mathcal{D}'_k$  has no dilations. Let  $\mathcal{D} = \bigcup_k \mathcal{D}_k$  be the union of all minimal dilation sets of  $\mathcal{G}(A, C)$ .

**Theorem 4.3.** A triple  $(A, C, F)$  is structurally functionally observable if and only if the corresponding graph  $\mathcal{G}(A, C)$  satisfies both of the following conditions:

1. every state variable  $x_i \in \mathcal{T}$  has a path to some output  $y_i \in \mathcal{S}$ ;
2.  $\mathcal{T} \cap \mathcal{D} = \emptyset$ , where  $\mathcal{D}$  is the union of all minimal dilation sets of  $\mathcal{G}(A, C)$ .

*Proof.* See Appendix C. □

**Remark 4.4.** The above result can be seen as a significant generalization of Lin's theory of structural controllability (Lin, 1974). Note that if  $F = I$ , or equivalently  $\mathcal{T} = \mathcal{X}$ , then  $\mathcal{T} \cap \mathcal{D} = \emptyset$  if and only if  $\mathcal{D} = \emptyset$ , which implies that the graph  $\mathcal{G}(A, C)$  has no dilations. Therefore, conditions for structural functional observability (Theorem 4.3) reduces to the conditions of structural observability (Theorem 3.3).

We illustrate this graphical characterization of complete observability and functional observability in Fig. 4.1a. The graph has no dilations due to the presence of self-edges and the pair  $(A, C)$  is unobservable because node  $x_9$  does not have a direct path to a sensor node. Even so, the system  $(A, C, F)$  is structurally functionally observable since from the target node  $x_{10}$  there is path to a sensor node  $x_5$  or  $x_8$ .

The above result lays the foundation of functional observer design on large-scale dynamical networks. To enable the algorithm development, we further investigate two main design problems:

1. How to select the minimum set of sensor nodes  $\mathcal{S}$  such that a triple  $(A, C, F)$  is structurally functionally observable?
2. What is the minimum set of "auxiliary" state nodes that must be estimated along with the desired target nodes so that a procedural functional observer design is

possible (see Fig. 4.1b)? In other words, given  $(A, C, F)$ , what is the minimum size  $F_0$  such that (4.4)–(4.5) are satisfied for  $(A, C, F_0)$ ?

Clearly, both questions are intertwined and intrinsically related to the network structure (illustrated in Figs. 4.1c,d). In the following sections, we assume that no target node is an element of a minimal subset of nodes with a dilation. A sufficient condition for this latter assumption is that every target node has a self-edge (i.e.,  $\mathcal{G}(A)$  has no dilations). The importance of including self-edges in dynamical networks models, specially for state control and estimation applications, has been thoroughly discussed in the literature (Cowan et al., 2012; Leitold et al., 2017; Montanari and Aguirre, 2020). Indeed, dilations are not found in a broad range of dynamical networks, especially those described by a set of diffusively coupled subsystems. This includes applications in networks of coupled oscillators (Arenas et al., 2008; Eroglu et al., 2017; Rodrigues et al., 2016), power grids (Dorfler et al., 2013; Nishikawa and Motter, 2015), neuronal models (Aguirre et al., 2017; Izhikevich, 2004), combustion networks (Haber et al., 2018; Perini et al., 2012), regulatory networks (Mirsky et al., 2009; Mochizuki et al., 2013), consensus problems (Olfati-Saber and Murray, 2004) and multi-group epidemiological models (Colizza et al., 2006).

## 4.3 Methods

With large-scale complex networks in mind, this section provides a highly scalable solution for the two design problems raised in Section 4.2: (1) the minimum sensor placement problem, and (2) the minimum-order functional observer design.

A MATLAB implementation of Algorithms 1, 2 and 5 are publicly available at <https://github.com/montanariarthur/FunctionalObservability>.

### 4.3.1 Minimum sensor placement for sets of target nodes

According to the theory and assumption discussed above, the minimal sensor placement problem is to determine a minimum set  $\mathcal{S}$  such that there is a direct path from every target node to some sensor node. We show that the minimum sensor placement problem can be formulated as a *set cover problem*. For each candidate sensor node, let  $\mathcal{R}_i$  denote the set of target nodes in  $\mathcal{G}(A)$  that have a direct path to the state node  $x_i \in \mathcal{X}$ . By this definition, the minimal sensor placement amounts to identifying the minimal node set  $\mathcal{S}$  such that the union of the sets  $\mathcal{R}_i$ ,  $x_i \in \mathcal{S}$ , covers the target set  $\mathcal{T}$ , i.e.,  $\cup_{x_i \in \mathcal{S}} \mathcal{R}_i \supseteq \mathcal{T}$ . This is an NP-hard problem, to which we provide an approximate but

highly scalable solution via Algorithm 1, in which a breadth-first search determines  $\mathcal{R}_i$  for each node  $x_i \in \mathcal{X}$  and a greedy algorithm solves the set cover problem.

---

**Algorithm 1** Minimum sensor placement

---

**input:** graph  $\mathcal{G}(A^\top)$ , target set  $\mathcal{T}$ , candidate set  $\mathcal{C}$

**output:** sensor set  $\mathcal{S}$

initialize  $\mathcal{R}_i \leftarrow \emptyset, \forall i = 1, \dots, |\mathcal{C}|$ ;

**for** all  $x_i \in \mathcal{T}$

starting at node  $x_i$  in graph  $\mathcal{G}(A)$ , find the set of reachable nodes  $\mathcal{R}'_i \subseteq \mathcal{X}$  using a breadth-first search algorithm;

**for** all  $x_j \in \mathcal{C}$

if  $x_j \in \mathcal{R}'_i$ , then  $\mathcal{R}_j \leftarrow \mathcal{R}_j \cup \{x_i\}$ ;

**end**

**end**

initialize  $\mathcal{S} \leftarrow \emptyset$ .

**do**

for all  $x_i \in \mathcal{C} \setminus \mathcal{S}$ , compute gain

$$\Delta(x_i) = \left| \bigcup_{j: x_j \in \mathcal{S} \cup \{x_i\}} \mathcal{R}_j \right| - \left| \bigcup_{j: x_j \in \mathcal{S}} \mathcal{R}_j \right|; \quad (4.6)$$

add the element with highest gain

$$\mathcal{S} \leftarrow \mathcal{S} \cup \{\arg \max_{x_i} \Delta(x_i) | x_i \in \mathcal{C} \setminus \mathcal{S}\}; \quad (4.7)$$

**while**  $\bigcup_{j: x_j \in \mathcal{S}} \mathcal{R}_j \neq \mathcal{T}$ .

---

Algorithm 1 provides an approximate solution to the minimum sensor placement problem in polynomial time. Firstly, a breadth-first search is run for each target node (for-loop), allowing one to determine the sets of target nodes  $\mathcal{R}_i \subseteq \mathcal{T}$  that have a direct path to each state node  $x_i \in \mathcal{C} \subseteq \mathcal{X}$  in  $\mathcal{G}(A)$ , where  $\mathcal{C}$  is a set of candidate nodes for sensor placement. Secondly, a greedy algorithm is used (while-loop) to find an approximation of the minimum set of sensor nodes such that structurally functional observability is guaranteed. Note that a breadth-first search has a complexity order  $O(n + |\mathcal{E}|)$  (Newman, 2010), where  $|\mathcal{E}|$  is the cardinality of the set of edges  $\mathcal{E}$  in  $\mathcal{G}(A)$ , and can be run in parallel for each  $x_i \in \mathcal{T}$ . Meanwhile, the greedy search, in a worst-case scenario, has a complexity order  $O(n^2)$ . The computational complexity of both algorithmic searches are suitable for large-scale complex networks with thousands of state nodes.

### 4.3.2 Minimum order functional observer design

After the sensor nodes are selected, we need to further choose matrix  $F_0$  to enable the design of a functional observer (as discussed in Section 4.1). In the last decade, [Fernando et al. \(2010b\)](#) provided a theoretical solution to the problem of designing a minimum-order functional observer, that is, finding a minimum order matrix  $F_0$  such that conditions (4.4)–(4.5) are satisfied for a triple  $(A, C, F_0)$ , where  $F_0$  is subjected to  $\text{row}(F_0) \supseteq \text{row}(F)$ . However, the method provided in ([Fernando et al., 2010a](#)) is not scalable for high-dimensional systems, because, as further detailed below, it iteratively invokes singular value decomposition (SVD) to numerically check rank condition (4.4) followed by a combinatorial search to determine additional rows to  $F_0$  such that condition (4.5) is satisfied. To circumvent these issues, adopting the structural approach described in previous sections, we can convert the rank-based condition (4.4)–(4.5) onto equivalent graph-theoretical ones, providing a highly scalable solution in the context of large-scale networks.

**Scalability issues of previous numerical procedures.** For completeness, we show the scalability issues present in the most conventional way to numerically implement, as shown by [Fernando et al. \(2010a\)](#), the theoretical results of [Fernando et al. \(2010b\)](#) for the design of a minimum order functional observer. Despite the fact that we only investigate this single algorithm, we note that other numerical procedures proposed in the literature ([Fernando and Trinh, 2014](#); [Mohajerpoor et al., 2016](#); [Rotella and Zambettakis, 2016a](#)), aside from their different numerical performances, have reported no improvement in the scalability of the design algorithm.

A two-stage algorithm is proposed by [Fernando et al. \(2010a\)](#), where a recursive augmentation of  $F_0$  with extra row vectors is carried out in each stage until condition (4.4) and (4.5) are satisfied. The numerical rank condition in (4.4)–(4.5) is computed using singular value decomposition (SVD), which is not very scalable, having complexity order  $O(n^3)$  (and also being unstable for high-dimensional matrices). In a worst-case scenario, one has  $q = 1$  and  $r = 1$ , but—in order to design a stable functional observer—has to estimate  $r_0 = n - q \approx n$  functions. Under these circumstances, the worst-case scenario for the first stage of this algorithm requires finding the minimum  $F_0$  that satisfies (4.4) with  $n$  recursive iterations. Since each iteration requires at least one SVD computation, the first stage of this algorithm has complexity order  $O(n^4)$ . For the second stage, the worst-case scenario requires checking the rank condition (4.5) for up to  $\sum_{k=1}^n \binom{n}{k} = 2^n - 1$  possible submatrices—which has complexity order  $O(2^n)$ . Despite the worst-case analysis, usually the second stage algorithm requires checking

(4.5) for just a few submatrices, or none at all. Thus, the complexity order of the first stage of this algorithm is a more “honest” metric for comparison purposes and the one that we mainly adopt throughout the main text. Note that the low scalability of the numerical procedure in (Fernando et al., 2010a) is a direct consequence of the use of SVD methods to compute the numerical rank.

**Minimum order functional observer design for large-scale systems.** To circumvent the scalability issues reported above, adopting the structural approach described in previous sections, we convert the rank-based condition (4.4)–(4.5) onto equivalent graph-theoretical ones. To achieve this, we first make the observation that, if the corresponding graph of a dynamical system has a self-edge in every target node, condition (4.4) structurally implies (4.5) (see Corollary 4.1 for a proof). In light of this, only condition (4.4) needs to be considered to determine  $F_0$  and hence the combinatorial search is no longer needed. We then propose Algorithm 2 as a highly scalable solution to determine matrix  $F_0$  with the *smallest* order possible by simply augmenting the rows of  $F$  in such a way that (4.4) is satisfied (see Corollary 4.2 for a proof). In Algorithm 2, instead of invoking SVD, the rank condition (4.4) is verified by computing the maximum matching set associated with the corresponding bipartite graph of its matrices. As shown below, the algorithm has a (worst-case scenario) computational complexity of  $O(n^{2.5})$ , which brings a significant improvement compared to the complexity order  $O(n^4)$  of the numerical procedure provided in (Fernando et al., 2010b).

With applications in network systems in mind, notice that, as stated in Section 4.2,  $C$  and  $F$  only have one nonzero element per row and Assumption 4.1 holds throughout this chapter. Firstly, we state Corollary 4.1.

**Corollary 4.1.** *A stable functional observer of order  $r$  exists for some numerical realization  $(\tilde{A}, \tilde{C}, \tilde{F})$  if: (i) condition (4.4) is true for a structured triple  $(A, C, F)$ , and (ii) every target node  $x_i \in \mathcal{T}$  has a self-edge in  $\mathcal{G}(A, C)$ .*

*Proof.* A stable functional observer of order  $r$  exists if and only if Darouach’s conditions (4.4)–(4.5) are true for some numerical realization of a structured triple  $(A, C, F)$  (Darouach, 2000, Theorem 2). We show that, structurally speaking, condition (4.4) implies (4.5) under the given condition (ii).

Consider condition (4.5) for  $\lambda = 0$ . Since, by assumption, condition (4.4) is true, then condition (4.5) can be restated for  $\lambda = 0$  as

$$\text{rank} \begin{bmatrix} FA \\ C \\ CA \end{bmatrix} = \text{rank} \begin{bmatrix} C \\ CA \\ F \\ FA \end{bmatrix}. \quad (4.8)$$

In other words, condition (4.8) holds true if  $\text{row}(F) \subseteq \text{row}([C^\top \ (CA)^\top \ (FA)^\top]^\top)$ . Since condition (4.4) is true, then  $\text{row}(FA) \subseteq \text{row}([C^\top \ (CA)^\top \ F^\top]^\top)$ , i.e.,

$$FA = D_1 \begin{bmatrix} C \\ CA \end{bmatrix} + D_2 F, \quad (4.9)$$

for some matrices  $D_1 \in \mathbb{R}^{r \times 2q}$  and  $D_2 \in \mathbb{R}^{r \times r}$ . If  $D_2$  is invertible, from equation (4.9), we have

$$F = D_2^{-1} FA - D_2^{-1} D_1 \begin{bmatrix} C \\ CA \end{bmatrix}. \quad (4.10)$$

This means that  $\text{row}(F) \subseteq \text{row}([C^\top \ (CA)^\top \ (FA)^\top]^\top)$ . We now show that  $D_2$  is indeed invertible under condition (ii). If every target node  $x_i \in \mathcal{T}$  has a self-edge in  $\mathcal{G}(A, C)$ , then the  $i$ -th entry of at least one row of  $FA$  is a non-zero entry. Since  $x_i$  is a target node, then, by assumption, the  $i$ -th entry of one row of  $F$  is a non-zero entry. As a result, there is always a non-zero value that can be assigned to  $[D_2]_{ii}$  such that (4.9) holds true. Since this result holds for all target nodes  $x_i \in \mathcal{T}$ , by induction,  $D_2$  has non-zero entries in all its diagonal elements. Moreover, since  $D_2$  is a map between structured matrices  $F$  and  $FA$ , then it is also a structured matrix with generic entries (Yamada and Luenberger, 1985). As a result, there is always some numerical realization of  $(A, C, F)$ , and hence of  $D_2$ , such that  $\text{rank}(D_2) = r$  and  $D_2$  is invertible.

Consider now condition (4.5) for  $\lambda \neq 0$ . From the results above, we have that

$$\text{rank} \begin{bmatrix} C \\ CA \\ F \end{bmatrix} = \text{rank} \begin{bmatrix} C \\ CA \\ FA \end{bmatrix}. \quad (4.11)$$

Since  $[(\lambda F - FA)^\top C^\top (CA)^\top]^\top$  is a linear combination of  $[F^\top C^\top (CA)^\top]^\top$  and  $[FA^\top C^\top (CA)^\top]^\top$ , both of which have the same rank, then

$$\text{rank} \begin{bmatrix} \lambda F - FA \\ C \\ CA \end{bmatrix} \leq \text{rank} \begin{bmatrix} F \\ C \\ CA \end{bmatrix}. \quad (4.12)$$

Because  $(A, C, F)$  are structured matrices, there is always some numerical realization of  $(A, C, F)$  such that this upper bound holds true for all  $\lambda \neq 0$ .  $\square$

If we assume that every target node  $x_i \in \mathcal{T}$  has a self-edge in  $\mathcal{G}(A, C)$ , then, in order to solve the minimum order functional observer design problem, it is only necessary to design an algorithm that finds a minimum order  $F_0$  that satisfies condition (4.4). Indeed, it is often the case that a dynamical network has a self-edge in every target node, as seen in the application examples in power grids and epidemiological models in Section 4.5.

---

**Algorithm 2** Minimum-order functional observer design

---

**input:** functionally observable triple  $(A, C, F)$

**output:** functional observer matrices  $(F_0, N, J, H, D, L)$

initialize  $F_0 \leftarrow F$ ,  $r_0 \leftarrow \text{rank}(F_0)$ ,  $\mathcal{M}_1 \leftarrow \emptyset$ ,  $\mathcal{M}_2 \leftarrow \emptyset$ ;

**do**

    update  $G \leftarrow [C^\top (CA)^\top F_0^\top]^\top$ ;

    build a bipartite graph  $\mathcal{B}(\mathcal{V}, \mathcal{X}, \mathcal{E}_{\mathcal{V}, \mathcal{X}})$ , where  $\mathcal{V} = \{v_1, \dots, v_{2q+r_0}\}$  is a set of nodes where each element corresponds to a row of  $G$ ,  $\mathcal{X} = \{x_1, \dots, x_n\}$  is the set of state nodes (where each element also corresponds to a column of  $G$ ), and  $(v_i, x_j)$  is an undirected edge in  $\mathcal{E}_{\mathcal{V}, \mathcal{X}}$  if  $G_{ij}$  is a non-zero entry;

    find the maximum matching set  $\mathcal{E}_m$  associated with  $\mathcal{B}(\mathcal{V}, \mathcal{X}, \mathcal{E}_{\mathcal{V}, \mathcal{X}})$  (e.g., via the Hopcroft-Karp algorithm);

    for all  $x_i \in \mathcal{X}$ , if  $x_i$  is connected to an edge in  $\mathcal{E}_m$ , then update the set of right-matched nodes  $\mathcal{M}_1 \leftarrow \mathcal{M}_1 \cup \{x_i\}$ ;

    define the set of candidate nodes  $\mathcal{C} = \mathcal{M}_2 \setminus \mathcal{M}_1$ , where  $x_j \in \mathcal{M}_2$  if  $[F_0 A]_{ij}$  is a non-zero entry;

    draw an element  $x_k \in \mathcal{C}$  and update  $F_0 \leftarrow [F_0^\top (F')^\top]^\top$  ( $r_0 \leftarrow r_0 + 1$ ), where  $F' \in \mathbb{R}^{1 \times n}$  and  $[F']_{1j} = 1$  if  $j = k$ , and 0 otherwise;

**while**  $\mathcal{C} \neq \emptyset$ ;

design the functional observer matrices  $(N, J, H, D, L)$  for a triple  $(A, C, F_0)$ .

---

For cases where  $(A, C, F)$  is functionally observable, Algorithm 2 provides a scalable solution to the problem of determining  $F_0$  with *minimum* order such that condition (4.4)

is satisfied for the triple  $(A, C, F_0)$ . In this algorithm, we avoid numerical computation of the rank condition in (4.4), which is numerically unstable and computationally demanding for high-dimensional matrices. The numerical rank computation based on SVD methods have a complexity order  $O(n^3)$ . Instead, we compute the structural (or generic) rank of a matrix by finding the maximum matching of the corresponding *bipartite graph* of such a matrix. This is a highly scalable alternative since the maximum matching problem can be solved by the Hopcroft-Karp algorithm, which has a complexity order  $O(\sqrt{n_b}|\mathcal{E}_b|)$ , where  $n_b$  and  $|\mathcal{E}_b|$  are the number of nodes (columns and rows) and edges (non-zero entries) in the bipartite graph (matrix). Fig. 4.2 presents an illustrative example, where it becomes clear how we take advantage of the structural properties of a dynamical system to augment  $F_0$  at every iteration until condition (4.4) satisfied.

Algorithm 2 finds the minimum order  $F_0$  in  $O(n^{2.5})$  time. This complexity order can be estimated in a worst-case scenario where one has a single sensor node ( $q = 1$ ) and target node ( $r = 1$ ), but—in order to satisfy (4.4)—all other unmeasured nodes must be estimated, hence  $r_0 = n - q \approx n$ . This means that Algorithm 2 provides a solution with approximately  $n$  recursive iteration, where a maximum matching algorithm is run at each iteration, yielding  $O(n \cdot \sqrt{n_b}|\mathcal{E}_b|)$ . Note that we have at most  $n_b = 2q + r_0 + n \approx 2n$  nodes in  $\mathcal{B}$ , and let  $|\mathcal{E}_b| = n_b k_{\text{avg}}$ , where  $k_{\text{avg}}$  is the average node degree in  $\mathcal{B}$ . Thus, the complexity order is  $O(n^{2.5})$  if we assume that  $k_{\text{avg}} \ll n$ . Note that this is still a very conservative estimate since usually  $r_0 \ll n$ .

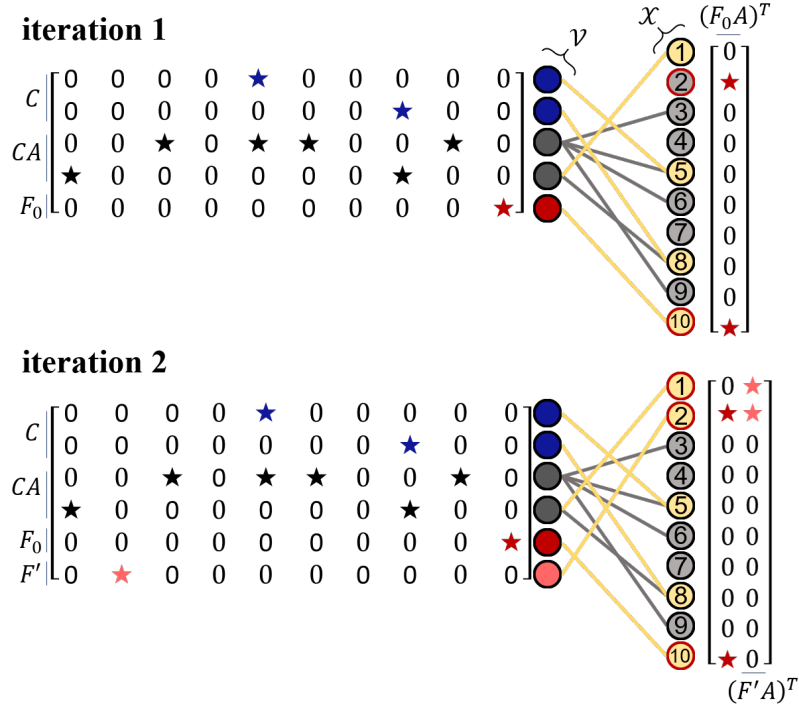
Corollary 4.2 proves that Algorithm 2 finds a matrix  $F_0$  with the smallest order possible under the assumptions stated throughout this chapter.

**Corollary 4.2.** *If  $(A, C, F)$  is structurally functionally observable, then Algorithm 2 returns a matrix  $F_0$  with the smallest order possible such that the rank condition (4.4) is structurally satisfied, under the constrain that  $F_0$  has only one nonzero entry per row.*

*Proof.* From (Fernando et al., 2010b, Lemma 1), condition (4.4) can be satisfied for a  $F_0$  of minimum order by incrementally augmenting  $F_0$  with row vectors orthogonal to

$$\text{row} \begin{bmatrix} C \\ CA \\ F_0 \\ F_0 A \end{bmatrix} \cap \text{row} \begin{bmatrix} C \\ CA \\ F_0 \end{bmatrix}. \quad (4.13)$$





**Figure 4.2:** Illustrative example of Algorithm 2 running for the dynamical network in Fig. 4.1b. On iteration 1, the algorithm builds the corresponding bipartite graph  $\mathcal{B}$  for  $[C^T \ (CA)^T \ F_0^T]^T$ . Note that the rows (nodes in  $\mathcal{V}$ ) corresponding to  $C$  and  $F_0$  are connected, respectively, to the sensor and target nodes in  $\mathcal{X}$ . Rows of  $CA$  and  $F_0A$  are connected to the “predecessors” of the sensor and target nodes (i.e., the nodes that point to them). Using some maximum matching algorithm, the matched edges  $\mathcal{E}_m$  and right-matched nodes in  $\mathcal{M}_1$  are highlighted in yellow. We highlight the elements in  $\mathcal{M}_2$  (predecessors of  $F_0$ ) with a red outline, and define  $\mathcal{C} = \{x_2\}$  (elements in  $\mathcal{M}_2$  that are not right-matched). After drawing element  $x_2$  in  $\mathcal{C}$ , and updating  $F_0$ , the algorithm proceeds to iteration 2. The same steps are repeated. Since  $x_1 \in \mathcal{M}_2$  is already a right-matched node, then  $\mathcal{C} = \emptyset$  and the process terminates. Note that the cardinality of  $\mathcal{V}$  increases at every iteration, also increasing the computational burden in the maximum matching computation. To avoid this, we provide a more efficient MATLAB implementation of Algorithm 2 that uses an incremental procedure to avoid computation of the maximum matching for the *whole* bipartite network at every iteration.

The maximum matching search in Algorithm 2 determines the set of right-matched nodes  $\mathcal{M}_1$  such that each of its elements corresponds to a set of basis vectors that spans  $\text{row}([C^T \ (CA)^T \ F_0^T]^T)$ . Likewise, elements of  $\mathcal{M}_2$  corresponds to basis vectors that spans  $\text{row}(F_0A)$ . Thus, the elements of  $\mathcal{C} = \mathcal{M}_2 \setminus \mathcal{M}_1$  corresponds to a set of basis vectors that spans the orthogonal complement of (4.13). Remind that each of these basis vectors have only one non-zero entry per row. Thus, from (Fernando et al., 2010b, Lemma 1), by recursively adding the basis vector corresponding to elements of  $\mathcal{C}$  to  $F_0$ ,

Algorithm 2 returns the minimum order  $F_0$  subject to the constraint that  $F_0$  (and  $F$ , by assumption) have only one nonzero entry per row.  $\square$

## 4.4 Numerical Results in Large-Scale Complex Networks

In this section, we show numerical results of Algorithms 1 and 2, developed in Section 4.3, applied to random complex networks and real-world networks. In what follows, we describe the numerical setup for the generated random networks and the real-world networks datasets used throughout this section.

**Generation of random complex dynamical networks** For the random generation of the complex networks, we adopt the following choice of parameters:  $N$  is the number of nodes;  $m = \{1, 3, 5, 7\}$  for a Barabási-Albert scale-free (SF) network (Barabási, 1999); and  $k = 2$ ,  $p = \{0, 0.2, 0.5, 1\}$  for a Newman-Watts small-world (SW) network (Newman and Watts, 1999). Parameter  $m$  is the number of edges that a new node attaches to existing nodes,  $k$  is the number of nearest neighbors in a ring graph and  $p$  is the probability of adding a new edge. For each one of these undirected networks, a *directed* model is generated by randomly assigning a single direction to each edge.

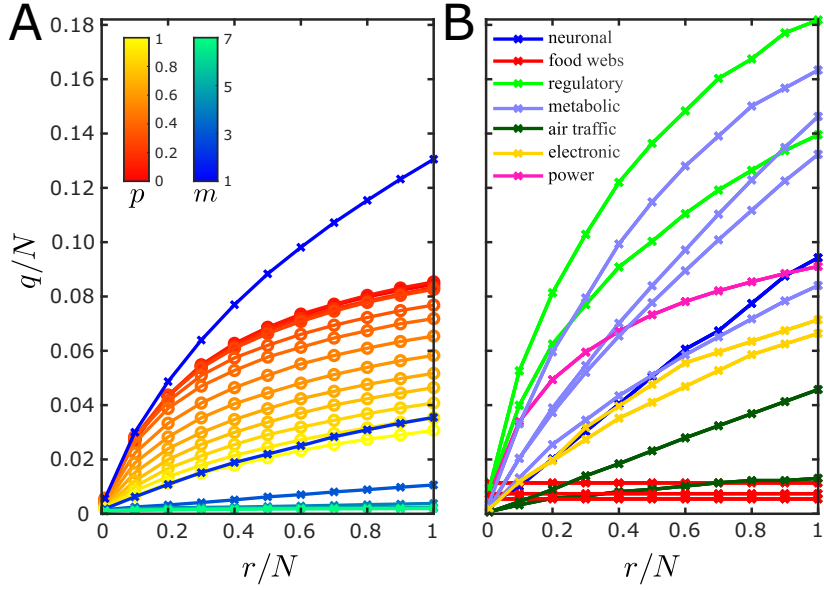
Since we are concerned with *dynamical networks*, we assume that, in each node of a generated complex network, there is a 3-dimensional subsystem with the following general structure:

$$A_{\text{node}} = \begin{bmatrix} -1 & -1 & 0 \\ 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix}. \quad (4.14)$$

To include the effects of heterogeneity in the nodal dynamics of the generated dynamical networks, we let each subsystem's dynamics be defined by  $A_i = \lambda_i A_{\text{node}}$ , for  $i = 1, \dots, N$ , where  $\lambda_i \sim \mathcal{U}[2, 5]$ . Thus, similar to (2.7) the dynamical matrix  $A$  describing the whole dynamical network is given by

$$A = \text{diag}(\lambda_1, \dots, \lambda_N) \otimes A_{\text{node}} - L \otimes M, \quad (4.15)$$

where  $\otimes$  is the Kronecker product operator,  $L$  is the Laplacian matrix of the generated complex network, and  $M = \{0, 1\}^{3 \times 3}$  is defined by  $M_{ij} = 1$  if  $i = j = 2$  and 0 otherwise. The term  $L \otimes M$  means that the second state variable of all subsystems are diffusively coupled according to  $L$ . Note that  $A$  has dimension  $n = 3N$ .



**Figure 4.3:** Minimum sensor placement in large-scale networks. Minimum number of sensors  $q/N$  as a function of the number of target nodes  $r/N$  in (a) randomly generated directed small-world (SW) and scale-free (SF) networks, and (b) real-world networks. Results are shown for an average over 100 realizations of randomly selected target nodes on each network. The minimum set of sensor nodes is determined using Algorithm 1 (with  $\mathcal{C} = \mathcal{X}$ ). Random complex networks were generated with  $N = 10^4$  nodes, where each node has a 3-dimensional subsystem (i.e.,  $n = 3N$ ), while real-world networks are assumed to have 1-dimensional subsystems (i.e.,  $n = N$ ). Parameter  $p$  is the probability of adding a new edge in a SW network and parameter  $m$  is the number of existing nodes a new node is connected to in a BA network.

State nodes were chosen as sensor or target nodes in this section under Assumption 4.1 and that only the first state variable of each subsystem  $A_i$  can be chosen as a sensor or target node (i.e.,  $C_{ij}$  or  $F_{ij}$  is a non-zero entry only if  $(j + 2)/3$  is integer), hence  $\mathcal{S}$  and  $\mathcal{T}$  can have at most  $N$  elements.

**Real-world networks datasets.** For the real-world networks used in Section 4.4.1, we take several adjacency matrices  $A_{\text{adj}}$  available in different real-world datasets shown in Table 4.1. For each real-world network, we define a dynamical matrix  $A$  as the Laplacian matrix of  $A_{\text{adj}}^T$ . We use the Laplacian matrix in order to model the energy/information flow in  $A_{\text{adj}}$  as diffusive processes, and we use  $A_{\text{adj}}^T$  since  $A_{\text{adj}}$  is defined, in the studied databases, under a different convention where  $x_i$  has a directed arrow to  $x_j$  in the corresponding graph  $\mathcal{G}(A_{\text{adj}})$  if  $[A_{\text{adj}}]_{ij}$  is a non-zero entry.

Table 4.1: Dataset of real-world networks studied in the paper.

Type	Name	$N$	$ \mathcal{E} $	Description
Neuronal	<i>C. elegans</i> (Liu et al., 2011b; Watts and Strogatz, 1998)	297	2,345	Neuronal network of <i>C. elegans</i> .
Food web	Grassland (Dunne et al., 2002; Liu et al., 2011b)	88	137	Food web in Grassland.
	Ythan (Dunne et al., 2002; Liu et al., 2011b)	135	601	Food web in Ythan.
	Little Rock lake (Liu et al., 2011b; Martinez, 1991)	183	2,494	Food web in Little Rock lake.
Regulatory	TRN-Yeast-2 (Liu et al., 2011b; Milo et al., 2002)	688	1,079	Transcriptional regulatory network of <i>S. cerevisiae</i> .
	TRN-EC-2 (Liu et al., 2011b; Milo et al., 2002)	418	519	Transcriptional regulatory network of <i>E. coli</i> .
Metabolic	<i>C. elegans</i> (Duch and Arenas, 2005; Kunegis, 2013)	453	2,040	Metabolic network of <i>C. elegans</i> .
	<i>E. coli</i> (iAF1260) (Liu et al., 2013; Schellenberger et al., 2010)	1,668	6,142	Metabolic network of <i>E. coli</i> .
	<i>S. cerevisiae</i> (iND750) (Liu et al., 2013; Schellenberger et al., 2010)	1,059	4,347	Metabolic network of <i>S. cerevisiae</i> .
	<i>H. sapiens</i> (RECON1) (Liu et al., 2013; Schellenberger et al., 2010)	2,766	10,280	Metabolic network of <i>H. sapiens</i> .
Air traffic	Air traffic control (Kunegis, 2013)	1,226	26,615	Federal Aviation Administration's command center.
	US airports (Kunegis, 2013)	1,574	28,236	Air traffic network between US airports.
Electronics	s420 (Liu et al., 2011b; Milo et al., 2002)	252	399	Sequential logic circuit.
	s838 (Liu et al., 2011b; Milo et al., 2002)	512	819	Sequential logic circuit.
Power grid	Western US power grid (Kunegis, 2013; Watts and Strogatz, 1998)	4,941	6,594	Power grid in the Western states of the US.

### 4.4.1 Minimum sensor placement

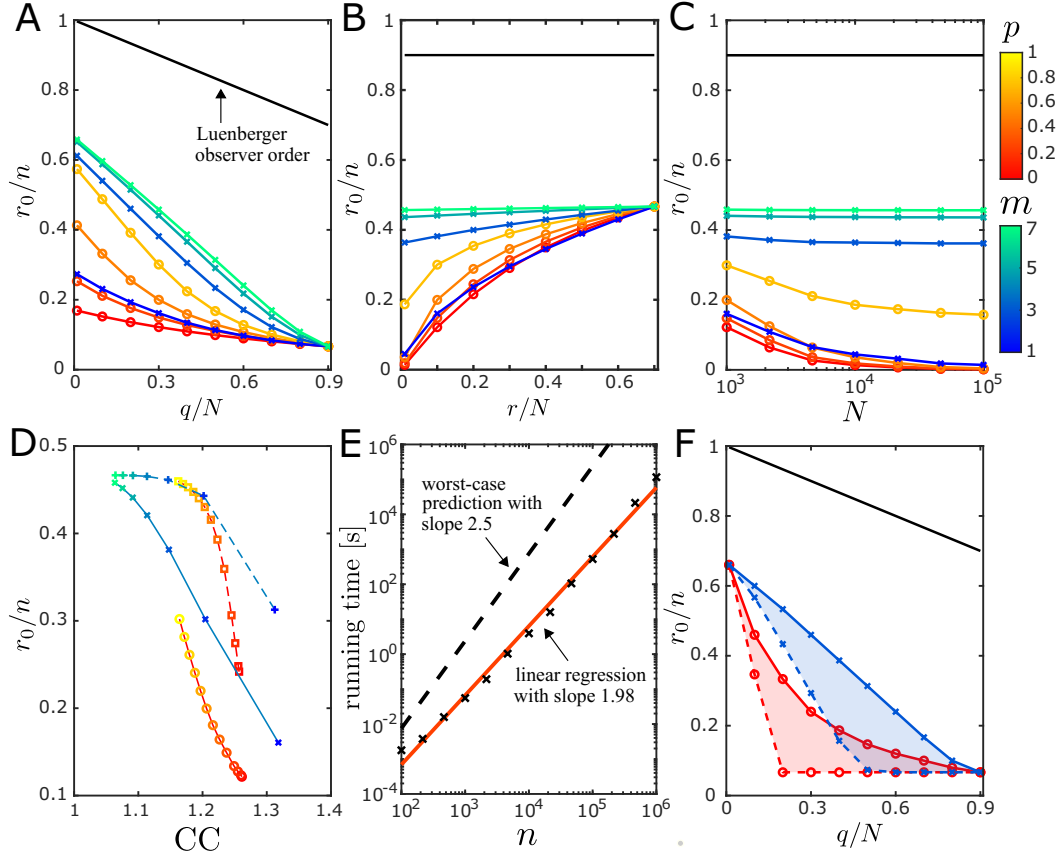
The application of Algorithm 1 to several randomly generated complex networks as well as real-world networks is illustrated in Figure 4.3. The results show that, generally speaking, the fewer target nodes, the fewer sensor nodes are required to guarantee the functional observability of a system. Indeed, as  $r$  approaches  $n$ , the minimum number of sensor nodes tends to be the one required for complete observability of a system. In the studied metabolic networks, Fig. 4.3b shows that monitoring around 8 to 15% of the total number of metabolites is *sufficient* for complete observability of the network system ( $r = n$ ). This is consistent with results found in (Liu et al., 2013, Table 1), where it was shown that a *necessary* number of sensor nodes for complete observability of metabolic networks lies around 5 to 10% of the total number of metabolites.

Complete observability is often unnecessary for many biomedical applications since the number of biomarkers (target nodes whose activities are altered by some disease) are usually much smaller than the network size ( $r \ll n$ ) (Barabási et al., 2011). Our results show that a functional observability approach is more feasible in such applications, especially because a significantly smaller number of sensor nodes is actually needed for estimation of the biomarkers concentrations (Fig. 4.3b). This is also true for cyber-physical systems in engineering applications (e.g., power grids and transportation networks), where one might be interested in monitoring and detecting potential failures or cyber-attacks in specific nodes. Overall, Fig. 4.3a shows that many other systems described by complex network models can have a substantially small set of sensor nodes if only a few target nodes are of interest, specially if the network connectivity is larger (i.e., higher parameters  $p$  and  $m$  in SW and SF networks, respectively).

### 4.4.2 Minimum order functional observer design

For different randomly generated complex networks, Fig. 4.4a,b illustrate the order of the minimum functional observer, determined via Algorithm 2, as a function of numbers of sensor and target nodes. Overall, functional observers are of much lower order compared with the corresponding Luenberger observers, leading to significant improvement in computation efficiency and scalability when designing and implementing the observers in large-scale networks. Generically speaking, the larger sensor set  $\mathcal{S}$  leads to lower order  $r_0$ , whereas the larger target set  $\mathcal{T}$  results in higher order  $r_0$ , as shown in Fig. 4.4a,b.

For a fixed number of target nodes, Fig. 4.4c shows that the functional observer order to the system dimension ratio  $r_0/n$  decreases as we increase the network size,



**Figure 4.4:** Minimum order functional observer design in large-scale networks. Minimum functional observer order  $r_0/n$  as a function of the (a) number of sensor nodes  $q/N$ , (b) number of targets  $r/N$ , and (c) the network size  $N$ , in different directed random networks. Results are color coded for SW and SF networks with different generation parameters  $p$  and  $m$ , respectively. Other parameters are set as  $(N, r) = (10^4, 0.1N)$  for (a),  $(N, q) = (10^4, 0.3N)$  for (b), and  $(q, r) = (0.3N, 100)$  for (c). Sensor and target nodes were randomly placed in the network. The black solid line shows the Luenberger observer order  $(n - q)$  for comparison purposes. (d) functional observer order  $r_0/N$  as a function of the global clustering coefficient (2.4) of  $\mathcal{G}(A)$  in directed (solid line) and undirected (dashed line) SW and SF networks, with  $(N, q, r) = (10^4, 0.3N, 0.1N)$ . (e) Running time of Algorithm 2 (in seconds) as a function of  $N$  in a directed SW network with  $(q, r, p) = (0.3N, 0.1N, 0.2)$ . (f)  $r_0/N$  as a function of  $q/N$  in undirected SW and SF networks, with  $(N, r) = (10^2, 0.1N)$ , for cases where sensor nodes are randomly placed (solid line) and optimally placed (dashed line) using a greedy algorithm to solve (A.3) (see Appendix (A)). Results are the average of 100 Monte Carlo runs in all plots.

which means that the order reduction gained by the functional observer compared to the Luenberger observer increases with the network size. The magnitude of this gain, however, depends more intrinsically on other system properties, including the network structure  $\mathcal{G}(A)$ , the choice of target nodes in  $\mathcal{T}$ , and how sensor nodes in  $\mathcal{S}$  are placed.

We find that more clustered and directed networks seems to lead to a larger order reduction in the design of a functional observer compared to a Luenberger one. This is illustrated in Fig. 4.4d, where we can see that as  $p$  and  $m$  increases in SW and SF networks, respectively, the clustering coefficient increases and the functional observer order decreases more sharply. Interestingly, although directed networks require a larger minimum set of sensor nodes to guarantee the structural functional observability of a system compared to undirected networks (which only requires one node), the directed network allows the design of functional observers of smaller orders. This result also highlights that Algorithm 2 brings computational improvement for both directed and undirected network applications. Finally, Fig. 4.4e illustrates how the running time of Algorithm 2 scales with the network size, showing that it does not surpass our worst-case prediction of  $O(n^{2.5})$ .

Figures 4.4a–e show results considering that sensors and targets are randomly placed in the network. However, the functional observer order  $r_0$  can be further reduced by optimizing the sensor placement (as previously seen in Fig. 4.1b–d). Given a dynamical network  $\mathcal{G}(A)$  and a set  $\mathcal{T}$ , one might be interested not only in solving the minimum sensor placement problem but also in finding, for a fixed number of additional sensors, the best placement of  $\mathcal{S}$  such that the order of  $F_0$  is minimized, thereby minimizing the computational costs in the functional observer design and implementation. This is, however, a difficult bi-level optimization problem, which—for illustration purposes only—we provide a non-scalable greedy algorithm (see Appendix A.2 for details). Such algorithm allows us to show that one can design a functional observer with *even smaller* order if the sensor nodes in  $\mathcal{S}$  are optimally placed in some way instead of randomly placed, shown in Fig. 4.4f. Albeit this specific result is illustrated in a low-dimensional setting, we can extrapolate from Fig. 4.4c that this optimal sensor placement is also relevant in a high-dimensional setting.

### 4.4.3 Performance comparison between observers

This section provides a performance comparison between (reduced-order) Luenberger observers and functional observers in large-scale complex networks. In what follows, we describe the numerical setup for the observer design and simulation.

For each generated complex dynamical network  $A$  and some given sets  $\mathcal{S}$  and  $\mathcal{T}$ , we have a functionally observable triple  $(A, C, F)$  in which we follow three steps to evaluate the observer performance:

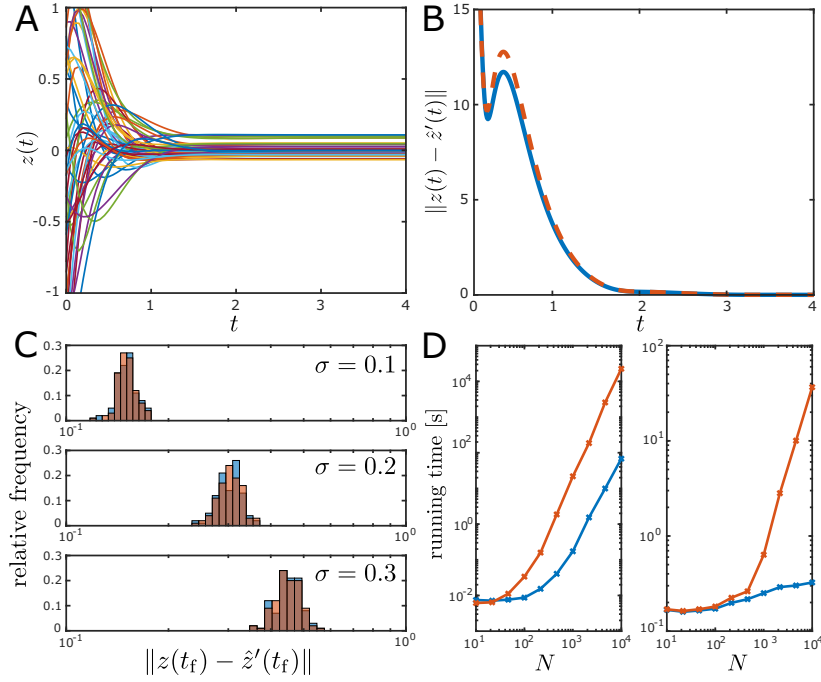
1. we design a Luenberger observer of form (B.5) (using Algorithm 4) and a minimum order functional observer of form (B.8) (using Algorithm 2 to determine a minimum order  $F_0$  then Algorithm 5 to determine the auxiliary matrices);
2. using a fourth-order Runge Kutta method, we simulate the dynamics of the physical system (2.1), the Luenberger observer (B.5) and the functional observer (B.8) excited by a a step input of  $\mathbf{u}(t) = 10, \forall t \geq 0$  (where  $B = [1 \dots 1]^\top$ ), with integration step  $\delta t = 0.01$ , initial conditions  $x_i(0) \sim \mathcal{N}(0, 1), \forall i$ , and  $w_i(0) \sim \mathcal{N}(10, 1), \forall i$ , and a total simulation time  $t_f = 4$  s (which is sufficient to reach a steady-state regime);
3. and, at each integration step  $k\delta t$ , where  $k = 1, 2, \dots, t_f/\delta t$ , we compute the estimation error of each observer as  $\|\mathbf{z}(k) - \hat{\mathbf{z}}'(k)\|$ , where  $\mathbf{z}(k) = F\mathbf{x}(k) \in \mathbb{R}^r$  is the (true) value of the target vector and  $\hat{\mathbf{z}}'(k) \in \mathbb{R}^r$  is the corresponding observer estimate of the target vector.

Note that, in the case of a Luenberger observer,  $\hat{\mathbf{z}}'(k)$  is inferred from the  $(n - q)$ -dimensional vector  $\hat{\mathbf{x}}_u$  estimated in (B.5), while, in the case of a functional observer,  $\hat{\mathbf{z}}'(k)$  is inferred from the  $r_0$ -dimensional vector  $\hat{\mathbf{z}}(k)$  estimated in (B.8).

In this work, we use the linear-quadratic regulator (LQR) as a pole-placement algorithm for the observer design, which requires solving the algebraic Riccati equation  $X^\top P + PX - PYR^{-1}Y^\top P + Q = 0$  for  $P$ . Let  $E \leftarrow P$ ,  $X \leftarrow A_{22}^\top + \alpha I$ , and  $Y \leftarrow A_{12}^\top$  for a Luenberger observer design, and  $Z \leftarrow P$ ,  $X \leftarrow N_2^\top + \alpha I$ ,  $Y \leftarrow N_1^\top$  for a functional observer design. In both cases, we define  $\alpha = -100$ ,  $Q = 10^{-3} \cdot I$  and  $R = I$ . The diagonal terms in  $X$  guarantee that  $Z$  and  $E$  are designed to have the right-most eigenvalues equal to  $\alpha$  with minimum estimation energy ( $R \gg Q$ ). This ensures that their dynamics are dominated by the same slowest eigenvalue, allowing a consistent comparison of observer performances despite their different orders.

Figure 4.5 compares the performance of the functional observer and the Luenberger observer when estimating the target state evolution shown in Fig. 4.5a. The transients of the target state estimation error  $\|\mathbf{z}(t) - \hat{\mathbf{z}}'(t)\|$  in Fig. 4.5b show that the functional observer has similar dynamical behavior as the Luenberger observer and the output of both observers converge to the accurate internal states of the system. Statistical analysis in Fig. 4.5c further reveals that both observers perform asymptotically close even under the effects of modelling errors. Overall, numerical results show that the considerable order reduction in the functional observer design does not compromise its efficacy. Figure 4.5d, on the other hand, shows that such order reduction significantly reduces the computational costs both in the design and online implementation of the functional observer. Such computational advantage of functional observer makes





**Figure 4.5:** Performance of functional observers for target state estimation in large-scale networks. (a) Target variables  $\mathbf{z}(t) = F\mathbf{x}(t)$  dynamical evolution over time  $t$ . (b) Target state estimation error  $\|\mathbf{z}(t) - \hat{\mathbf{z}}'(t)\|$  evolution over time  $t$ , where  $\mathbf{z}(t)$  is the target state vector’s “true values” and  $\hat{\mathbf{z}}'(t)$  is the estimated target state vector by a functional observer (solid line) or Luenberger observer (dashed line). (c) Histogram of the steady-state estimation error  $\|\mathbf{z}(t_f) - \hat{\mathbf{z}}'(t_f)\|$ , where  $t_f = 4$  s, of a functional observer (blue) and Luenberger observer (orange), for different modelling errors  $\sigma$  in the dynamical matrix  $A$ . In this simulation, each observer is designed using a dynamical matrix  $\tilde{A}$  where each entry is drawn from a uniform distribution as  $\tilde{A}_{ij} \sim \mathcal{U}[(1 - \frac{\sigma}{2})A_{ij}, (1 + \frac{\sigma}{2})A_{ij}]$ . (d) Running time of the design algorithm (left) and simulation time of the dynamics (right) of a functional observer (blue) and Luenberger observer (orange) as a function of the network size  $N$ . Results are shown for a directed SW network where each node has a 3-dimensional subsystem (i.e.,  $n = 3N$ ), and sensors and targets were randomly chosen. Parameters were set as  $(p, q, r) = (0.2, 0.3N, 0.1N)$ , and  $N = 10^3$  in  $(A, B, C)$ . Simulations in (c,d) are the average of 100 Monte Carlo runs.

it superior or even indispensable in large-scale networks, especially when constant re-design of the observer is expected due to continuous evolution of system equilibrium and network structure.

## 4.5 Applications

This section provides two application examples of the developed algorithms in the context of power grids and epidemics.

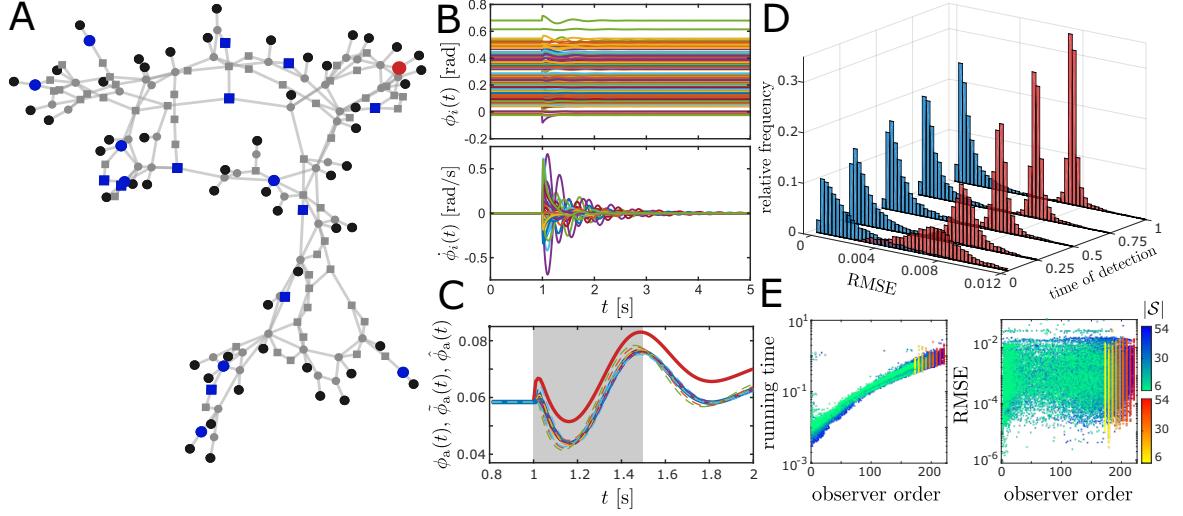
### 4.5.1 Cyber-attacks detection in power grids

The control of man-made technological systems, such as power grids, supply networks, interconnected autonomous vehicles and swarms of robots, is mediated by a sensory and communication infrastructure that captures and exchange measurements from the physical system between areas that are geographically apart. Decentralized control strategies (Bakule, 2008), such as wide-area damping control in power grids (Xue and Chakraborty, 2018), are important to improve the system stability and suppress perturbations which often lead to system-wide failures (Yang et al., 2017). Such control strategies, however, rely on a resilient communication network, which is arguably more vulnerable to potential failures and cyber-attacks than the physical system itself. There have been growing threats to cyber-security, among which are cyber-attacks to supervisory control and data acquisition (SCADA) systems leading to the massive power outages in Ukraine (Lee et al., 2016), the Maroochu Water Services breach in Australia (Slay and Miller, 2007), the substantial damage on Iran’s nuclear program with the Stuxnet computer worm (Farwell and Rohozinski, 2011), and, more recently, the short communication outages in the Western U.S. power grid on March 2019 (NAE, 2019).

Two common types of cyber-attacks (and failures) are based on jamming or corruption of measurement signals being transmitted in some communication channel—also known as denial-of-service and deception attacks, respectively (Amin et al., 2009). Depending on the set of attacks, however, knowledge of the physical system dynamics and the transmitted data can still be used to design observers capable of recovering lost data and also detecting the presence of stealth deception attacks through state estimation (Giraldo et al., 2018; Liu et al., 2011a; Pasqualetti et al., 2013a; Teixeira et al., 2010). In what follows, we show, in the context of power grids, how functional observers can be implemented for cyber-attack monitoring and detection, and how they are a computationally efficient alternative to such applications compared to the use of full-state estimators, e.g., Luenberger observers and Kalman filters.

The power grid dynamics can be modeled as a structure-preserving network of coupled Kuramoto oscillators (Dorfler et al., 2013; Nishikawa and Motter, 2015), where the generators dynamics are governed by the so-called swing equation,

$$\frac{2H_i}{\omega_R} \ddot{\phi}_i + \frac{D_i}{\omega_R} \dot{\phi}_i = P_i + \sum_{j=1, j \neq i}^N K_{ij} \sin(\phi_j - \phi_i), \quad (4.16)$$



**Figure 4.6:** Target state estimation for cyber-attack detection in power grids. (a) Schematic network of the IEEE-118 power grid, which comprises of 118 buses with  $n_g = 54$  generator oscillators (black circle), each one connected to a corresponding generator terminal (gray circle), and  $n_l = 64$  load oscillators (gray square). PMUs placed in load buses are shown in blue. Phase oscillator whose transmitted measurement is corrupted by a deception cyber-attack ( $|\mathcal{C}| = 1$ ) is shown in red. (b) Dynamics of the oscillators phase  $\phi_i(t)$ , for  $i = 1, \dots, N$ , and the generators frequency  $\dot{\phi}_i(t)$ , for  $i = 1, \dots, n_g$ , over time  $t$ . (c) A deception cyber-attack framework is illustrated. Transmitted oscillator phase  $\phi_a(t) \in \mathcal{C}$  (blue solid line) is replaced with false data  $\tilde{\phi}_a(t)$  (red solid line). The colored dashed lines show the state estimation, from multiple observers, of the lost data. (d) Performance of the cyber-attack detection framework. Histogram of the RMSE between the functional observers' state estimates  $\hat{\phi}_a(t)$  and the possibly attacked measurement  $\phi_a(t)$ , as a function of the *time of detection*  $t_d$ . Left histogram (blue) shows results when  $\phi_a(t)$  is not under attack, and right histogram (red) shows results when  $\phi_a(t)$  is under attack, i.e.,  $\phi_a(t) \leftarrow \tilde{\phi}_a(t)$ . (e) Running time of the observer design and state estimation error (RMSE between  $\phi_a$  and  $\hat{\phi}_a$ ) as a function of the observer order. Results are color coded for different scenarios where functional (blue scale) and Luenberger (red scale) observers are designed with different number of sensors  $|\mathcal{S}|$ . Simulations are shown for 100 Monte Carlo runs, where elements of  $\mathcal{S}$  and  $\mathcal{C}$  were randomly chosen in each run, and 100 observers are designed from distinct subsets  $\mathcal{S}'$  per run.

for  $i = 1, \dots, n_g$ , and the dynamics of load buses and generator terminals are described as first-order phase oscillators,

$$\frac{D_i}{\omega_R} \dot{\phi}_i = P_i + \sum_{j=1, j \neq i}^N K_{ij} \sin(\phi_j - \phi_i), \quad (4.17)$$

for  $i = n_g + 1, \dots, N$ , where  $n_g$  is the number of generators,  $n_l$  is the number of load buses,  $N = 2n_g + n_l$  is the number of oscillators, and  $n = N + n_g$  is the system dimension. Let  $\phi_i(t)$  be the phase angle of oscillator  $i$  at time  $t$  relative to frame

rotating at the reference frequency  $\omega_R$  in rad/s,  $H_i$  and  $D_i$  be the inertia and damping constants. In addition,  $K_{ij} = V_i V_j B_{ij}$  where  $V_i$  and  $V_j$  are the voltage levels at bus  $i$  and  $j$  and  $B_{ij}$  is the susceptance of the transmission line connecting bus  $i$  and  $j$ . If there is no line connecting bus  $i$  and  $j$ ,  $K_{ij} = 0$ . The power injection  $P_i > 0$  represents power generation whereas  $P_i < 0$  indicates power consumption by the loads.

We illustrate our framework on the IEEE-118 benchmark system with a diagram shown in 4.6a. Parameters  $(K_{ij}, P_i, H_i, D_i, \omega_R)$  in (4.16)–(4.17) were computed based on the IEEE-118 dataset using the MATLAB toolbox provided by Nishikawa and Motter (2015). Power flow equations were numerically solved using the MATPOWER toolbox (Zimmerman et al., 2011). Initial conditions were set assuming that the power system was in a synchronized steady-state, i.e.,  $\dot{\phi}_i(0) = 0, \forall i$ , and each  $\phi_i(0)$  is determined by the power flow solution.

We assume that the power grid is equipped with a set of phasor measurement units (PMUs) randomly placed on a small number of load or generator terminal buses, i.e., the set of sensor nodes  $\mathcal{S} \subseteq \{\phi_{n_g+1}, \dots, \phi_N\}$ , where  $|\mathcal{S}| = 0.3(n_g + n_l)$ . These PMU measurements are transmitted to a control center in real-time to support automated control actions, human-based decision-making, and cyber-attack detection, etc. In this study, the system is initially operating in steady-state regime when, right after the power system is hit by a small perturbation at time  $t = 1$  s (Fig. 4.6b), a coordinated cyber-attack corrupts one of the sensor measurements  $\phi_a \in \mathcal{C} \subset \mathcal{S}$ , transmitting false information instead to the control center, illustrated in Fig. 4.6c. To simulate the small perturbation, an additive perturbation was applied to the phase of each generator in steady-state  $t = 1$  s, drawn from the Gaussian distribution  $\mathcal{N}(0, 0.01)$ . We assume that a single attack replaces the transmitted measurement of some oscillator phase  $\phi_a(t) \in \mathcal{C}$  with some false data  $\tilde{\phi}_a(t)$ —which, for illustration purposes, we assume to be copied from some another neighboring measurement:  $\tilde{\phi}_a(t) = \{\phi_j(t) : \phi_j \notin \mathcal{C}, K_{aj} > 0\}$ . Data is reconstructed from the state estimation  $\hat{\phi}_a(t)$  of multiple functional observers, each one designed from a distinct subset  $\mathcal{S}' \subset \mathcal{S}$ , with cardinality  $|\mathcal{S}'| = |\mathcal{S}|/2$ . For all simulations, observers were designed using knowledge of the system dynamics (4.16)–(4.17) linearized around the equilibrium point (steady-state) and a subset of measurements  $\mathcal{S}'$  as defined above, with the target set defined as  $\mathcal{T} = \mathcal{C}$ . We assume that  $\hat{\phi}_a(t) = \phi_a(t)$  for  $t < 1$  s (before perturbation).

We show that this cyber-attack can be effectively detected by designing a stable functional observer and cross-validating the transmitted measurements against the state estimates during the short transient dynamics. Since one has no access to the true state estimation error  $\phi_a(t) - \hat{\phi}_a(t)$ , such cross-validation is actually performed

statistically, relying on the state estimation of *multiple* distinct functional observers designed from  $\mathcal{S}' \subset \mathcal{S}$ , as shown in Fig. 4.6d. We define the root mean-square error (RMSE) as

$$\text{RMSE} = \sqrt{\frac{1}{t_d} \int_1^{t_d+1} \|\phi_a(t) - \hat{\phi}_a(t)\|^2 dt}, \quad (4.18)$$

where the RMSE is a function of the time of detection window (e.g., the gray box in Fig. 4.6c). Simulations are shown for 100 Monte Carlos runs, where the additive perturbations to the system were randomly drawn from  $\mathcal{N}(0, 0.01)$  in each run, and 100 distinct functional observers were designed in each run (as illustrated in Fig. 4.6c). Therefore, each histogram comprises 10,000 data points.

The performance of methods for cyber-attack detection depends on the *time of detection* window, i.e., the time it takes to reliably detect an attack after it is launched, or equivalently, the interval of time under which the system is left *unprotected* waiting for a decision. Fig. 4.6d shows that, after a short period of time ( $\approx 0.25$ s), the separation between the histograms corresponding to the attacked and un-attacked systems becomes statistically significant so that reliable judgement can be made. Since the separation between the two histograms becomes even stronger as time increases, the judgement can be made more robust if a larger detection window is allowed.

In applications with multiple and constantly changing operation points, such as smart power grids, algorithms for the design of controllers and observers have to be sufficiently fast so that they can be implemented in real-time after a subsequent change of equilibrium (operation point). A cyber-attack detection framework can only be performed statistically if the algorithms for the observer design are sufficiently fast so that *multiple* observers can be implemented in a relatively short amount of time for real-time use. Algorithm 2 provides a fast and scalable solution for the design of a minimum-order functional observer, which effectively allows one to design significantly more functional observers in a same time frame. Figure 4.6e shows that for the same number of PMUs  $|\mathcal{S}|$ , the functional observer usually has a much smaller dimension than the Luenberger observer, and thereby has a significantly smaller running time for design. This has a cumulative impact on the computational resources for such application, specially when hundreds of (functional) observers are desired to provide estimates. Despite this large improvement in computational costs, no significant differences between the observers performance (state estimation error) are perceptible (Fig. 4.6e). These conclusions reassure the findings in Figs. 4.4 and 4.5, and, as in Fig. 4.5d, we can expect that the larger the power grid size, the more expressive the computational resources used by functional and Luenberger observers.

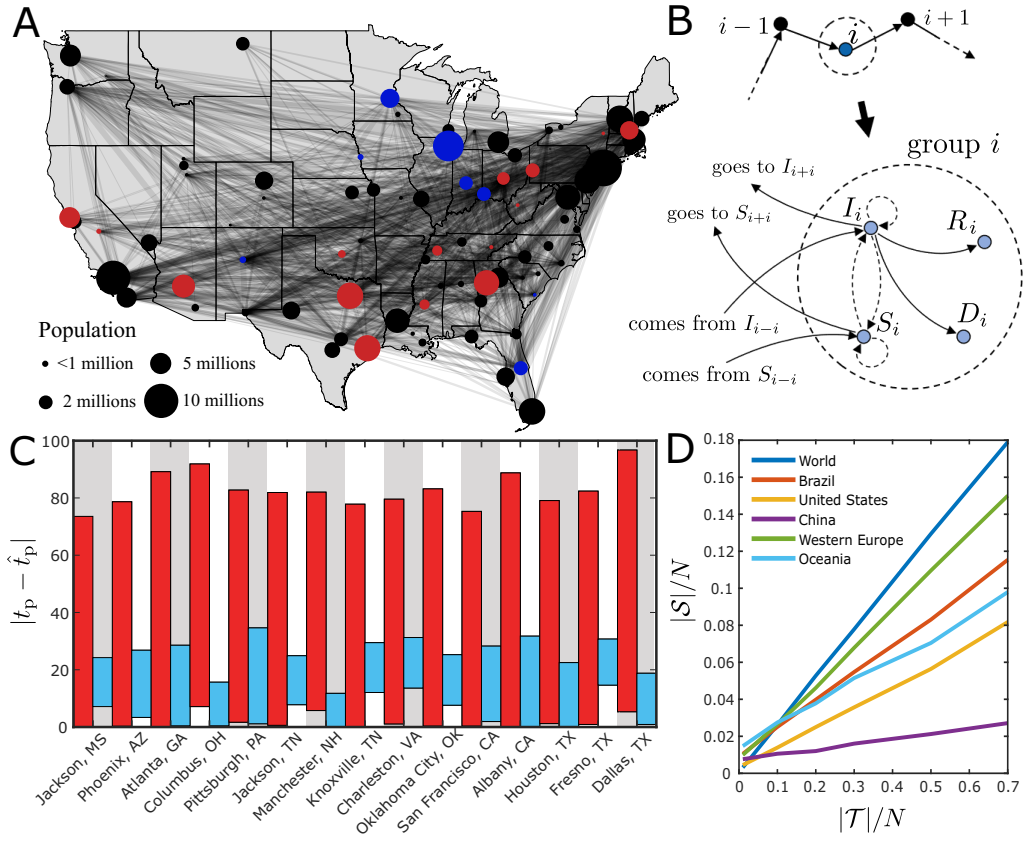
## 4.5.2 Estimation of epidemic spreading in target populations

Motivated by the unprecedented spreading of the coronavirus SARS-CoV-2 and the increasing number of fatalities associated with the COVID-19 disease, recent works in the literature highlighted the importance of epidemiological models for containment measures against the epidemic spreading, which often come at significant socioeconomic costs. For instance, epidemiological models are useful for our understanding of growth patterns and scaling laws governing the epidemic spreading (Blasius, 2020; Singer, 2020), as well as for the development of optimal control strategies (Lesniewski, 2020; Morris et al., 2020; Tsay et al., 2020), which ultimately support the decision of *when* social distancing and quarantine policies should be implemented (Hethcote, 2000). Such policy-making strategies depend on the knowledge of the initial and/or current state of the epidemic in a population. However, when a pandemic like COVID-19 starts, it takes time to mobilize medical resources and ramp up testing capacity. At the initial stage of the pandemic, it is almost impossible to ensure every city to have sufficient testing. Therefore, it is important to develop the ability to infer as much information as possible from available testing data. Previous results showed that state estimators can infer the true current state of the number of infected, susceptible and recovered individuals to some disease (Iggidr and Souza, 2019).

In this work, we show how functional observers can be designed to provide a state estimation of the infected population in a set of “target” cities (where testing is inadequate) from the known infection rate of a set of “sensor” cities (where sufficient testing has been done). To this end, consider the following multi-group epidemiological model (Colizza et al., 2006):

$$\begin{cases} \dot{S}_i = -\beta_i \frac{S_i I_i}{P_i} - \sum_{j=1, j \neq i}^N [A_{\text{adj}}]_{ij} \frac{S_i}{P_i} + \sum_{j=1, j \neq i}^N [A_{\text{adj}}]_{ji} \frac{S_j}{P_j}, \\ \dot{I}_i = \beta_i \frac{S_i I_i}{P_i} - \gamma_i I_i - \sum_{j=1, j \neq i}^N [A_{\text{adj}}]_{ij} \frac{I_i}{P_i} + \sum_{j=1, j \neq i}^N [A_{\text{adj}}]_{ji} \frac{I_j}{P_j}, \\ \dot{R}_i = (1 - \eta) \gamma I_i, \\ \dot{D}_i = \eta \gamma I_i, \end{cases} \quad (4.19)$$

for  $i = 1, \dots, N$ , where  $N$  is the number of groups (populations, nodes),  $(S_i, I_i, R_i, D_i)$  are the susceptible, infected, recovered and dead (SIRD) individuals of group  $i$  with population size  $P_i = S_i + I_i + R_i + D_i$ ,  $(\gamma, \eta)$  are the recovery and fatality rate,  $\beta_i$  is the contact rate per group  $i$ , and  $A_{\text{adj}}$  is the adjacency matrix of a transportation



**Figure 4.7:** Target state estimation in epidemics. (a) Air traffic network of the United States, where nodes represent cities and edges represent the domestic flights between two cities. Target cities are highlighted in red, and the minimum set of sensor cities required for functional observability is highlighted in blue. (b) Transportation network flow between three groups (top) and its corresponding structural graph representation (bottom) as a dynamical network described by (4.19). The nodal dynamics are taken into account by expanding each group  $i$  as a set of SIRD states, while edges represent the linear (solid lines) and nonlinear (dashed lines) interactions between the state variables in the ODE. (c) Difference between the time instant  $t_p$  when the epidemic peak actually happens in each target city and the predicted  $\hat{t}_p$  time instant (in days). Red bars show predictions (per city) done by numerically integrating (4.19) with arbitrary initial conditions, while blue bars show predictions given by the estimates of a functional observer. (d) Minimum sensor placement  $|S|/N$  as a function of the number of targets  $|\mathcal{T}|/N$  for different countries and regions with distinct air transportation networks. Simulations are presented for an average of 100 Monte Carlo runs, where target cities were randomly chosen.

network where  $[A_{\text{adj}}]_{ij}$  describes the number of people traveling from group  $i$  to  $j$  on a daily basis.

In this study, we assume each group to be an individual city and the adjacency matrix  $A_{\text{adj}}$  to describe the air traffic network between the cities' airports (Fig. 4.7a), which is defined according to the TranStats database for international and domestic

flights (all carriers) (Tra, 2004). Multiple airports belonging to the same city are combined in a single group  $i$  (node), with the corresponding city's population size  $P_i$ . Parameters in (4.19) were set as  $(\beta, \gamma, \eta) = (0.4, 0.16, 0.01)$  in order to mimic the coronavirus spreading in a given group, according to results reported in (Blasius, 2020). To incorporate the heterogeneity of each group in the simulation, we define the contact rate of each group as  $\beta_i \sim \mathcal{N}(\beta, 0.01)$ . In the study of the United States air traffic, only domestic flights were considered, while in the study of sets of countries, as in Fig. 4.7d, only flights within the countries and regions were considered.

Simulations in Fig. 4.7 are shown for 100 Monte Carlo runs, where we assume that the epidemic outbreak was in Miami, setting the initial conditions  $I_m(0) = 100$  and  $S_m(0) = P_m - 100$ , where  $I_m$  represents the infected population in Miami. To simulate the lack of information about the outbreak start (first infection), the simulation-based predictions and the functional observers were initiated with a false guess of  $I_j(0) = 1$  and  $S_j(0) = P_j - 1$ , where  $j$  is a random city chosen in each Monte Carlo run.

Ideally, if the true initial conditions of an epidemic were known, containment measures to “flatten the curve” could be established based on a straightforward simulation of model (4.19). Unfortunately, this is not the case, because not only the current number of infected individuals in a given group might be wrong but also the starting point of the outbreak might be far from the assumed. This carries larger errors to the simulations of “when” the epidemic peak in a given population group happens, shown by the red bars in Fig. 4.7c. To circumvent these limitations, we design a functional observer to provide more reliable estimates of the number of infected individuals in a set of 15 “target” cities. To find the minimum set of sensor nodes required for the system functional observability and actually design a minimum order functional observer for the nonlinear system (4.19), we represent the linear and nonlinear interactions in (4.19) as a structural graph (as in Fig. 4.7b), thereafter using Algorithms 1 and 2.

In more details, we show how the structural results derived in this paper for the minimum sensor placement and the minimum functional observer design of large-scale linear systems can be extended to a nonlinear model such as (4.19). Firstly, we draw a “nonlinear” graph representation  $\bar{\mathcal{G}} = \{\bar{\mathcal{X}}, \bar{\mathcal{E}}\}$  of (4.19) (Fig. 4.7B), where  $S_i, I_i, R_i, D_i \in \bar{\mathcal{X}}$  is the set of state nodes and  $(x_i, x_j) \in \bar{\mathcal{E}}$  is a directed edge from  $x_j$  to  $x_i$  if  $\dot{x}_i$  is an explicit function of  $x_j$  in (4.19), as described in Section 3.3.3. Let the (unweighted) adjacency matrix  $\bar{A}$  be a representation of  $\bar{\mathcal{G}}$ . Secondly, we define the set of “target” cities whose number of infected individuals are desired to be known (i.e.,  $I_i \in \mathcal{T}$  if  $i$  is a target city). Thirdly, we define the set of candidate nodes  $\mathcal{C}$  available



for sensor placement. In this simulation, we assume that all cities are available for complete medical testing, but only associated with the number of fatalities due to the epidemic disease, thus  $\mathcal{C} = \{D_1, \dots, D_N\}$ . Note that because there is always a path from  $I_i$  to  $D_i$ ,  $I_i$  is always structurally functionally observable from  $D_i$ . Fourthly, given  $(\bar{\mathcal{G}}, \mathcal{T}, \mathcal{C})$ , we use Algorithm 1 to find the minimum sensor placement  $\mathcal{S} \subseteq \mathcal{C}$  necessary to guarantee structural functional observability with respect to  $\mathcal{T}$  (if possible). Let  $C$  and  $F$  be the corresponding matrices to  $\mathcal{S}$  and  $\mathcal{T}$ , as in Fig. 4.1. Fifthly, given  $(\bar{A}, C, F)$ , we use Algorithm 2 to find the structure of the minimum order functional observer required to estimate  $\mathcal{T}$  (i.e.,  $F_0$ ). And finally, we use the results derived in (Trinh et al., 2006) to determine the functional observer parameters with Algorithm 6 (see Section B.3 for more details).

Figure 4.7a shows the minimum set of 8 sensor nodes required for the system to be functionally observable, while the blue bars in Fig. 4.7c show the time estimate of when the epidemic peak happens in the target cities. We define the time instant  $[t_p]_i$  when the epidemic peak happens in a given target city  $i$  as  $[t_p]_i = \arg \max_t I_i(t)$ , and the corresponding estimated time instant  $\hat{t}_p$  as  $[\hat{t}_p]_i = \arg \max_t \hat{I}_i(t)$ , where  $\hat{I}_i(t)$  is the estimated infected population over time given by: (i) the free simulation of (4.19) with false initial conditions, and (ii) the functional observer simulation.

As seen in Fig. 4.7c, there is a great improvement in the estimation accuracy with a smaller deviation between realizations of different initial conditions, which highlights the observer resilience to false initial predictions. *Note that, in this case, a functional observer is designed in a situation where the system is unobservable but functionally observable.* This highlights a fundamental advantage of our approach: when the conventional full-state observers are not applicable to the unobservable epidemic dynamics, our approach still provides high-quality estimation of the current state of the epidemics with limited information from a small number of “sensor” cities.

In addition, there is an interesting fact that, different from most networks studied in Fig. 4.3, the minimum number of sensors in this case is not a logarithmic function but rather a linear function of the number of targets. This is illustrated in Fig. 4.7d using the transportation network data of different countries (regions), where the slope of this function is shown to be dependent on the structural properties of the countries’ transportation network. Air traffic networks with high connectivity, such as the Chinese one, are shown to require smaller sets of sensor nodes to guarantee functional observability, compared to others more sparsely connected networks such as the Brazilian, European and the global transportation networks—which are intrinsically affected by their geopolitical interests. Once again, the network structural connectivity

has a fundamental impact in the functional observability of a system, including the number of sensor nodes (as seen in Fig. 4.3) and the functional observer order (as seen in Fig. 4.4d).

## 4.6 Conclusion

In many large-scale complex networks, it is physically impossible to ensure complete observability and computationally prohibitive to design full-state observers, posing fundamental challenges to our ability to observe, understand, and control the dynamical processes. On the other hand, many practical applications only need the observation of a small subset of key state variables, i.e. requiring the system to be functionally observable. Our theoretical work establishes a connection between the functional observability and the network structure, enabling highly scalable graph algorithms to optimally allocate sensors and design functional observers that achieve accurate estimation of a target subset of system variables. It is noteworthy that the concept of “target observability” (or, by duality, target controllability) was previously explored in the literature (Commault et al.; Czeizler et al., 2018; Gao et al., 2014; Klickstein et al., 2017; Li et al., 2019; Wu et al., 2015), where conditions were proposed for the existence of an estimator (controller) capable of estimating (controlling) a selected subset of target nodes. Though it appears similar, it is fundamentally different from the functional observability property: the target observability theory does not lead to an order reduction in the designed state observers, still having the same computational burden as the conventional full-state observers (see Appendix D for a more detailed analysis).

Hence, our results advance the theory and methods for state estimation on large-scale dynamical networks, which could have significant implications to cyber-physical systems, metabolic engineering, drug re-purposing, autonomous robotics, intervention ecology, chemical engineering, etc. For example, intervention measures designed to be taken on specific genetic biomarkers, metabolites or ecological populations often require precise quantification of such states. Even though physically accessible, such variables cannot be directly measured without the use of equipment or human-based actions that interfere with the underlying process or environment, compromising the intervention measures to be taken thereafter. Our results provide a framework under which such target nodes can be estimated from measures taken from different variables that do not interfere with the control actions, dismissing the need for direct measures. On a different front, our results also allows technological supply networks and multi-

agent systems to benefit economically from a smaller number of sensor units, and the communication infrastructure built-in, while relying on a robust state estimation of the system state for feedback control and operation purposes.

This work also gives rise to further fundamental questions worthy pursuing in the future. First, the functional observability is a qualitative concept which does not yield a quantitative measure of how functionally observable the system is. By developing a quantitative metric of functional observability, Algorithm 1 can be re-designed to maximize the functional observability metric, so that the estimation accuracy and dynamical performance of observers are accounted for during the sensor placement. Second, our application example in epidemiology illustrates how our knowledge of the model structure can be used to design functional observers in nonlinear systems. But our analysis is based on a specific system model and how our method can be extended to more general nonlinear systems requires further studies. Third, the cyber-attack detection problem can be extended to more general problem of how to design a communication network resilient to cyber-attacks in a large-scale cyber-physical system. We believe the graph-theoretical approach to functional observability developed in this manuscript could shed light on this intriguing problem as well. Finally, the design of functional observers is based on the knowledge of accurate system models. What if only the system structural model is available but its specific parameters are unknown? How can we still design and implement functional observers that achieve accurate estimation of the internal state variables of the system? Further empowered by machine learning techniques, our graph-based theory and methods could enable plug-and-play state estimation for complex systems without prior knowledge of the system models.



# Chapter 5

## Conclusion

The structural approach to analyze the observability and controllability of large-scale network systems generated interest in applying it to the most diverse areas. New findings came out in the past decade from combining theoretical results and techniques developed in the context of control theory, graph theory and, more recently, network theory. However, as more works find applications away from technological infrastructures with well-defined and precise models (e.g., power grids, multi-agent systems) to more broad network applications with imprecise models and parameters (e.g., neuronal networks, food webs, epidemiological models), it becomes increasingly more important to avoid pitfalls that might arise from such rich combination of fields.

Observability is a property originally defined in control theory with low-dimensional systems in mind. Albeit a crisp “yes-no” definition in origin, this property is a discontinuous function of the systems parameters, where a small change in the parameter space can move a dynamical system from observable to unobservable, and vice-versa. The classical definition of observability, therefore, becomes “trickier” as the system dimension grows large and the number of parameters increase, especially when the modelling uncertainty is reasonably large (see Sections [3.1](#) and [3.2](#)).

The graph-based approach to observability, entitled structural observability, circumvents this problem by establishing a definition that depends solely on the system structure, without taking into account the specific values of the parameters. Not only the static structure of a network is usually a highly reliable information gathered in large datasets when compared to the system dynamical information, but the structural approach also provides a more reliable characterization by guaranteeing the observability of a system even in the case of “missing links” in the system modelling. Moreover, it provides an intuitive framework where graph-based techniques can be applied to

solve problems related to the observability of a system, such as the optimal sensor placement and state observer design (see Section 3.3).

The structural observability avoids fundamental pitfalls related to the dimensionality of a large-scale network system, but it also creates some other problems of its own. For instance, it does not take into account symmetries inherent to the dynamical system, and the optimal sensor placement determined by algorithms grounded on this formulation does not guarantee that the state estimation of a high-dimensional state vector will perform within the desired estimation error for any given application. As seen in Section 3.6, if the system dimension is reasonably large, a dynamical system observed from a single node, even if observable, is practically speaking almost unobservable from a conditioning point-of-view. Finally, observability only guarantees the existence of a  $n$ -dimensional state estimator, which is usually computationally intractable for its design or expensive for real-time applications.

Chapter 3 investigates the advantages and disadvantages of different approaches to investigate the observability in dynamical systems. In order to avoid most of the pitfalls exposed throughout this text, in Chapter 4, we explore upon the concept of structural observability generalizing its property to the notion of structural functional observability. By establishing graph-theoretical conditions under which only a given subset of targeted state variables can be reconstructable from a set of sensor nodes, we circumvent the curse of dimensionality by reducing the dimension of the state vector to be estimated. Grounded on this theoretical development, we develop two algorithms for the optimal sensor placement and minimum-order functional observer design. Although the sensor placement algorithm shares the same disadvantages of other structural-based approaches, it is capable—in the context of directed networks—of returning smaller sets of sensor nodes depending on the number of target variables. Likewise, the designed functional observer achieves significantly smaller dimension compared to traditional full-state observers, being more computationally efficient for large-scale applications while also achieving the same estimation quality. The interdisciplinary problems explored in Section 4.5 are an exemplification of the broad range of network systems over which the developed techniques can find application.

Better techniques for state estimation in high-dimensional systems can further advance our capability of achieving precise control of large-scale networks, and we believe that the results presented in this thesis were able to address some of the challenges in this field and may provide interesting applications in a broad scope of network applications, ranging from the development of cyber-security infrastructures to the medical diagnosing of potential biomarkers in biological systems.

# References

- TranStats: the intermodal transportation database. United States Bureau Of Transportation Statistics, 2004. URL <http://transtats.bts.gov>.
- Lesson Learned: Risks Posed by Firewall Firmware Vulnerabilities. Technical report, North American Electric Reliability Corporation, 2019.
- L. A. Aguirre. Controllability and Observability of Linear Systems: Some Noninvariant Aspects. *IEEE Transactions on Education*, 38(1):33–39, 1995.
- L. A. Aguirre and C. Letellier. Observability of multivariate differential embeddings. *Journal of Physics A: Mathematical and General*, 38(28):6311–6326, 2005.
- L. A. Aguirre and C. Letellier. Investigating observability properties from data in nonlinear dynamics. *Physical Review E*, 83:066209, 2011.
- L. A. Aguirre, S. B. Bastos, M. A. Alves, and C. Letellier. Observability of nonlinear dynamics: Normalized results and a time-series approach. *Chaos*, 18:013123, 2008.
- L. A. Aguirre, L. L. Portes, and C. Letellier. Observability and synchronization of neuron models. *Chaos*, 27:103103, 2017.
- L. A. Aguirre, L. L. Portes, and C. Letellier. Structural, dynamical and symbolic observability: From dynamical systems to networks. *PloS ONE*, 13(10):e0206180, 2018.
- S. Amin, A. Cárdenas, and S. Sastry. Safe and Secure Networked Control Systems under Denial-of-Service Attacks. *Hybrid Systems: Computation and Control*, 5469: 31–45, 2009.
- A. Arenas, A. Díaz-Guilera, J. Kurths, Y. Moreno, and C. Zhou. Synchronization in complex networks. *Physics Reports*, 469(3):93–153, 2008.
- L. Bakule. Decentralized control: An overview. *Annual Reviews in Control*, 32:87–98, 2008.
- T. Baldwin, L. Mili, M. Boisen, and R. Adapa. Power-system observability with minimal phasor measurement placement. *IEEE Transactions on Power Systems*, 8(2):707–715, 1993.
- A.-L. Barabási. Emergence of Scaling in Random Networks. *Science*, 286(5439): 509–512, 1999.

- A.-L. Barabási. Scale-Free Networks: A Decade and Beyond. *Science*, 325(5939): 412–413, 2009.
- A.-L. Barabási and M. Pósfasi. *Network Science*. Cambridge University Press, 1st edition, 2016.
- A.-L. Barabási, N. Gulbahce, and J. Loscalzo. Network medicine: A network-based approach to human disease. *Nature Reviews Genetics*, 12:56–68, 2011.
- E. Bianco-Martinez, M. S. Baptista, and C. Letellier. Symbolic computations of nonlinear observability. *Physical Review E*, 91:06912, 2015.
- B. Blasius. Power-law distribution in the number of confirmed COVID-19 cases. *arXiv:2004.00940*, 2020.
- S. Boccaletti, J. Kurths, G. Osipov, D. Valladares, and C. Zhou. The synchronization of chaotic systems. *Physics Reports*, 366(1–2):1–101, 2002.
- S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4–5):175–308, 2006.
- N. Bof, G. Baggio, and S. Zampieri. On the role of network centrality in the controllability of complex networks. *IEEE Transactions on Control of Network Systems*, 4(3):643–653, 2017.
- R. Bouffanais. *Design and Control of Swarm Dynamics*. Springer, 2016.
- N. Bretas and J. London. Network observability: the critical measurement identification using the symbolic Jacobian matrix. *POWERCON '98. 1998 International Conference on Power System Technology. Proceedings*, pages 1222–1226, 1998.
- F. Bullo. *Lectures on Network Systems*. CreateSpace, 2016.
- T. L. Carroll. Testing Dynamical System Variables for Reconstruction. *Chaos*, 28: 103117, 2018.
- E. Castillo, P. Jiménez, J. M. Menéndez, and A. J. Conejo. The observability problem in traffic models: Algebraic and topological methods. *IEEE Transactions on Intelligent Transportation Systems*, 9(2):275–287, 2008.
- A. Chapman, M. Nabi-Abdolyousefi, and M. Mesbahi. Controllability and Observability of Network-of-Networks via Cartesian Products. *IEEE Transactions on Automatic Control*, 59(10):2668–2679, 2014.
- C.-T. Chen. *Linear System Theory and Design*. Oxford University Press, 3rd edition, 1999.
- D. Chen and H. Qi. Controllability and observability of Boolean control networks. *Automatica*, 45(7):1659–1667, 2009.
- G. Chen, X. Wang, and X. Li. *Fundamentals of complex networks*. Wiley, 2013.



- G.-R. Chen. Problems and Challenges in Control Theory under Complex Dynamical Network Environments. *Acta Automatica Sinica*, 39(4):312–321, 2014.
- V. Colizza, A. Barrat, M. Barthelemy, and A. Vespignani. The role of the airline transportation network in the. *PNAS*, 103(7):2015–2020, 2006.
- C. Commault, J. Van der Woude, and P. Frasca. Functional target controllability of networks: structural properties and efficient algorithms. *IEEE Transactions on Network Science and Engineering*, 7:1521–1530.
- S. P. Cornelius, W. L. Kath, and A. E. Motter. Realistic control of network dynamics. *Nature Communications*, 4:1942, 2013.
- L. d. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242, 2007.
- N. J. Cowan, E. J. Chastain, D. A. Vilhena, J. S. Freudenberg, and C. T. Bergstrom. Nodal dynamics, not degree distributions, determine the structural controllability of complex networks. *PLoS ONE*, 7(6):e38398, 2012.
- M. A. d. R. Cruz and H. R. O. Rocha. Planning Metering for Power Distribution Systems Monitoring with Topological Reconfiguration. *Journal of Control, Automation and Electrical Systems*, 28:135–146, 2017.
- 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*, 15(4):1217–1228, 2018.
- M. Darouach. Existence and Design of Functional Observers for Linear Systems. *IEEE Transactions on Automatic Control*, 45(5):940–943, 2000.
- F. Derakhshan and S. Yousefi. A review on the applications of multiagent systems in wireless sensor networks. *International Journal of Distributed Sensor Networks*, 15(5), 2019.
- F. Dorfler and F. Bullo. Synchronization in Complex Networks of Phase Oscillators: A Survey. *Automatica*, 50(6):1539–1564, 2014.
- F. Dorfler, M. Chertkov, and F. Bullo. Synchronization in complex oscillator networks and smart grids. *PNAS*, 110(6):2005–2010, 2013.
- J. Duch and A. Arenas. Community detection in complex networks using extremal optimization. *Physical Review E*, 72:027104, 2005.
- J. A. Dunne, R. J. Williams, and N. D. Martinez. Food-web structure and network theory: The role of connectance and size. *PNAS*, 99(20):12917–12922, 2002.
- D. Eroglu, J. S. Lamb, and T. Pereira. Synchronisation of chaos and its applications. *Contemporary Physics*, 58(3):207–243, 2017.
- J. P. Farwell and R. Rohozinski. Stuxnet and the future of cyber war. *Survival*, 53(1): 23–40, 2011.

- T. Fernando and H. Trinh. A system decomposition approach to the design of functional observers. *International Journal of Control*, 87(9):1846–1860, 2014.
- T. Fernando, L. Jennings, and H. Trinh. Numerical implementation of a functional observability algorithm: A singular value decomposition approach. *IEEE Asia-Pacific Conference on Circuits and Systems (APCCAS)*, pages 796–799, 2010a.
- T. L. Fernando, H. M. Trinh, and L. Jennings. Functional Observability and the Design of Minimum Order Linear Functional Observers. *IEEE Transactions on Automatic Control*, 55(5):1268–1273, 2010b.
- B. Fiedler, A. Mochizuki, G. Kurosawa, and D. Saito. Dynamics and Control at Feedback Vertex Sets. I: Informative and Determining Nodes in Regulatory Networks. *Journal of Dynamics and Differential Equations*, 25(3):563–604, 2013.
- S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.
- B. Friedland. Controllability index based on conditioning number. *Journal of Dynamic Systems, Measurement, and Control*, 97(4):444–445, 1975.
- Y. Fu, L. Wang, and M. Chen. Robustness of Controllability for Scale-free Networks Based on a Nonlinear Load-Capacity Model. *IFAC Proceedings Volumes*, 49(4):37–42, 2016.
- J. Gao, Y.-Y. Liu, R. M. D’Souza, and A.-L. Barabási. Target control of complex networks. *Nature Communications*, 5:5415, 2014.
- A. J. Gates and L. M. Rocha. Control of complex networks requires both structure and dynamics. *Scientific Reports*, 6:24456, 2015.
- L. J. Gilarranz, B. Rayfield, G. Liñán-Cembrano, J. Bascompte, and A. Gonzalez. Effects of network modularity on the spread of perturbation impact in experimental metapopulations. *Science*, 357:199–201, 2017.
- J. Giraldo, D. Urbina, A. Cardenas, J. Valente, M. Faisal, J. Ruths, N. O. Tippenhauer, H. Sandberg, and R. Candell. A survey of physics-based attack detection in cyber-physical systems. *ACM Computing Surveys*, 51(4):76, 2018.
- S. Gu, F. Pasqualetti, M. Cieslak, Q. K. Telesford, A. B. Yu, A. E. Kahn, J. D. Medaglia, J. M. Vettel, M. B. Miller, S. T. Grafton, and D. S. Bassett. Controllability of structural brain networks. *Nature Communications*, 6:8414, 2015.
- J. Guan, T. Berry, and T. Sauer. Limits on reconstruction of dynamical networks. *Physical Review E*, 98:022318, 2018.
- Y. Guan and L. Wang. Controllability of multi-agent systems with directed and weighted signed networks. *Systems and Control Letters*, 116:47–55, 2018.
- R. Gutiérrez, I. Sendiña-Nadal, M. Zanin, D. Papo, and S. Boccaletti. Targeting the dynamics of complex networks. *Scientific reports*, 2:396, 2012.
- A. Haber and M. Verhaegen. Subspace identification of large-scale interconnected systems. *IEEE Transactions on Automatic Control*, 59(10):2754–2759, 2014.

- A. Haber, F. Molnar, and A. E. Motter. State Observation and Sensor Selection for Nonlinear Networks. *IEEE Transactions on Control of Network Systems*, 5(2):694 – 708, 2018.
- C. Hammond, H. Bergman, and P. Brown. Pathological synchronization in Parkinson’s disease: networks, models and treatments. *Trends in Neurosciences*, 30(7):357–364, 2007.
- R. Hermann and A. J. Krener. Nonlinear Controllability and Observability. *IEEE Transactions on Automatic Control*, 22(5):728–740, 1977.
- H. W. Hethcote. The Mathematics of Infectious Diseases. *SIAM Review*, 42(4):599–653, 2000.
- T. Hieu and F. Tyrone. *Functional Observers for Dynamical Systems*. Springer Berlin Heidelberg, 2012.
- A. Iggidr and M. O. Souza. State estimators for some epidemiological systems. *Journal of Mathematical Biology*, 78(1-2):225–256, 2019.
- F. L. Iudice, F. Sorrentino, and F. Garofalo. On Node Controllability and Observability in Complex Dynamical Networks. *IEEE Control Systems Letters*, 3(4):847–852, 2019.
- E. M. Izhikevich. Which model to use for cortical spiking neurons? *IEEE Transactions on Neural Networks*, 15(5):1063–1070, 2004.
- L. S. Jennings, T. L. Fernando, and H. M. Trinh. Existence conditions for functional observability from an eigenspace perspective. *IEEE Transactions on Automatic Control*, 56(12):2957–2961, 2011.
- T. Jia, Y.-Y. Liu, E. Csoka, M. Posfai, J.-J. Slotine, and A.-L. Barabási. Emergence of bimodality in controlling complex networks. *Nature Communications*, 4:2002, 2013.
- J. Jiang and Ying-Cheng Lai. Irrelevance of linear controllability to nonlinear dynamical networks. *Nature Communications*, 10:3961, 2019.
- C. D. Johnson. Optimization of a certain quality of complete controllability and observability for linear dynamical systems. *Journal of Basic Engineering*, 91(2): 228–238, 1969.
- R. Kalman. On the general theory of control systems. *IRE Transactions on Automatic Control*, 4(3):110–110, 1959.
- E. H. Kerner. Universal formats for nonlinear ordinary differential systems. *Journal of Mathematical Physics*, 22(7):1366–1371, 1981.
- H. K. Khalil. *Nonlinear Systems*. Prentice Hall, 3rd edition, 2002.
- R. Klaus and K. J. Reinschke. An efficient method to compute Lie derivatives and the observability matrix for nonlinear systems. *Int. Symposium on Nonlinear Theory and its Applications (NOLTA)*, 2, 1999.

- I. Klickstein, A. Shirin, and F. Sorrentino. Energy scaling of targeted optimal control of complex networks. *Nature Communications*, 8:15145, 2017.
- M. Komareji and R. Bouffanais. Controllability of a swarm of topologically interacting autonomous agents. *International Journal of Complex Systems in Science*, 3(1): 11–19, 2014.
- S. K. Korovin and V. V. Fomichev. *State Observers for Linear Systems with Uncertainty*. Walter de Gruyter, 1st edition, 2009.
- J. Kunegis. KONECT - The Koblenz Network Collection. *Proceedings of the International Web Observatory Workshop*, pages 1343–1350, 2013.
- Y. Kuramoto. Self-entrainment of a population of coupled non-linear oscillators. *International Symposium on Mathematical Problems in Theoretical Physics*, pages 420–422, 1975.
- D. Laschov, M. Margaliot, and G. Even. Observability of Boolean networks: A graph-theoretic approach. *Automatica*, 49(8):2351–2362, 2013.
- R. Lee, M. Assante, and T. Conway. Analysis of the Cyber Attack on the Ukrainian Power Grid. Technical report, SANS Industrial Control Systems, 2016.
- K. Lehnertz, S. Bialonski, M. T. Horstmann, D. Krug, A. Rothkegel, M. Staniek, and T. Wagnet. Synchronization phenomena in human epileptic brain networks. *Journal of Neuroscience Methods*, 183(183):42–48, 2009.
- D. Leitold, A. Vathy-Fogarassy, and J. Abonyi. Controllability and observability in complex networks – the effect of connection types. *Scientific Reports*, 7:151, 2017.
- D. Leitold, A. Vathy-Fogarassy, and J. Abonyi. Network-based Observability and Controllability Analysis of Dynamical Systems: the NOCAD toolbox. *F1000Research*, 8:646, 2019.
- A. Lesniewski. Epidemic control via stochastic optimal control. *arXiv:2004.06680*, 2020.
- C. Letellier and L. A. Aguirre. Investigating nonlinear dynamics from time series: The influence of symmetries and the choice of observables. *Chaos*, 12(3):549–558, 2002.
- C. Letellier and L. A. Aguirre. Symbolic observability coefficients for univariate and multivariate analysis. *Physical Review E*, 79:066210, 2009.
- C. Letellier, J. Maquet, L. L. Sceller, G. Gouesbet, and L. A. Aguirre. On the non-equivalence of observables in phase-space reconstructions from recorded time series. *Journal of Physics A: Mathematical and General*, 31:7913–7927, 1998.
- C. Letellier, L. A. Aguirre, and J. Maquet. Relation between observability and differential embeddings for nonlinear dynamics. *Physical Review E*, 71:066213, 2005.
- C. Letellier, I. Sendiña-Nadal, and L. A. Aguirre. A nonlinear graph-based theory for dynamical network observability. *Physical Review E*, 98:020303, 2018.

- A. Li, S. P. Cornelius, Y. Y. Liu, L. Wang, and A. L. Barabási. The fundamental advantages of temporal networks. *Science*, 358:1042–1046, 2017.
- J. Li, X. Chen, S. Pequito, G. J. Pappas, and V. M. Preciado. Resilient Structural Stabilizability of Undirected Networks. *2019 American Control Conference (ACC)*, pages 5173–5178, 2019.
- C. T. Lin. Structural Controllability. *IEEE Transactions on Automatic Control*, 19(3): 201–208, 1974.
- Y. Liu, P. Ning, and M. K. Reiter. False data injection attacks against state estimation in electric power grids. *ACM Transactions on Information and System Security*, 14(1):13, 2011a.
- Y. Y. Liu and A. L. Barabási. Control principles of complex systems. *Reviews of Modern Physics*, 88:035006, 2016.
- Y.-Y. Liu, J.-J. Slotine, and A.-L. Barabási. Controllability of complex networks. *Nature*, 473:167–73, 2011b.
- Y.-Y. Liu, J.-J. Slotine, and A.-L. Barabási. Observability of complex systems. *PNAS*, 110(7):2460–2465, 2013.
- A. Lombardi and M. Hörnquist. Controllability analysis of networks. *Physical Review E*, 75:056110, 2007.
- Z. Lu, L. Zhang, and L. Wang. Observability of Multi-Agent Systems with Switching Topology. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 64(11): 1317–1321, 2017.
- X. Luan and P. V. Tsvetkov. Novel consistent approach in controllability evaluations of point reactor kinetics models. *Annals of Nuclear Energy*, 131:496–506, 2019.
- G. Luenberger. Observers for multivariable systems. *IEEE Transactions on Automatic Control*, AC-II(2):190–197, 1966.
- O. P. Malik. Evolution of power systems into smarter networks. *Journal of Control, Automation and Electrical Systems*, 24:139–147, 2013.
- N. D. Martinez. Artifacts or attributes? Effects of resolution on the little rock lake food web. *Ecological Monographs*, 61:367–392, 1991.
- A. Mesbahi, J. Bu, and M. Mesbahi. Nonlinear Observability via Koopman Analysis: Characterizing the Role of Symmetry. *arXiv:1904.08449 [cs.SY]*, 2019.
- 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.
- H. P. Mirsky, A. C. Liu, D. K. Welsh, S. A. Kay, and F. J. Doyle. A model of the cell-autonomous mammalian circadian clock. *PNAS*, 106(27):11107–11112, 2009.

- A. Mochizuki, B. Fiedler, G. Kurosawa, and D. Saito. Dynamics and control at feedback vertex sets. II: A faithful monitor to determine the diversity of molecular activities in regulatory networks. *Journal of Theoretical Biology*, 335:130–146, 2013.
- R. Mohajerpoor, H. Abdi, and S. Nahavandi. A new algorithm to design minimal multi-functional observers for linear systems. *Asian Journal of Control*, 18(3):842–857, 2016.
- A. N. Montanari and L. A. Aguirre. Particle filtering of dynamical networks: Highlighting observability issues. *Chaos*, 29:033118, 2019.
- A. N. Montanari and L. A. Aguirre. Observability of Network Systems: A Critical Review of Recent Results. *Journal of Control, Automation and Electrical Systems*, 31:1348–1374, 2020.
- A. N. Montanari, L. Freitas, L. A. B. Torres, and L. A. Aguirre. Phase synchronization analysis of bridge oscillators between clustered networks. *Nonlinear Dynamics*, 97(4):2399–2411, 2019.
- A. N. Montanari, E. I. Moreira, and L. A. Aguirre. Effects of network heterogeneity and tripping time on the basin stability of power systems. *Communications in Nonlinear Science and Numerical Simulation*, 89:105296, 2020.
- L. H. A. Monteiro. *Sistemas Dinâmicos Complexos*. Editora Livraria de Física, 2nd edition, 2014.
- A. Monticelli and F. F. Wu. Network Observability: Theory. *IEEE Transactions on Power Apparatus and Systems*, PAS-104(5):1042–1048, 1985.
- E. I. Moreira and L. A. Aguirre. Resiliência de Sistemas Elétricos de Potência Representados por Redes de Kuramoto. In *14<sup>o</sup> Simpósio Brasileiro de Automação Inteligente (SBAI)*, 2019.
- Y. Moreno and F. Pacheco. Synchronization of Kuramoto oscillators in scale-free networks. *Europhysics Letters*, 68(4):603–609, 2004.
- Y. Moreno, R. Pastor-Satorras, and A. Vespignani. Epidemic Outbreaks in Complex Heterogeneous Networks. *European Physical Journal B*, 26(4):521–529, 2002.
- D. H. Morris, F. W. Rossine, J. B. Plotkin, and S. A. Levin. Optimal, near-optimal, and robust epidemic control. *arXiv:2004.02209*, 2020.
- A. E. Motter. Networkcontology. *Chaos*, 25:097621, 2015.
- J. C. Nacher and T. Akutsu. Structural controllability of unidirectional bipartite networks. *Scientific Reports*, 3:1647, 2013.
- H. Nakao and A. S. Mikhailov. Turing Patterns in Network-Organized Activator-Inhibitor Systems. *Nature Physics*, 6:544–550, 2010.
- G. Nemhauser, L. Wolsey, and M. Fisher. An analysis of approximations for maximizing submodular set functions-I. *Mathematical Programming*, 14(1):265–294, 1978.

- T. Nepusz and T. Vicsek. Controlling edge dynamics in complex networks. *Nature Physics*, 8:568–573, 2012.
- M. Newman. *Networks: An Introduction*. OUP Oxford, 1st edition, 2010.
- M. E. Newman and D. J. Watts. Renormalization group analysis of the small-world network model. *Physics Letters, Section A: General, Atomic and Solid State Physics*, 263(4-6):341–346, 1999.
- T. Nishikawa and A. E. Motter. Comparative analysis of existing models for power-grid synchronization. *New Journal of Physics*, 17:015012, 2015.
- G. Notarstefano and G. Parlangeli. Controllability and observability of grid graphs via reduction and symmetries. *IEEE Transactions on Automatic Control*, 58(7):1719–1731, 2013.
- K. Ogata. *Modern Control Engineering*. Prentice Hall, 5th edition, 2010.
- T. Ohtsuka. Model structure simplification of Nonlinear Systems via immersion. *IEEE Transactions on Automatic Control*, 50:607–618, 2005.
- R. Olfati-Saber. Flocking for multi-agent dynamic systems: Algorithms and theory. *IEEE Transactions on Automatic Control*, 51(3):401–420, 2006.
- 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.
- A. Olshevsky. Minimal controllability problems. *IEEE Transactions on Control of Network Systems*, 1(3):249–258, 2014.
- J. O’Reilly. *Observers for Linear Systems*. Academic Press, 1st edition, 1983.
- P. Overschee and B. De Moor. *Subspace Identification for Linear Systems*. Springer US, 1st edition, 1996.
- S. P. Pang, W. X. Wang, F. Hao, and Y. C. Lai. Universal framework for edge controllability of complex networks. *Scientific Reports*, 7:4224, 2017.
- G. Parlangeli and G. Notarstefano. On the reachability and observability of path and cycle graphs. *IEEE Transactions on Automatic Control*, 57(3):743–748, 2012.
- F. Pasqualetti, F. Dorfler, and F. Bullo. Attack detection and identification in cyber-physical systems. *IEEE Transactions on Automatic Control*, 58(11):2715–2729, 2013a.
- F. Pasqualetti, S. Zampieri, and F. Bullo. Controllability, Limitations and Algorithms for Complex Networks. *IEEE Transactions on Control of Network Systems*, 1(1):40–52, 2013b.
- F. Pasqualetti, C. Favaretto, S. Zhao, and S. Zampieri. Fragility and Controllability Tradeoff in Complex Networks. *2018 Annual American Control Conference (ACC)*, pages 216–221, 2018.

- D. J. Pearce, A. M. Miller, G. Rowlands, and M. S. Turner. Role of projection in the control of bird flocks. *Proceedings of the National Academy of Sciences of the United States of America*, 111(29):10422–10426, 2014.
- L. M. Pecora and T. L. Carroll. Synchronization in chaotic systems. *Physical Review Letters*, 64(8):821–824, 1990.
- L. M. Pecora and T. L. Carroll. Master stability functions for synchronized coupled systems. *Physical Review Letters*, 80(10):2109–2112, 1998. ISSN 10797114.
- J. Peng, Y. Sun, and H. F. Wang. Optimal PMU placement for full network observability using Tabu search algorithm. *International Journal of Electrical Power and Energy Systems*, 28(4):223–231, 2006.
- F. Perini, E. Galligani, and R. D. Reitz. An analytical Jacobian approach to sparse reaction kinetics for computationally efficient combustion modeling with large reaction mechanisms. *Energy and Fuels*, 26(8):4804–4822, 2012.
- A. Phadke and J. Thorp. *Synchronized Phasor Measurements and Their Applications*. Springer US, 1st edition, 2008.
- M. Pósfai, Y.-Y. Liu, J.-J. Slotine, and A.-L. Barabási. Effect of correlations on controllability transition in network control. *Scientific Reports*, 3:1067, 2013.
- C. L. Pu, W. J. Pei, and A. Michaelson. Robustness analysis of network controllability. *Physica A: Statistical Mechanics and its Applications*, 391(18):4420–4425, 2012.
- J. Qi, K. Sun, and W. Kang. Optimal PMU Placement for Power System Dynamic State Estimation by Using Empirical Observability Gramian. *IEEE Transactions on Power Systems*, 30(4):2041–2054, 2015.
- A. Rahmani, M. Ji, M. Mesbahi, and M. Egerstedt. Controllability of Multi-Agent Systems from a Graph-Theoretic Perspective. *SIAM Journal on Control and Optimization*, 48(1):162–186, 2009.
- H. R. O. Rocha, J. A. Silva, J. C. de Souza, and M. B. Do Coutto Filho. Fast and Flexible Design of Optimal Metering Systems for Power Systems Monitoring. *Journal of Control, Automation and Electrical Systems*, 29(2):209–218, 2018.
- F. A. Rodrigues, T. K. Peron, P. Ji, and J. Kurths. The Kuramoto model in complex networks. *Physics Reports*, 610:1–98, 2016.
- O. E. RöSSLer. An equation for continuous chaos. *Physics Letters*, 57(5):397–398, 1976.
- F. Rotella and I. Zambettakis. A direct design procedure for linear state functional observers. *Automatica*, 70:211–216, 2016a.
- F. Rotella and I. Zambettakis. A note on functional observability. *IEEE Transactions on Automatic Control*, 61(10):3197–3202, 2016b.
- J. Ruths and D. Ruths. Control profiles of complex networks. *Science*, 343(6177):1373–1376, 2014.



- P. W. Sauer, M. A. Pai, and J. H. Chow. *Power System Dynamics and Stability: With Synchronphasor Measurement and Power System Toolbox*. Wiley, 2nd edition, 2017.
- B. Schäfer, D. Witthaut, M. Timme, and V. Latora. Dynamically induced cascading failures in power grids. *Nature Communications*, 9:1975, 2018.
- J. Schellenberger, J. O. Park, T. M. Conrad, and B. Ø. Palsson. Database BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics*, 11:213, 2010.
- P. H. T. Schimit and L. H. A. Monteiro. On the Basic Reproduction Number and the Topological Properties of the Contact Network: An Epidemiological Study in Mainly Locally Connected Cellular Automata. *Ecological Modelling*, 220(7):1034–1042, 2009.
- H. M. Singer. The COVID-19 pandemic: growth patterns, power law scaling, and saturation. *arXiv:2004.03859*, 2020.
- A. K. Singh and J. Hahn. Determining optimal sensor locations for state and parameter estimation for stable nonlinear systems. *Industrial and Engineering Chemistry Research*, 44(15):5645–5659, 2005.
- A. K. Singh and B. C. Pal. Decentralized dynamic state estimation in power systems using unscented transformation. *IEEE Transactions on Power Systems*, 29(2):794–804, 2014.
- J. Slay and M. Miller. Lessons Learned from the Maroochy Water Breach. *Critical Infrastructure Protection*, 253:73–82, 2007.
- J.-J. Slotine and Y.-Y. Liu. Complex networks: The missing link. *Nature Physics*, 8(7):512–513, 2012.
- I. W. Slutsker and J. M. Scudder. Network observability analysis through measurement jacobian matrix reduction. *IEEE Transactions on Power Systems*, 2(2):331–336, 1987.
- E. D. Sontag. Kalman’s Controllability Rank Condition: From Linear to Nonlinear. In A. C. Antoulas, editor, *Mathematical System Theory: The Influence of R. E. Kalman*, pages 453–462. Springer Berlin Heidelberg, 1991.
- T. Stankovski, T. Pereira, P. V. McClintock, and A. Stefanovska. Coupling functions: Universal insights into dynamical interaction mechanisms. *Reviews of Modern Physics*, 89:045001, 2017.
- F. Su, J. Wang, H. Li, B. Deng, H. Yu, and C. Liu. Analysis and application of neuronal network controllability and observability. *Chaos*, 27:023103, 2017.
- T. H. Summers, F. L. Cortesi, and J. Lygeros. On Submodularity and Controllability in Complex Dynamical Networks. *IEEE Transactions on Control of Network Systems*, 3(1):91–101, 2016.
- J. Sun and A. E. Motter. Controllability transition and nonlocality in network control. *Physical Review Letters*, 110:208701, 2013.

- S. Sundaram. *Fault-Tolerant and Secure Control Systems Acknowledgments*. Class notes, Department of Electrical and Computer Engineering, University of Waterloo, 2012.
- F. Takens. Detecting strange attractors in turbulence. In D. Rand and L. Young, editors, *Dynamical Systems and Turbulence*, pages 366–381. Springer, Berlin, Heidelberg, 1981.
- H. Tanner. On the controllability of nearest neighbor interconnections. *43rd IEEE Conference on Decision and Control*, pages 2467–2472, 2004.
- A. Teixeira, S. Amin, H. Sandberg, K. H. Johansson, and S. S. Sastry. Cyber security analysis of state estimators in electric power systems. *Proceedings of the IEEE Conference on Decision and Control*, pages 5991–5998, 2010.
- V.-k. Tran and H.-s. Zhang. Optimal PMU Placement Using Modified Greedy Algorithm. *Journal of Control, Automation and Electrical Systems*, 29(1):99–109, 2018.
- H. Trinh, T. Fernando, and S. Nahavandi. Partial-State Observers for Nonlinear Systems. *IEEE Transactions on Automatic Control*, 51(11):1808–1812, 2006.
- C. Tsay, F. Lejarza, M. A. Stadtherr, and M. Baldea. Modeling, state estimation, and optimal control for the US COVID-19 outbreak. *arXiv:2004.06291v1*, 2020.
- J. W. van der Woude. A graph-theoretic characterization for the rank of the transfer matrix of a structured system. *Mathematics of Control, Signals, and Systems*, 4(1):33–40, 1991.
- J. W. van der Woude. The generic number of invariant zeros of a structured linear system. *SIAM Journal on Control and Optimization*, 38(1):1–21, 1999.
- T. Vicsek, A. Czirok, E. Ben-Jacob, I. Cohen, and O. Shochet. Novel Type of Phase Transition in a System of Self-Driven Particles. *Physical Review Letters*, 75(6):1226–1229, 1995.
- M. Vidyasagar. *Nonlinear Systems Analysis*. Prentice Hall, 2nd edition, 1978.
- V. Vittal. Transient stability test systems for direct stability methods. *IEEE Transactions on Power Systems*, 7(1):37–43, 1992.
- S. Vivek, D. Yanni, P. J. Yunker, and J. L. Silverberg. Cyberphysical risks of hacked internet-connected vehicles. *Physical Review E*, 100:012316, 2019.
- L.-Z. Wang, Y.-Z. Chen, W.-X. Wang, and Y.-C. Lai. Physical controllability of complex networks. *Scientific Reports*, 7:40198, 2017.
- X. F. Wang and G. Chen. Synchronization in scale-free dynamical networks: Robustness and fragility. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 49(1):54–62, 2002a.
- X. F. Wang and G. Chen. Pinning control of scale-free complex networks. *Physica A*, 310:521–531, 2002b.

- X. F. Wang and G. Chen. Complex networks: Small-world, scale-free and beyond. *IEEE Circuits and Systems Magazine*, 3(1):6–20, 2003.
- D. J. Watts and S. H. Strogatz. Collective dynamics of ‘small-world’ networks. *Nature*, 393:440–442, 1998.
- A. J. Whalen, S. N. Brennan, T. D. Sauer, and S. J. Schiff. Observability and controllability of nonlinear networks: The role of symmetry. *Physical Review X*, 5: 011005, 2015.
- N. Wiener. *Cybernetics or Control and Communication in the Animal and the Machine*. MIT press, 2019.
- J. L. Willems. Structural controllability and observability. *Systems and Control Letters*, 8:5–12, 1986.
- M. Wolfrum. The Turing Bifurcation in Network Systems: Collective Patterns and Single Differentiated Nodes. *Physica D: Nonlinear Phenomena*, 241(16):1351–1357, 2012.
- L. Wu, Y. Shen, M. Li, and F. X. Wu. Network Output Controllability-Based Method for Drug Target Identification. *IEEE Transactions on Nanobioscience*, 14(2):184–191, 2015.
- N. Xue and A. Chakraborty. Control Inversion: A Clustering-Based Method for Distributed Wide-Area Control of Power Systems. *IEEE Transactions on Control of Network Systems*, 6(3):937–949, 2018.
- T. Yamada and D. G. Luenberger. Generic properties of column structured matrices. *Linear Algebra and its Applications*, 65:189–206, 1985.
- G. Yan, J. Ren, Y.-C. Lai, C.-H. Lai, and B. Li. Controlling complex networks: How much energy is needed? *Physical Review Letters*, 108(21):218703, 2012.
- G. Yan, G. Tsekenis, B. Barzel, J.-j. Slotine, Y.-y. Liu, and A.-L. Barabási. Spectrum of controlling and observing complex networks. *Nature Physics*, 11(9):779–786, 2015.
- Y. Yang, J. Wang, and A. E. Motter. Network observability transitions. *Physical Review Letters*, 109:258701, 2012.
- Y. Yang, T. Nishikawa, and A. E. Motter. Small vulnerable sets determine large network cascades in power grids. *Science*, 358(6365), 2017.
- Z. Yuan, C. Zhao, Z. Di, W.-X. Wang, and Y.-C. Lai. Exact controllability of complex networks. *Nature Communications*, 4:2447, 2013.
- J. Zabczyk. *Mathematical Control Theory: An Introduction*. Birkhäuser Boston, 2nd edition, 1995.
- J. G. T. Zañudo, G. Yang, R. Albert, and H. Levine. Structure-based control of complex networks with nonlinear dynamics. *PNAS*, 114(28):7234–7239, 2017.

- 
- S. Zhang and V. Vittal. Design of wide-area power system damping controllers resilient to communication failures. *IEEE Transactions on Power Systems*, 28(4):4292–4300, 2013.
- W. Zhang, W. Pei, and T. Guo. An efficient method of robustness analysis for power grid under cascading failure. *Safety Science*, 64:121–126, 2014.
- S. Zhao and F. Pasqualetti. Networks with diagonal controllability Gramian: Analysis, graphical conditions, and design algorithms. *Automatica*, 102:10–18, 2019.
- A. Zhirabok and A. Shumsky. An approach to the analysis of observability and controllability in nonlinear systems via linear methods. *International Journal of Applied Mathematics and Computer Science*, 22(3):507–522, 2012.
- R. D. Zimmerman, C. E. Murillo-Sanchez, and R. J. Thomas. MATPOWER: Steady-State Operations, Planning, and Analysis Tools for Power Systems Research and Education. *IEEE Transactions on Power Systems*, 26(1):12–19, 2011.

# Appendix A

## Sensor Placement Algorithm

This appendix describes a greedy algorithm for optimal sensor placement in dynamical networks. Section A.1 is based on the work of Summers et al. (2016) and describes the use of greedy algorithm, applied in Section 3.6.1, to minimize to find the optimal set of sensor nodes that minimizes a given coefficient of observability. Section A.2, applied in Section 4.4.2, uses the same greedy algorithm to solve a slightly different optimization problem which finds the set of sensor nodes that minimizes the functional observer order with respect to a given set of target nodes.

### A.1 Minimization of Coefficient of Observability

Summers et al. (2016) formulated the problem of sensor placement as a *set function optimization problem*. Consider a finite set  $\mathcal{X} = \{x_1, \dots, x_n\}$ , where  $x_i$  is the  $i$ -th state variable of the state vector  $\mathbf{x} = [x_1 \dots x_n]^\top$ . Let a *set function*  $J : 2^{\mathcal{X}} \mapsto \mathbb{R}^1$  assign a real number to each subset of  $\mathcal{X}$ . In the context of dynamical systems,  $\mathcal{X}$  represents the set of state variables which are potential locations for sensor placement,  $\mathcal{S} = \{y_1, \dots, y_q\} \subseteq \mathcal{X}$  represents the set of output variables, and  $J(\mathcal{S})$  could be a coefficient of observability that quantifies the degree of observability conveyed by a particular set  $\mathcal{S}$  (as in Section 3.2). In the context of networks, the state and output variables can represent nodes and sensors, respectively.

The set function optimization problem can be defined as

$$\max_{\mathcal{S} \subseteq \mathcal{X}, |\mathcal{S}|=q} J(\mathcal{S}) \tag{A.1}$$

where  $|\mathcal{S}|$  denotes the cardinality of a set. The goal is to find the set of  $q$ -elements of  $\mathcal{X}$  that maximizes  $J$ . This combinatorial problem, however, grows exponentially with the cardinality of  $\mathcal{S}$ . Nevertheless, Summers et al. (2016) demonstrated that some coefficients of observability based on the Gramian of observability are submodular functions. Hence, if  $J$  is a submodular function, (A.1) can be solved using a greedy algorithm, which is guaranteed to have a performance within a well-known bound (Summers et al., 2016, Theorem 3). Summers *et al.* showed numerically that, given a submodular set function  $J$ , the greedy optimization performs very close to the global optimal solution. In fact, a good performance was achieved even when using cost functions that are not submodular, such as the minimum eigenvalue of the Gramian of observability. The greedy algorithm is implemented as follows.

---

**Algorithm 3** Greedy algorithm.

---

For  $k = 1, \dots, q..$

1. Let  $\mathcal{S}^{(0)} \leftarrow \emptyset$  be an empty starting set, where the superscript represents an iteration counter.
2. Compute the gain  $\Delta(x_i | \mathcal{S}^{(k)}) = J(\mathcal{S}^{(k)} \cup \{x_i\}) - J(\mathcal{S}^{(k)})$ , for all elements  $x_i \in \mathcal{X} \setminus \mathcal{S}^{(k)}$ .
3. Add the element with the highest gain

$$\mathcal{S}^{(k+1)} \leftarrow \mathcal{S}^{(k)} \cup \left\{ \arg \max_{x_i} \Delta(x_i | \mathcal{S}^{(k)}) \mid x_i \in \mathcal{X} \setminus \mathcal{S}^{(k)} \right\}. \quad (\text{A.2})$$


---

## A.2 Minimization of Functional Observer Order

Given a functionally observable triple  $(A, C, F)$ , one can be interested in how to optimally place additional sensor nodes in a network such that the functional observer order  $r_0$  is minimized. This is a bilevel optimization problem

$$\min_{\mathcal{S} \subseteq \mathcal{C} \setminus \mathcal{T}, |\mathcal{S}|=q} J(\mathcal{S}), \quad (\text{A.3})$$

where  $\mathcal{C}$  is a set of candidate nodes for sensor placement and  $J(\mathcal{S})$  is a cost function that returns the minimum order  $r_0$  of a functional observer. However, finding  $r_0$  is “embedded” in a lower-level optimization task which requires, for instance, use of Algorithm 2. This is a hard to solve problem but, for illustration purposes in Fig. 4.4f (Section 4.4.2), we implement the same (non-scalable) Algorithm 3 where the cost function  $J(\mathcal{S})$  is the order of  $F_0$  returned by Algorithm 2 for this set  $\mathcal{S}$ .

# Appendix B

## State Observer Design

This appendix reviews the design of full-order state observers, reduced-order state observers and functional observers.

### B.1 Luenberger Observer

The classic linear state observer was introduced by [Luenberger \(1966\)](#). The following concepts and theorems were further discussed and proven in many textbooks such as ([Hieu and Tyrone, 2012](#); [Korovin and Fomichev, 2009](#); [O'Reilly, 1983](#)).

Consider the linear dynamical system (2.1) and assume that not all state variables  $\mathbf{x}$  are measured directly in  $\mathbf{y}$ . Define an observer with the following dynamics ([Luenberger, 1966](#)):

$$\dot{\hat{\mathbf{x}}} = A\hat{\mathbf{x}} + B\mathbf{u} + L(\mathbf{y} - C\hat{\mathbf{x}}), \quad (\text{B.1})$$

where  $\hat{\mathbf{x}} \in \mathbb{R}^n$  is an approximation of  $\mathbf{x}$ , i.e., an estimation by the state observer, and  $L \in \mathbb{R}^{n \times q}$  is the observer gain matrix. This state observer is classified as a *full-order* state observer. In the RHS of (B.1), the first two terms model the system dynamics while the third term is proportional to the difference between the measured and estimated output, i.e., the mismatch between the real system and the observer dynamics.

Let the observer estimation error  $\mathbf{e} := \mathbf{x} - \hat{\mathbf{x}}$  dynamics be given by

$$\begin{aligned} \dot{\mathbf{e}} &= \dot{\mathbf{x}} - \dot{\hat{\mathbf{x}}} \\ &= (A - LC)\mathbf{x} - (A - LC)\hat{\mathbf{x}} \\ &= (A - LC)\mathbf{e}. \end{aligned} \quad (\text{B.2})$$

Thus, the estimation error converges to zero, i.e.,  $\hat{\mathbf{x}}$  approaches  $\mathbf{x}$  as  $t \rightarrow \infty$ , if  $(A - LC)$  is Hurwitz. The eigenvalue assignment of  $(A - LC)$  determines the observer performance: the higher the negative real component of its eigenvalues, the faster  $\hat{\mathbf{x}}$  approaches  $\mathbf{x}$ ; however this also results in higher entries of  $L$  amplifying the measurement noise levels in  $\mathbf{y}$ .

The following theorem states that observability is a necessary and sufficient condition for the existence of a full-order state observer.

**Theorem B.1.** (*O'Reilly, 1983, Theorem 1.16*) *Corresponding to the real matrices  $A$  and  $C$ , there is a real matrix  $L$  such that the set of eigenvalues of  $(A - LC)$  can be arbitrarily assigned (subject to complex eigenvalues occurring in conjugate pairs) if and only if the pair  $(A, C)$  is observable. ■*

There is redundancy in the design of a full-order state observer. A full-order state observer has the same order  $n$  that the original linear dynamical system it aims to reconstruct. However, since the output  $\mathbf{y}$  contains  $q$  linear combinations of  $\mathbf{x}$ , the remaining  $n - q$  state variables can be reconstructed from an observer of order  $n - q$ —entitled *reduced-order* state observer.

The reduced-order observer is based on the following partitioned form<sup>1</sup> of (2.1):

$$\begin{aligned} \begin{bmatrix} \dot{\mathbf{x}}_q \\ \dot{\mathbf{x}}_u \end{bmatrix} &= \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{x}_q \\ \mathbf{x}_u \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \cdot \mathbf{u} \\ \mathbf{y} &= \begin{bmatrix} I_q & 0 \end{bmatrix} \cdot \begin{bmatrix} \mathbf{x}_q \\ \mathbf{x}_u \end{bmatrix} = \mathbf{x}_q, \end{aligned} \quad (\text{B.3})$$

where  $\mathbf{x}_q \in \mathbb{R}^q$  and  $\mathbf{x}_u \in \mathbb{R}^{n-q}$  are, respectively, the observed (measured) and unobserved (unmeasured) states from the output  $\mathbf{y}$ , and  $I_q$  is an identity matrix of order  $q$ . In this case, the goal is to estimate only  $\mathbf{x}_u$ . Thus, analogous to (B.1), define the reduced-order state observer with the following dynamics:

$$\dot{\hat{\mathbf{x}}}_u = A_{21}\mathbf{x}_p + A_{22}\hat{\mathbf{x}}_u + B_2\mathbf{u} + E(\bar{\mathbf{y}} - A_{12}\hat{\mathbf{x}}_u), \quad (\text{B.4})$$

where  $\bar{\mathbf{y}} := A_{12}\mathbf{x}_u = \dot{\mathbf{x}}_q - A_{11}\mathbf{x}_q - B_1\mathbf{u} \in \mathbb{R}^q$  and  $E \in \mathbb{R}^{(n-q) \times n}$  is a reduced-order observer gain matrix. Nevertheless,  $\bar{\mathbf{y}}$  requires differentiation of  $\mathbf{x}_q = \mathbf{y}$ , which amplifies noise in  $\mathbf{y}$ , compromising the observer performance. To circumvent this issue,

<sup>1</sup>This can be achieved through a linear transformation  $\mathbf{x} = P \cdot [\mathbf{x}_q^\top \ \mathbf{x}_u^\top]^\top$ , where  $P = [C^\dagger \ C^\perp] \in \mathbb{R}^{n \times n}$ ,  $C^\dagger \in \mathbb{R}^{n \times p}$  denotes the Moore-Penrose inverse (pseudoinverse) of  $C$ , and  $C^\perp \in \mathbb{R}^{n \times (n-p)}$  denotes an orthogonal basis for the null-space of  $C$  ( $CC^\perp = 0$ ).



a reduced-order observer can be defined as follows:

$$\begin{cases} \dot{\mathbf{w}} = N\mathbf{w} + J\mathbf{y} + H\mathbf{u}, \\ \hat{\mathbf{x}}_u = \mathbf{w} + E\mathbf{y}, \end{cases} \quad (\text{B.5})$$

$\mathbf{w} \in \mathbb{R}^{(n-q)}$ , and  $(E, N, J, H)$  are design matrices of appropriate dimension. Thus, the estimation error  $\mathbf{e}_u := \mathbf{x}_u - \hat{\mathbf{x}}_u$  dynamics is given by

$$\begin{aligned} \dot{\mathbf{e}} &= \dot{\mathbf{x}}_u - \dot{\hat{\mathbf{x}}}_u \\ &= A_{21}\mathbf{x}_q + A_{22}\mathbf{x}_u + B_2\mathbf{u} - (N\mathbf{w} + J\mathbf{y} + H\mathbf{u} + E\dot{\mathbf{y}}) \\ &= (A_{22} - EA_{12})\mathbf{x}_u - N\hat{\mathbf{x}}_u + (A_{21} + NE - J - EA_{11})\mathbf{y} + (B_2 - H - EB_1)\mathbf{u} \\ &= (A_{22} - EA_{12})\mathbf{e}_u, \end{aligned} \quad (\text{B.6})$$

where  $N = A_{22} - EA_{12}$ ,  $J = A_{21} - EA_{11} + NE$  and  $H = B_2 - EB_1$  were chosen properly to provide cancellations. As in (B.2), the reduced-order state observer error converges asymptotically to zero if  $(A_{22} - EA_{12})$  is Hurwitz. Analogous to Theorem B.1, the reduced-order observer exists if and only if the pair  $(A_{22}, A_{12})$  is observable.

**Lemma B.1.** *If the pair  $(A, C)$  is observable, then the pair  $(A_{22}, A_{12})$  is also observable.*

**Remark B.1.** Since the output signal  $\mathbf{y}$  is present in the dynamics of the observer  $\dot{\mathbf{w}}$  as well as in the state estimate  $\hat{\mathbf{x}}_u$  in (B.5), the reduced-order observer is more susceptible to measurement errors in  $\mathbf{y}$  than in the full-order observer. This is a consequence of the lack of redundancy in the reduced-order observer.

**Remark B.2.** In practical applications, an order reduction of  $q$  is only significant in multiple output systems where  $q \gg 1$ .

Algorithm 4 summarizes the steps to design a reduced-order observer of form (B.5). Further work expands the scope of the Luenberger observers to applications with unknown inputs, time-delay systems, Lipschitz nonlinear systems and others. The reader is referred to (Hieu and Tyrone, 2012) for further details.

## B.2 Functional Observer

Consider the linear dynamical system (2.1)–(4.1). In this section, the goal is to design a state observer—known as a *functional observer*—that is capable of reconstructing a linear functional of the unknown (unmeasured) state vector  $\mathbf{x}$ . We assume that a triple  $(A, C, F)$  is functional observable and satisfies conditions (4.4)–(4.5) (thus,  $F = F_0$ ).

---

**Algorithm 4** Design of a Luenberger reduced-order observer.

---

1. Compute sub-matrices  $A_{11}, A_{12}, A_{21}, A_{22}, F_1, F_2$ , where

$$P^{-1}AP = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad FP = \begin{bmatrix} F_1 & F_2 \end{bmatrix}, \quad P = \begin{bmatrix} C^\dagger & C^\perp \end{bmatrix}. \quad (\text{B.7})$$

2. Check the *detectability*<sup>2</sup> of the pair  $(A_{22}, A_{12})$ , i.e. if  $\text{rank}[(\lambda I_n - A_{22}) \ A_{12}] = n$  for every  $\text{Re}(\lambda_i) \geq 0$ , where  $\lambda_i$ , for  $i = 1, \dots, n$ , is an eigenvalue of  $A$ . If not, stop as  $N$  is not stable.
  3. Determine  $E$  such that  $N = A_{22} - EA_{12}$  is stable.
  4. Compute  $J = A_{21} - EA_{11} + NE$  and  $H = B_2 - EB_1$ .
- 

Define a functional observer with the following dynamics ([Hieu and Tyrone, 2012](#); [Korovin and Fomichev, 2009](#)):

$$\begin{cases} \dot{\mathbf{w}} = N\mathbf{w} + J\mathbf{y} + H\mathbf{u} \\ \hat{\mathbf{z}} = D\mathbf{w} + E\mathbf{y}, \end{cases} \quad (\text{B.8})$$

where  $\mathbf{w} \in \mathbb{R}^{\bar{r}}$ ,  $\hat{\mathbf{z}} \in \mathbb{R}^r$  is an estimate of  $\mathbf{z}$ , and  $(D, E, N, H)$  are matrices of appropriate dimensions to be determined such that  $\hat{\mathbf{z}}$  converges asymptotically to  $\mathbf{z}$ . Since conditions (4.4)–(4.5) holds for the triple  $(A, C, F)$ , the functional observer is of  $r$ -th order.

Note that since  $\hat{\mathbf{z}}$  estimates  $F\mathbf{x}$ , then  $\mathbf{w}$  estimates some other linear combination  $T\mathbf{x}$ , where  $T \in \mathbb{R}^{\bar{r} \times n}$ . Thus, to guarantee that  $\hat{\mathbf{z}} \rightarrow F\mathbf{x}$  as  $t \rightarrow \infty$ , then  $\mathbf{w} \rightarrow T\mathbf{x}$  as  $t \rightarrow \infty$ . The conditions for the observer asymptotic stability are determined by the following theorem.

**Theorem B.2.** ([Hieu and Tyrone, 2012, Theorem 3.1](#)) *The estimate  $\hat{\mathbf{z}}(t)$  will converge asymptotically to  $F\mathbf{x}(t)$  for any initial condition  $\mathbf{w}(t_0)$  and any known input  $\mathbf{u}(t)$  if and only if the following conditions hold.*

$$N \text{ is Hurwitz}, \quad (\text{B.9})$$

$$NT + JC - TA = 0, \quad (\text{B.10})$$

$$H - TB = 0, \quad (\text{B.11})$$

$$F - DT - EC = 0. \quad (\text{B.12})$$

---

<sup>2</sup>Detectability is a weaker notion of observability, where a system is detectable if all the unobservable (unmeasured) states are stable. The stated condition is straightforwardly derived from [Theorem 3.1.4](#).

*Proof.* Let the estimation error  $\mathbf{e}_w \equiv \mathbf{w} - T\mathbf{x}$  dynamics be given by

$$\begin{aligned}\dot{\mathbf{e}}_w &= \dot{\mathbf{w}} - T\dot{\mathbf{x}} \\ &= N\mathbf{w} + JC\mathbf{x} + H\mathbf{u} - TA\mathbf{x} - TB\mathbf{u} + (+NT\mathbf{x} - NT\mathbf{x}) \\ &= N\mathbf{e}_w + (NT + JC - TA)\mathbf{x} + (H - TB)\mathbf{u} \\ &= N\mathbf{e}_w,\end{aligned}\tag{B.13}$$

where conditions (B.10) and (B.11) were applied. Clearly,  $\mathbf{e}_w$  is asymptotically stable if condition (B.9) holds. From this result, the estimation error  $\mathbf{e}_z := \hat{\mathbf{z}} - F\mathbf{x}$  asymptotic behavior can be proven as follows:

$$\begin{aligned}\dot{\mathbf{e}}_z &= \dot{\hat{\mathbf{z}}} - F\dot{\mathbf{x}} \\ &= D\mathbf{w} + EC - DT\mathbf{x} - EC\mathbf{x} \\ &= D\mathbf{e}_w,\end{aligned}\tag{B.14}$$

where condition (B.12) was applied. Since  $\mathbf{e}_w$  is asymptotically stable, so is  $\mathbf{e}_z$ .  $\square$

---

**Algorithm 5** Design of a multi-functional observer.

---

1. Compute sub-matrices  $A_{11}, A_{12}, A_{21}, A_{22}, F_1, F_2$ , where

$$P^{-1}AP = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad FP = \begin{bmatrix} F_1 & F_2 \end{bmatrix}, \quad P = \begin{bmatrix} C^\dagger & C^\perp \end{bmatrix}.\tag{B.15}$$

2. Check if condition

$$\text{rank} \begin{bmatrix} F_2 A_{22} \\ A_{12} \\ F_2 \end{bmatrix} = \text{rank} \begin{bmatrix} A_{12} \\ F_2 \end{bmatrix}\tag{B.16}$$

- is satisfied. If not, stop as an  $r$ -th order observer does not exist.
  3. Compute  $N_1 = (\Phi\Omega^\dagger A_{12} + F_2 A_{22})F_2^\dagger$  and  $N_2 = (\Omega\Omega^\dagger - I_q)A_{12}F_2^\dagger$ , where  $\Omega = A_{12}F_2^\perp$  and  $\Phi = -F_2 A_{22} F_2^\perp$ .
  4. Check the detectability of the pair  $(N_2, N_1)$ , i.e. if  $\text{rank}[(\lambda I_n - N_2) N_1] = n$  for every  $\text{Re}(\lambda_i) \geq 0$ , where  $\lambda_i$ , for  $i = 1, \dots, n$ , is an eigenvalue of  $A$ . If not, stop as  $N$  is not stable.
  5. Determine  $Z$  such that  $N = N_1 - ZN_2$  is stable.
  6. Compute  $L_1 = \Phi\Omega^\dagger + Z(I_q - \Omega\Omega^\dagger)$ ,  $L_2 = F_2$ ,  $D = I_r$ ,  $J = L_1 A_{11} + L_2 A_{21} - NL_1$ , and  $E = F_1 - DL_1$ .
  7. Compute  $H = LB$ , where  $L = [L_1 \ L_2]$ .
- 

Designing the parameters of  $(D, E, J, N, T)$  so that it guarantees a functional observer of minimum order is a nontrivial problem (Darouach, 2000; Fernando et al., 2010b), specially when multi-functional observers ( $r > 1$ ) are concerned. As presented

by (Hieu and Tyrone, 2012, Section 3.5.1), Algorithm 5 provides a step-by-step design procedure to systematically derive the parameters of a functional observer assuming that  $(A, C, F)$  satisfies conditions (4.4)–(4.5). If such conditions are not satisfied but  $(A, C, F)$  is functional observable, one can determine an augmented matrix  $F_0 \in \mathbb{R}^{r_0 \times n}$ , where  $\text{row}(F_0) \supseteq \text{row}(F)$ , such that  $(A, C, F_0)$  satisfies (4.4)–(4.5) and a functional observer of order  $r_0 > r$  can be designed. This is one of the main problems discussed in Chapter 4.

### B.3 Functional Observer for Nonlinear Systems

In the study of the epidemiological model (4.19) in Section 4.5.2, we illustrate how our contributions can lead to the design of a functional observer for nonlinear systems by: (i) determining, via Algorithm 1, the minimum sensor placement  $\mathcal{S}$  for the functional observability of a given set of target states  $\mathcal{T}$  (i.e., given  $\tilde{\mathcal{G}}$  and  $\mathcal{T}$ , finding  $\mathcal{S}$ ); and (ii) determining, via Algorithm 2, the structure of the minimum order functional observer (i.e., given  $\tilde{\mathcal{G}}$ ,  $\mathcal{S}$  and  $\mathcal{T}$ , finding the structure matrix  $F_0$ ). However, in order to actually complete the design of the functional observer, we need to determine the parameters of the functional observer system. To that end, we apply the results presented in (Trinh et al., 2006), and also reported in (Hieu and Tyrone, 2012, Section 6.2), for the design of stable functional observers for nonlinear systems as detailed below.

Consider the following class of nonlinear systems studied in (Trinh et al., 2006):

$$\begin{cases} \dot{\mathbf{x}} = A\mathbf{x} + \mathbf{f}(\mathbf{x}), \\ \mathbf{y} = C\mathbf{x}, \\ \mathbf{z}_0 = F_0\mathbf{x}, \end{cases} \quad (\text{B.17})$$

where  $\mathbf{f}(\mathbf{x}) : \mathbb{R}^n \mapsto \mathbb{R}^n$  is a nonlinear function not required to be Lipschitz. The epidemiological model (4.19) can be described as (B.17) by defining the state vector

$$\mathbf{x} = [S_1 \ \dots \ S_N \ I_1 \ \dots \ I_N \ R_1 \ \dots \ R_N \ D_1 \ \dots \ D_N]^\top \in \mathbb{R}^n, \quad (\text{B.18})$$

where  $n = 4N$ , the nonlinear function

$$\mathbf{f}(\mathbf{x}) = [-\beta_1 S_1 I_1 / P_1 \ \dots \ -\beta_N S_N I_N / P_N \ \beta_1 S_1 I_1 / P_1 \ \dots \ \beta_N S_N I_N / P_N \ \mathbf{0}^\top]^\top, \quad (\text{B.19})$$

where  $\mathbf{0} \in \{0\}^{1 \times 2N}$ , and  $A \in \mathbb{R}^{n \times n}$  as a matrix defined by the linear functions in (B.17). As described above, matrices  $C$  and  $F_0$  are determined by Algorithms 1 and 2, respectively.

Given  $(A, C, F_0)$  and  $\mathbf{f}(\mathbf{x})$ , a reduced-order functional observer can be designed to estimate the partial state vector  $\mathbf{z}_0$  and thereby  $\mathbf{z} = F\mathbf{x}$  (since  $F$  defines the first  $r$  rows of  $F_0$ , see Algorithm 2). Consider the following structure for the functional observer:

$$\begin{cases} \dot{\mathbf{w}} = N\mathbf{w} + J\mathbf{y} + L\mathbf{f}_1(\mathbf{y}, \mathbf{z}_0), \\ \hat{\mathbf{z}} = D\mathbf{w} + E\mathbf{y}, \end{cases} \quad (\text{B.20})$$

where  $(N, J, L, D, E)$  and  $\mathbf{f}_1(\mathbf{x})$  are to be determined such that  $\hat{\mathbf{z}}(t)$  converges asymptotically to  $\mathbf{z}_0(t)$ . This can be achieved by satisfying the conditions stated in (Hieu and Tyrone, 2012, Theorem 6.1) following Algorithm 6.

Note that measures  $\mathbf{y}$  are not an argument of  $\mathbf{f}_1(\mathbf{z}_0)$  since we defined in the main manuscript that measures are taken only on the number of dead cases  $D_i$  (if a given group  $i$  is chosen as a “sensor” city), and  $\mathbf{f}(\mathbf{x})$  is not a function of  $D_i$ . We note that, in the design of the functional observer (B.20), the nonlinear function  $\mathbf{f}_2(\mathbf{x})$  is treated as an unknown input, and that  $W$  has a fixed structure (with full-column rank), despite not being a uniquely defined matrix.

---

**Algorithm 6** Design of nonlinear functional observer for a epidemiological model.

---

1. Decompose  $\mathbf{f}(\mathbf{x}) = \mathbf{f}_1(\mathbf{y}, \mathbf{z}_0) + W\mathbf{f}_2(\mathbf{x}) = \mathbf{f}_1(\mathbf{z}_0) + W\mathbf{f}_2(\mathbf{x})$ , where

$$[\mathbf{f}_1(\mathbf{z}_0)]_i = \begin{cases} -\beta_i \frac{z_j z_k}{P_i}, & \text{if } [F_0]_{ji} = 1, \text{ for some } j, \\ & \text{and } [F_0]_{k(i+N)} = 1, \text{ for some } k, \text{ and } i \leq N, \\ +\beta_i \frac{z_j z_k}{P_i}, & \text{if } [F_0]_{ji} = 1, \text{ for some } j, \\ & \text{and } [F_0]_{k(i+N)} = 1, \text{ for some } k, \text{ and } N < i \leq 2N, \\ 0, & \text{otherwise,} \end{cases} \quad (\text{B.21})$$

$$[W\mathbf{f}_2(\mathbf{x})]_i = \begin{cases} -\beta_i \frac{x_i x_{(i+N)}}{P_i}, & \text{if } [F_0]_{ji} = 0, \text{ for all } j, \\ & \text{or } [F_0]_{k(i+N)} = 0, \text{ for all } k, \text{ and } i \leq N, \\ +\beta_i \frac{x_i x_{(i+N)}}{P_i}, & \text{if } [F_0]_{ji} = 0, \text{ for all } j, \\ & \text{or } [F_0]_{k(i+N)} = 0, \text{ for all } k, \text{ and } N < i \leq 2N, \\ 0, & \text{otherwise,} \end{cases} \quad (\text{B.22})$$

for all  $i = 1, \dots, n$ .

2. The nonlinear function  $\mathbf{f}_1(\mathbf{z}_0)$  is Lipschitz with a constant Lipschitz constant  $\kappa$ , i.e.,

$$\|\mathbf{f}_1(\mathbf{z}_0) - \mathbf{f}_1(\bar{\mathbf{z}}_0)\| \leq \kappa \|\mathbf{z}_0 - \bar{\mathbf{z}}_0\|, \quad (\text{B.23})$$

where  $\kappa = \max_i(\beta_i)$ .

3. Consider the linear transformation in (B.8),  $PW = [W_1^\top \ W_2^\top]^\top$ , and  $P^{-1} = [P_1^\top \ P_2^\top]^\top$ . Compute  $N_1 = (\Phi\Omega^\dagger A_{12} + F_2 A_{22})F_2^\dagger$ ,  $N_2 = (\Omega\Omega^\dagger - I_q)A_{12}F_2^\dagger$ ,  $\bar{L}_1 = \Phi\Omega^\dagger P_1 + F_2 * P_2$  and  $\bar{L}_2 = (\Omega\Omega^\dagger - I_q)P_1$ , where  $\Omega = [A_{12}F_2^\perp \ W_1]$  and  $\Phi = -[F_2 A_{22} F_2^\perp \ F_2 W_2]$ .
4. Find matrices  $Q = Q^\top \in \mathbb{R}^{r_0 \times r_0}$  and  $G \in \mathbb{R}^{r_0 \times q}$ , and the positive scalars  $\beta_1$  and  $\beta_2$ , such that the following linear matrix inequality (LMI) holds (Hieu and Tyrone, 2012, Theorem 6.2):

$$\begin{bmatrix} \Delta & Q\bar{L}_1 & G\bar{L}_2 \\ \bar{L}_1^\top Q & -\beta_1 I_n & \mathbf{0} \\ \bar{L}_2^\top G^\top & \mathbf{0} & -\beta_2 I_n \end{bmatrix} < 0, \quad (\text{B.24})$$

where  $\Delta = QN_1 + N_1^\top - GN_2 - N_2^\top G^\top + \kappa^2(\beta_1 + \beta_2)I_{r_0}$ .

5. Compute the auxiliary matrices  $Z = Q^{-1}G$ ,  $T_1 = \Phi\Omega^\dagger + Z(I_q - \Omega\Omega^\dagger)$  and  $T_2 = F_2$ .
6. Compute the functional observer matrices  $N = N_1 - ZN_2$ ,  $J = T_1 A_{11} + T_2 A_{21} - NT_1$ ,  $L = \bar{L}_1 - Z\bar{L}_2$ ,  $D = I_{r_0}$ , and  $E = F_1 - DT_1$ . This satisfies condition 2 of (Hieu and Tyrone, 2012, Theorem 6.1).
-

# Appendix C

## Proof of Structural Functional Observability

*Proof of Theorem 4.3.* This proof follows from the four following lemmas. Lemma C.1 and C.2 state the necessity of conditions 1 and 2, respectively. Lemmas C.4 and C.5, together, state the sufficiency of conditions 1 and 2. The proof of Lemma C.4 borrows basic idea from (Sundaram, 2012, Theorem C.3) and extends it to functional observability.

We remind the reader that an equivalent condition to (4.2) for functional observability is (Jennings et al., 2011):

$$\text{rank} \begin{bmatrix} A - \lambda I \\ C \\ F \end{bmatrix} = \text{rank} \begin{bmatrix} A - \lambda I \\ C \end{bmatrix}, \quad (\text{C.1})$$

for all  $\lambda \in \mathbb{C}$ . Since the equality holds trivially for  $\lambda \notin \text{spec}(A)$ , we only need to care about those  $\lambda \in \text{spec}(A)$  when trying to establish the equality (C.1) for given triple  $(A, C, F)$ .  $\square$

**Lemma C.1** (Necessity of condition 1). *If there is at least one state  $x_i \in \mathcal{T}$  that does not have a path to some output  $y_i \in \mathcal{S}$ , then for any generic choice of free parameters in the system matrices  $(A, C, F)$ , there is at least one  $\lambda \in \mathbb{C}$  such that  $\text{rank} [A^\top - \lambda I \quad C^\top \quad F^\top]^\top > \text{rank} [A^\top - \lambda I \quad C^\top]^\top$ .*

*Proof.* Suppose that some nodes in  $\mathcal{X}$  do not have a path to some output in  $\mathcal{S}$ . Let  $\mathcal{X}_1$  be the set of nodes that have a path to some output, and  $\mathcal{X}_2 = \mathcal{X} \setminus \mathcal{X}_1$  denote all state nodes that do not have a path to any output. Let  $|\mathcal{X}_1| = k$  and  $|\mathcal{X}_2| = n - k$ .

After applying a permutation of coordinates such that the nodes in  $\mathcal{X}_1$  come first, then matrices  $(A, C, F)$  have the form:

$$A = \begin{bmatrix} A_{11} & \mathbf{0} \\ A_{21} & A_{22} \end{bmatrix}, C = [C_1 \quad \mathbf{0}], F = [F_1 \quad F_2]. \quad (\text{C.2})$$

where  $A_{11} \in \mathbb{R}^{k \times k}$ ,  $A_{21} \in \mathbb{R}^{(n-k) \times k}$ ,  $A_{22} \in \mathbb{R}^{(n-k) \times (n-k)}$ ,  $C_1 \in \mathbb{R}^{q \times k}$ ,  $F_1 \in \mathbb{R}^{r \times k}$ ,  $F_2 \in \mathbb{R}^{r \times (n-k)}$ , and  $\mathbf{0}$  is a vector or matrix of appropriate dimension with zero entries. Note that there are no edges from a node in  $\mathcal{X}_2$  to  $\mathcal{X}_1 \cup \mathcal{S}$ . Thus, we have the following matrix pencil:

$$\begin{bmatrix} A - \lambda I \\ C \\ F \end{bmatrix} = \begin{bmatrix} A_{11} - \lambda I & \mathbf{0} \\ A_{21} & A_{22} - \lambda I \\ C_1 & \mathbf{0} \\ F_1 & F_2 \end{bmatrix}. \quad (\text{C.3})$$

Assume (C.1) is satisfied. From (C.3), we have  $\text{row}(F_2) \subseteq \text{row}(A_{22} - \lambda I)$  for all  $\lambda \in \text{spec}(A_{22})$ , which implies

$$\text{row}(F_2) \subseteq \bigcap_{\lambda \in \text{spec}(A_{22})} U_\lambda^\perp = \left( \bigoplus_{\lambda \in \text{spec}(A_{22})} U_\lambda \right)^\perp = \{\mathbf{0}\}, \quad (\text{C.4})$$

where  $\oplus$  is the direct sum operator,  $U_\lambda$  is the left eigen-space of  $A_{22}$  corresponding to eigenvalue  $\lambda$  and the second equality comes from the fact that, for a generic numerical realization,  $A$  has a complete set of eigenvectors. This shows that  $F_2$  contain only all-zero rows, which contradicts the assumption that  $\mathcal{T} \cap \mathcal{X}_2 \neq \emptyset$ . Therefore, we have  $\text{rank} [A^\top - \lambda I \quad C^\top \quad F^\top]^\top > \text{rank} [A^\top - \lambda I \quad C^\top]^\top$ .  $\square$

**Lemma C.2** (Necessity of condition 2).  $\text{rank}[A^\top \quad C^\top \quad F^\top] > \text{rank}[A^\top \quad C^\top]^\top$  if  $\mathcal{T} \cap \mathcal{D} \neq \emptyset$ .

*Proof.* Pick  $x_i \in \mathcal{T} \cap \mathcal{D}$  and thus there is a minimal dilation set  $\mathcal{D}' \subseteq \mathcal{D}$  that contains  $x_i$ . Let  $\mathcal{X}_2 = \mathcal{D}'$ , and  $\mathcal{X}_1 = \mathcal{X} \setminus \mathcal{X}_2$ , where  $|\mathcal{X}_1| = n - k$  and  $|\mathcal{X}_2| = k$ . After applying a permutation of coordinates such that the nodes in  $\mathcal{X}_1$  come first, then matrices  $(A, C)$  have the form:

$$A = [A_1 \quad A_2], C = [C_1 \quad C_2], F = [F_1 \quad F_2], \quad (\text{C.5})$$

where  $A_1 \in \mathbb{R}^{k \times n}$ ,  $A_2 \in \mathbb{R}^{n \times k}$ ,  $C_1 \in \mathbb{R}^{q \times (n-k)}$ ,  $C_2 \in \mathbb{R}^{q \times k}$ ,  $F_1 \in \mathbb{R}^{r \times (n-k)}$ , and  $F_2 \in \mathbb{R}^{r \times k}$ . Since  $\mathcal{D}'$  is a dilation set,  $[A_2^\top \quad C_2^\top]^\top$  has at most  $k - 1$  non-zeros rows due to Remark ???. In addition,  $F_2$  contains at least one non-zero row because  $x_i \in \mathcal{T}$ . Let us pick  $\phi \in \mathbb{R}^{1 \times k}$  which is the non-zero row of  $F_2$  that corresponds to the target state



$x_i$ . By assumption, each row of  $F$  has only one non-zero entry, thus row vector  $\phi$  only contains a single non-zero entry in the corresponding column of  $x_i$ . Now, we are ready to prove the Lemma by contradiction. Assume  $\text{rank}[A^\top \ C^\top \ F^\top]^\top = \text{rank}[A^\top \ C^\top]^\top$ . It gives  $\phi \in \text{row}([A_2^\top \ C_2^\top]^\top)$ , i.e. there exist a non-zero vector  $\mathbf{y} \in \mathbb{R}^{1 \times (n+q)}$  such that  $\phi = \mathbf{y}[A_2^\top \ C_2^\top]^\top$ . This implies  $\mathbf{y}[A_2^\top \ C_2^\top]_{\mathcal{D}' \setminus \{x_i\}}^\top = \mathbf{0}^{1 \times (k-1)}$  where  $[A_2^\top \ C_2^\top]_{\mathcal{D}' \setminus \{x_i\}}^\top$  is the matrix  $[A_2^\top \ C_2^\top]^\top$  removing the corresponding column of state  $x_i$ . It shows  $\text{rank}([A_2^\top \ C_2^\top]_{\mathcal{D}' \setminus \{x_i\}}^\top) < k - 1$ , i.e.  $\mathcal{D}' \setminus \{x_i\}$  is also a dilation set. This contradicts to the assumption that  $\mathcal{D}'$  is a minimal dilation set. Therefore, we conclude  $\text{rank}[A^\top \ C^\top \ F^\top]^\top > \text{rank}[A^\top \ C^\top]^\top$ .  $\square$

**Lemma C.3.** (*van der Woude, 1991, 1999*) *The generic rank of a matrix pencil*

$$P(\lambda) = \begin{bmatrix} A - \lambda I & B \\ C & D, \end{bmatrix} \quad (\text{C.6})$$

over all choices of free parameters in  $(A, B, C, D)$  and  $\lambda \in \mathbb{C}$ , is equal to  $n + l$ , where  $l$  is the largest number of disjoint paths from the input nodes  $u_i \in \mathcal{U}$  to the output nodes  $y_i \in \mathcal{S}$  in  $\mathcal{G}(A, B, C)$ .

*Proof.* See (*van der Woude, 1991, 1999*).  $\square$

**Lemma C.4** (Sufficiency of condition 1 for (C.1) for all  $\lambda \neq 0$ ). *If every state  $x_i \in \mathcal{T}$  has a path to some output  $y_i \in \mathcal{S}$ , then, for almost any generic choice of free parameters in  $(A, C, F)$ ,  $\text{rank}[A^\top - \lambda I \ C^\top \ F^\top]^\top = \text{rank}[A^\top - \lambda I \ C^\top]^\top$  for every  $\lambda \in \mathbb{C} \setminus \{0\}$ .*

*Proof.* Let matrix  $\bar{P}_i(\lambda)$  be formed by removing the  $i$ -th row of  $[A^\top - \lambda I \ C^\top \ F^\top]^\top$  and permuting the  $i$ -th column to the last column, i.e.

$$\bar{P}_i(\lambda) = \begin{bmatrix} A_i - \lambda I_{n-1} & \mathbf{b}_i \\ C_i & \mathbf{c}_i \\ F_i & \mathbf{f}_i \end{bmatrix}, \quad (\text{C.7})$$

where  $A_i$  is the matrix formed by removing the  $i$ -th row and  $i$ -th column of  $A$ ;  $C_i$  and  $F_i$  are formed by removing the corresponding  $i$ -th column of matrices  $C$  and  $F$ , respectively;  $\mathbf{b}_i$ ,  $\mathbf{c}_i$  and  $\mathbf{f}_i$  are the  $i$ -th column of  $A - \lambda I$ ,  $C$ , and  $F$ , respectively. In  $\mathcal{G}(A, C)$ , this corresponds to removing all incoming edges to the  $i$ -th state node  $x_i$ , and, by maintaining all outgoing edges from  $x_i$ , we thus can view  $x_i$  as an input node corresponding to the input vector  $\mathbf{b}_i$ . We further define  $P_i(\lambda)$  as the matrix formed by removing the  $i$ -th row from matrix  $[A^\top - \lambda I \ C^\top]^\top$ .

For any node  $x_i \in \mathcal{T} \cup \mathcal{S}$ , by assumption, there is a path from  $x_i \in \mathcal{T}$  to some node in  $\mathcal{S}$ . According to Lemma C.3, this leads to  $\text{rank } \bar{P}_i(\lambda) = n - 1 + 1 = n$ . In addition, the path from  $x_i$  to some node in  $\mathcal{S}$  is unaffected if we remove all rows corresponding to matrix  $F$  from  $\bar{P}_i(\lambda)$ . Therefore, for generic choices of  $(A, C, F)$  and  $\lambda \in \mathbb{C}$ ,

$$\text{rank } \bar{P}_i(\lambda) = \text{rank} \begin{bmatrix} A_i - \lambda I_{n-1} & \mathbf{b}_i \\ C_i & \mathbf{c}_i \end{bmatrix} = \text{rank } P_i(\lambda). \quad (\text{C.8})$$

For any node  $x_i \notin \mathcal{T} \cup \mathcal{S}$ , either there is a path from  $x_i$  to  $\mathcal{T}$  or there is no path to  $\mathcal{T}$ . We discuss these two cases separately. If there is a path from  $x_i$  to some node  $x_j$  in  $\mathcal{T}$ , by assumption there is a path from  $x_j$  to some node in  $\mathcal{S}$ , so we conclude that there is a path from  $x_i$  to some node in  $\mathcal{S}$ . As a result, equation (C.8) also holds true for this type of node  $x_i$ . Assume now that there is no path from  $x_i$  to  $\mathcal{T}$ , then there are two further possibilities: either there is a path from  $x_i$  to  $\mathcal{S}$  or there is no such path. In the former case,  $\text{rank } \bar{P}_i(\lambda) = \text{rank } P_i(\lambda) = n$  and, in the latter case,  $\text{rank } \bar{P}_i(\lambda) = \text{rank } P_i(\lambda) = n - 1$ . As a result, equation (C.8) still holds true for both cases. In summary, equation (C.8) holds for all  $n$  state nodes.

Now, assume that, for some particular choice of  $\lambda \in \mathbb{C}$ ,  $\text{rank} \begin{bmatrix} A^\top - \lambda I & C^\top & F^\top \end{bmatrix}^\top > \text{rank} \begin{bmatrix} A^\top - \lambda I & C^\top \end{bmatrix}^\top$  holds for any generic realization of  $(A, C, F)$ . Since the rank of both matrices are upper bounded by  $n$ , we have  $\text{rank} \begin{bmatrix} A^\top - \lambda I & C^\top \end{bmatrix}^\top < n$ . It further implies that, with this particular choice of  $\lambda$ ,  $\text{rank } P_i(\lambda) < n$  for any  $x_i \in \mathcal{X}$ . In other words,  $\lambda$  is the common root for polynomial  $\xi_i(\lambda) = \det(P_i(\lambda))$ ,  $\forall x_i \in \mathcal{X}$ . Meanwhile, a necessary condition for  $\text{rank} \begin{bmatrix} A^\top - \lambda I & C^\top \end{bmatrix}^\top < n$  is  $\text{rank}(A - \lambda I) < n$ , i.e.  $\lambda$  is also the root for polynomial  $\xi_0(\lambda) = \det(A - \lambda I)$ . Note that each polynomial  $\xi_i(\lambda)$  does not depend on the free parameters from the  $i$ -th row of  $A$ , and that the polynomial  $\xi_0(\lambda)$  does not contain free parameters from  $C$ . Thus, each free parameter in system matrices  $(A, C)$  does not appear in at least one of these polynomials. As a result, any common root of all polynomials cannot be a function of any of the free parameters. The only possible common root that does not depend on any numerical realization of the free variables is  $\lambda = 0$ , which exists only when  $\begin{bmatrix} A^\top & C^\top \end{bmatrix}^\top$  is rank deficient. Therefore, for any  $\lambda \in \mathbb{C} \setminus \{0\}$ ,  $\text{rank} \begin{bmatrix} A^\top - \lambda I & C^\top & F^\top \end{bmatrix}^\top = \text{rank} \begin{bmatrix} A^\top - \lambda I & C^\top \end{bmatrix}^\top$ .  $\square$

**Lemma C.5** (Sufficiency of condition 2 for (C.1) for  $\lambda = 0$ ).  $\text{rank}[A^\top \ C^\top \ F^\top] = \text{rank}[A^\top \ C^\top]^\top$  if  $\mathcal{T} \cap \mathcal{D} = \emptyset$ .

*Proof.* Let  $\mathcal{D}$  be the union of all minimal dilation sets of  $\mathcal{G}(A, C)$ . Let  $\mathcal{X}_2 = \mathcal{D}$ , and  $\mathcal{X}_1 = \mathcal{X} \setminus \mathcal{X}_2$ , where  $|\mathcal{X}_1| = k$  and  $|\mathcal{X}_2| = n - k$ . After applying a permutation of

coordinates such that the nodes in  $\mathcal{X}_1$  come first, then matrices  $(A, C, F)$  have the form:

$$A = \begin{bmatrix} A_1 & A_2 \end{bmatrix}, C = \begin{bmatrix} C_1 & C_2 \end{bmatrix}, F = \begin{bmatrix} F_1 & \mathbf{0} \end{bmatrix}. \quad (\text{C.9})$$

where  $A_1 \in \mathbb{R}^{k \times n}$ ,  $A_2 \in \mathbb{R}^{n \times (n-k)}$ ,  $C_1 \in \mathbb{R}^{q \times k}$ ,  $C_2 \in \mathbb{R}^{q \times (n-k)}$ , and  $F_1 \in \mathbb{R}^{r \times k}$ . The second block in  $F$  only contains zeros entries due to the assumption  $\mathcal{T} \cap \mathcal{D} = \emptyset$ . Assume that  $\text{rank}[A_1^\top \ C_1^\top]^\top < k$ , which implies that we can find a subset  $\mathcal{D}' \subseteq \mathcal{X}_1$  such that the submatrix formed by the corresponding columns of  $A_1$  contains less than  $|\mathcal{D}'|$  non-zero rows. This means  $\mathcal{X}_1$  contains a dilation and thus contains a minimal dilation set, which contradicts to the assumption that  $\mathcal{D}$  is the union of all minimal dilation sets. As a result,  $\text{rank}[A_1^\top \ C_1^\top]^\top = k$ , i.e.  $\text{row}(F_1) \subseteq \text{row}([A_1^\top \ C_1^\top]^\top)$  and also  $\text{row}(F) \subseteq \text{row}([A^\top \ C^\top]^\top)$  because the second block of  $F$  is all-zeros. Therefore,  $\text{rank}[A^\top \ C^\top \ F^\top] = \text{rank}[A^\top \ C^\top]^\top$ .  $\square$



# Appendix D

## Related Works on “Target Observability”

Under the motivation that controlling the entire state vector of a dynamical system is unfeasible for large-scale network applications, [Gao et al. \(2014\)](#) proposed the concept of *target controllability*, which is based on the concept of *output controllability* from control theory ([Ogata, 2010](#), Section 9.6). Formally, a triple  $(A, B, F)$  is said to be “target controllable” if, for any initial target state  $\mathbf{z}(0) = F\mathbf{x}(0)$  and final target state  $\mathbf{z}(t_1) = F\mathbf{x}(t_1)$ , there exists an input  $\mathbf{u}$  that transfers the output  $\mathbf{z}(0)$  to  $\mathbf{z}(t_1)$  in finite time. A triple  $(A, B, F)$  is output controllable if and only if

$$\text{rank}(FC) = r, \tag{D.1}$$

where  $C = [B \ AB \ A^2B \ \dots \ A^{n-1}B]$  is the controllability matrix, and  $F$  determines the target states desired to be controlled. By duality, a notion of “target observability” can be defined as follows: A triple  $(A, C, F)$  is said to be “target observable” if, for any unknown initial target state  $F\mathbf{x}(0)$ , there exists a finite time  $t_1 > 0$  such that knowledge of the input  $\mathbf{u}$  and output  $\mathbf{y}$  over  $t \in [0, t_1]$  suffices to uniquely determine  $F\mathbf{x}(0)$ . A triple  $(A, C, F)$  is target observable if and only if

$$\text{rank}(\mathcal{O}F^\top) = r. \tag{D.2}$$

Example [D.1](#) illustrates that the functional observability property studied in Chapter 4 and the above definition of target observability are not equivalent properties.

**Example D.1.** Consider the observability condition (??), the functional observability condition (4.2), and the target observability condition (D.2). Let the triple  $(A, C, F)$  be

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ a_{31} & a_{32} & 0 & 0 \\ 0 & 0 & a_{43} & 0 \end{bmatrix}, C = [0 \ 0 \ 0 \ 1], F = [0 \ 1 \ 0 \ 0]. \quad (\text{D.3})$$

Then, from the following conditions

$$\text{rank}(\mathcal{O}) := \text{rank} \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & a_{43} & 0 \\ a_{43}a_{31} & a_{43}a_{32} & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = 3 < 4 := n \quad (\text{observability}),$$

$$\text{rank}(\mathcal{O}F^\top) := \text{rank} \begin{bmatrix} 0 \\ a_{43} \\ 0 \\ 0 \end{bmatrix} = 1 := r \quad (\text{target obsv.}),$$

$$\text{rank} \begin{bmatrix} \mathcal{O} \\ F \end{bmatrix} := \text{rank} \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & a_{43} & 0 \\ a_{43}a_{31} & a_{43}a_{32} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} = 4 > 3 := \text{rank}(\mathcal{O}) \quad (\text{functional obsv.}),$$

we have that the triple  $(A, C, F)$  is neither observable nor functional observable for any choice of parameters in  $A$ , but it is target observable. This can also be verified, respectively, by checking the graph-theoretical conditions for structural observability in Theorem 3.3, structural functional observability in Theorem 4.3, and the dual definition of target controllability in (Li et al., 2019, Theorem 2).  $\triangle$

Compared to previous results on target observability/controllability (Commault et al.; Czeizler et al., 2018; Gao et al., 2014; Klickstein et al., 2017; Li et al., 2019; Wu et al., 2015), our contribution stands out to the research of scalable methods for target state estimation in large-scale systems, not only because our generalization was built under a different dynamical system property from control theory, but because of another subtle reason. As discussed in the main text, functional observability leads to a procedural design algorithm of a functional observer (Darouach, 2000; Fernando et al., 2010b; Hieu and Tyrone, 2012) capable of estimating the target states  $F\mathbf{x}$

*without* requiring state estimation of the *whole* state vector  $\mathbf{x}$ . Thus, there is an order reduction in the observer design that reduces its computational complexity, improving its computational and numerical performance for large-scale applications. On the other hand, target observability (defined, by duality, from (Gao et al., 2014)) is a necessary and sufficient condition for the existence of a *full-state* observer that guarantees the asymptotic convergence of  $\|F\hat{\mathbf{x}}(t) - F\mathbf{x}(t)\|$ , where  $\hat{\mathbf{x}}(t)$  is the observer estimate of the true values  $\mathbf{x}(t)$ . Thus, target observability is a property that, like observability, still leads to the design of a  $n$ -dimensional observer, with the difference that it is concerned in guaranteeing that *only* the estimates  $F\hat{\mathbf{x}}(t)$  approach the true value  $F\mathbf{x}(t)$  (without “caring” for the rest of the state estimates). In short, it is true that functional observability is a stronger condition than output observability, but it leads to a model reduction in the design of a “target observer”—an important feature for large-scale applications.

