ATTACKS AND COUNTERMEASURES
IN COMMUNICATIONS AND POWER NETWORKS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Jinsub Kim

January 2014

ATTACKS AND COUNTERMEASURES

IN COMMUNICATIONS AND POWER NETWORKS

Jinsub Kim, Ph.D.

Cornell University 2014

The threat of malicious network attacks has become significant ever since networking became pervasive in our life. When adversaries have enough control over the network measurements and control procedures, the effect of attacks can be as detrimental as the breakdown of the whole network operations. This dissertation studies possible adversarial effects under certain protection strategy, the conditions under which attacks can be detected, and protection strategies to render attacks detectable. Specifically, attacks on two types of networks are considered: communications networks and power networks.

First, we consider an attack on communications networks, where a pair of nodes are suspected to belong to the chain of compromised nodes used by the adversary. If the pair belongs to the compromised chain, it forwards attack packets along the chain, and thus there should exist an information flow between the pair. Detection of an information flow based on node transmission timings is formulated as a binary composite hypothesis testing. An unsupervised and nonparametric detector with linear complexity is proposed and tested with real-world TCP traces and MSN VoIP traces. The detector is proved to be consistent for a class of nonhomogeneous Poisson processes.

Secondly, the topology attack on power networks is studied. In a so-called man-in-the-middle topology attack, an adversary alters data from certain meters and network switches to mislead the control center with an incorrect network topology

while avoiding detection by the control center. A necessary and sufficient condition for the existence of an undetectable attack is obtained, and countermeasures to prevent undetectable attacks are presented. It is shown that any topology attack is detectable if a set of meters satisfying a certain branch covering property are protected from adversarial data modification. The proposed attacks are tested with IEEE 14-bus and IEEE 118-bus system, and their effect on real-time locational marginal pricing is examined.

Lastly, a new attack mechanism aimed at misleading the power system control center about the source of data attacks is proposed. As a man-in-the-middle state attack, a data framing attack is proposed to exploit the bad data detection and identification mechanisms at the control center. In particular, the proposed attack frames normal meters as sources of bad data and causes the control center to remove useful measurements from the framed meters. The optimal design of data framing attack is formulated as a quadratically constrained quadratic program (QCQP). It is shown that the proposed attack is capable of perturbing the power system state estimate by an arbitrary degree using only half of the critical measurements. Implications of this attack on power system operations are discussed, and the attack performance is evaluated using benchmark systems.

BIOGRAPHICAL SKETCH

Jinsub Kim was born in Seoul, Republic of Korea, in 1984. He earned the bachelor's degree in Electrical Engineering from KAIST, Republic of Korea, in 2007, and came to Ithaca for the graduate study in Electrical and Computer Engineering at Cornell University. Since January 2008, he has been with the Adaptive Communications and Signal Processing group under the guidance of Prof. Lang Tong. He has been working on statistical inference for anomaly detection in communications networks and power networks. Besides working on research puzzles, he enjoys spending time with his family, Soo Yeun and Noah, rambling around Ithaca, watching movies, and listening to all kinds of good old music.

For my parents, Young Jin Kim and Ok Hee Kim

For my wife and son, Soo Yeun Lim and Noah Jaehyun Kim

ACKNOWLEDGEMENTS

First of all, I would like to thank Prof. Lang Tong for his dedicated guidance and support throughout my life as a graduate student. He is definitely the best advisor I can imagine and one of the most enthusiastic person I have ever met. To me, he has been an excellent role model as a researcher, a teacher, and a leader. He has influenced me in so many good ways. I should also thank him for teaching me the fun of doing research. In addition, I thank him for being very supportive and considerate when I was going through hard times.

I would like to thank Prof. Aaron B. Wagner and Prof. Ping Li for serving as my committee members and providing helpful guidance. I also thank Prof. A. Kevin Tang for attending my A exam and providing many helpful suggestions.

I want to thank the members in our research group: Stefan Geirhofer, Animashree Anandkumar, Oliver Kosut, Brandon M. Jones, Shiyao Chen, Liyan Jia, Zhe Yu, Yuting Ji, Daniel Munoz Alvarez, and Tirza Routtenberg. They have provided me with helpful research discussions, personal supports, and lots of small talks. I thank them for all the fun we had together.

I also want to thank other professors, who taught me at Cornell, for giving me wonderful lectures: Prof. A. Salman Avestimehr, Prof. C. Richard Johnson, Prof. David Matteson, Prof. R. Keith Dennis, Prof. Michael Nussbaum, and Prof. Richard Durrett.

I would like to thank my parents, Young Jin Kim and Ok Hee Kim. They have helped me all the way since October 1984. I thank them for everything I am now.

I thank my wife, Soo Yeun Lim, for being there with me all the time and being a good mother. And, I thank my son, Noah Jaehyun Kim, for just being there.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Motivation and Overview

Since the advent of computer networks, networks among people and devices have grown rapidly in their sizes and capabilities. Nowadays, the majority of people are connected to cellular or computer networks most of time, and our reliance on communications networks has never been more tremendous. In addition to communications networks, power networks assume an extremely crucial role in supporting our daily life: power networks enable reliable delivery of electricity to our homes, work places, and physical infrastructures.

For proper operations, a network has to be protected from possible attacks. As the role of networks became important, potential effects of network attacks also became significant. For instance, an adversary in a data network may hack into a server to attain unauthorized data thereby possibly causing privacy data leakage. In power networks, its cyber-physical nature allows an adversary to create even worse consequences. For instance, an adversary may alter meter data to mislead the control center about the current operating condition. Such an attack may possibly leads to breakdown of power plants, electricity price perturbation, and even a blackout in the worst case. Such attacks on networks have been continuously reported thereby proving the presence of threats.

Fortunately, many attacks leave traces in the network measurements (*e.g.*, network log files, measurements from deployed sensors.) However, it is nontrivial how to detect presence of an attack based on the measurements. Smart adversaries will

attempt to hide their traces, and it is indeed possible if they have enough controls on the network or the measurements. Furthermore, sometimes, an attack may not leave a strong signature in the network measurements. Therefore, a detection algorithm needs to be carefully designed, and the fundamental limitation due to a strong adversary needs to be studied. This dissertation studies protection strategies to render attacks detectable and conditions under which a smart adversary can launch an attack without leaving any detectable trace.

We first consider an attack on communications networks in Chapter 2. We consider the so-called stepping stone attack [1], in which the attacker uses a chain of compromised nodes to access the victim. This strategy is often used to confuse the intrusion detection system about the adversary's location. If the adversary compromises a pair of supposedly independent nodes and use them as stepping stones to the victim, there should exist information flows (associated with the attack) between the pair. Given suspect nodes for stepping stones, detection of information flows can be applied to trace back the stepping stone chain. We formulate the problem as a binary composite hypothesis testing and present an unsupervised and nonparametric detection algorithm.

Then, we move to attacks on power networks in Chapter 3. In a power network, the control center periodically collects measurements from meters and sensors deployed throughout the network. These measurements are used in estimating the real-time system state and the network topology. We study a specific type of data attack that alters part of the measurements to mislead the control center with an incorrect network topology. Such attacks may cause the control center to believe in false contingency information or delay preventive actions when important transmission lines are tripped. Unless the adversary can control a sufficient number of

measurements, an attempt to disturb the topology estimate causes inconsistency among the measurements, and this anomaly can be well detected by the legacy bad data test. We provide a necessary and sufficient condition under which an attack can be successful without causing any detectable anomaly and present construction of undetectable attacks. Then, the necessary and sufficient condition is used to develop a graph-theoretical meter protection strategy.

In Chapter 4, a data framing attack on power system state estimation is presented. The framing attack is a new approach of data attack on state estimation which misleads the control center that certain normally operating meters are responsible for generating biased measurements. The bad data identification rule falsely identifies the data from these meters as bad and remove them from system state estimation. Such an attack may degrade the accuracy of state estimation and even make the network vulnerable to arbitrary perturbation of the state estimate. We formulate the optimal design of the framing attack as a quadratically constrained quadratic program and show that the framing attack needs to alter only half of the critical set of measurements to perturb the state estimate by an arbitrary degree. The proposed attack is evaluated with the IEEE 14-bus and 118-bus networks.

The following three sections give more details about related works and our contributions in the aforementioned three topics.

1.2 Detection of Information Flows

Detection of information flows between a pair of nodes has been studied in the context of network intrusion detection, especially in the detection of interactive

stepping-stone attacks [1]. The use of only transmission timing measurements for detection is motivated by the fact that packets involved in an attack can be easily encrypted. Even though transmission timings of nodes can be easily monitored, detecting information flows based on timing is non-trivial. One main source of difficulty is the presence of noise-like epochs. When an information flow exists between the two nodes, the two nodes may have transmissions that do not belong to the flow. They may multiplex transmissions of other flows that go through only one of the two nodes, or intentionally superpose dummy transmissions to avoid detection. We refer to such transmissions as *chaff* transmissions.

1.2.1   Related Works

Donoho *et al.* [1] were among the first to consider the flow model with a uniform delay bound. Following their model, many algorithms have been proposed to detect a flow with a delay constraint. As an *active* detection scheme, Wang *et al.* [2] proposed a watermark-based detector which embeds watermarks by slightly adjusting transmission timings of a node; if the same watermarks are detected in another node, two nodes are claimed to have flows between them. Their work was followed by a large number of watermark-based detectors [3–10]. The insertion of watermarks, however, requires the ability of the detector to modify traffic at different locations of the network, which may not be possible in practical situations.

If the network traffic cannot be modified to facilitate detection, the problem is referred to as *passive* flow detection, and it is the problem of our interest. In passive detection, a detector collects transmission timing measurements and analyzes them to draw a conclusion about presence of a flow. Many research efforts have been made to develop effective passive detectors. Zhang *et al.* [11, 12] proposed

matching-based algorithms. However, they assumed that only one of two nodes can insert chaff transmissions, and their algorithms are vulnerable to chaff insertion at both nodes. Donoho *et al.* [1] proposed a wavelet analysis with a claim that it can detect a flow in chaff if the chaff part is independent of the flow part and the sample size is sufficiently large. Blum *et al.* [13] presented a counting-based method which was shown to be able to detect a flow in chaff if the fraction of chaff is small enough. Under the Poisson traffic assumption, they characterized the sufficient sample size for satisfying a given false alarm probability constraint. However, their method may result in high miss detection probability if chaff transmissions are bursty. He and Tong [14] proposed a matching-based detector with better chaff tolerance and characterized the maximum tolerable fraction of chaff under the homogeneous Poisson traffic assumption. Their approach requires choosing a detection threshold which is a function of the parameter of the underlying Poisson traffic. When the traffic deviates from the Possion model, the detection algorithm is not always robust. The approach in [14] can be applied to the general traffic if a training data with a sufficiently long time span is available. Coskun and Memon [15, 16] presented detectors based on random projection of transmission processes. Similar to [14], their methods also require choosing an appropriate detection threshold, which can be successful only if a large volume of training data or an accurate parametric model is available.

1.2.2 Contributions

Our results include three parts: a nonparametric passive detection algorithm for unidirectional or bidirectional flows, the related performance analysis, and experiments with synthetic and real data. In developing an algorithm, our main contri-

bution is a new nonparametric technique that does not rely on knowledge of traffic distribution; nor does it require a training data for either hypothesis. The key idea lies in a particular transformation of the measurements that leads to distinct statistical behaviors under two different hypotheses. The proposed detector does not assume stationarity of traffic and hence is applicable in time-varying traffic conditions. Furthermore, it is memory-efficient and has linear computational complexity with respect to the sample size thereby making real-time inference feasible.

In algorithm analysis, we aim to give theoretical justifications for the proposed approach. To this end, we establish the consistency property of the proposed detector for a class of non-homogeneous Poisson traffic. Even though the detector is analyzed only for non-homogeneous Poisson traffic, the intuition behind it suggests that it may perform well on the traffic with more general distribution.

The performance of our detector is evaluated using synthetic Poisson traffic, LBL TCP traces [17], and real-world measurements from MSN VoIP sessions, and comparison with other passive detectors is provided. The use of synthetic data allows us to examine the trade-offs between miss detection and false alarm probabilities using Monte Carlo simulations. LBL TCP traces and MSN VoIP traces are of course not guaranteed to satisfy the assumptions made in our algorithm analysis, and our results indicate a level of robustness.

1.3   Topology Attack of a Power Grid

Liu, Ning, and Reiter [18] appear to be the first to introduce the concept of data injection attack (also referred to as malicious data attack) of a power grid. Assuming that the attacker is capable of altering data from a set of meters, a similar

scenario assumed in our problem setting, the authors of [18] show that if the set of compromised meters satisfies certain condition, the adversary can perturb the network state by an arbitrarily large amount without being detected by any detector. In other words, the data attack considered in [18] is undetectable. The main difference between [18] and our work is that the attacks considered in [18] perturb only the network state, not the network topology. It is thus most appropriate to refer to attacks in [18] and many follow-ups as *state attack*, in distinguishing the *topology attack* considered in our work.

1.3.1   Related Works

The work in [18] is influential; it has inspired many further developments, *e.g.*, [19–22] and references therein, all focusing on state attacks. A key observation is made by Kosut *et al.* in [23, 24], showing that the condition of non-existence of an undetectable attack is equivalent to that of network observability [25, 26]. This observation leads to graph theoretic techniques that characterize network vulnerability [24]. The condition to be presented in Chapter 3 on the non-existence of an undetectable topology attack mirrors the state attack counterpart in [24].

The problem of adding protection on a set of meters to prevent undetectable state attacks was considered by Bobba *et al.* [19]. We consider the same problem in the context of topology attack. While meter protection problem for state attacks is equivalent to protecting a sufficient number of meters to ensure observability [19, 24], the corresponding problem for topology attacks is somewhat different and more challenging.

The problem of detecting topology error from meter data is in fact a classical

problem, casted as part of the bad data detection problem [27–29]. Monticelli [30] pioneers the so-called generalized state estimation approach where, once the state estimate fails the bad data test, modifications of topology that best represent the meter data are considered. Abur *et al.* [31] extend this idea to the least absolute value state estimation formulation, and Mili *et al.* [32] apply the idea to the state estimation with the Huber M-estimator. Extensive works followed to improve computational efficiency, estimation accuracy, and convergence property over the aforementioned methods (*e.g.*, see [33–35] and references therein).

Finally, there is a limited discussion on the impact of a malicious data attack on power system operations. Should state estimates be used in closed-loop control of the power grid, such an attack may cause serious stability problems. The current state of the art, however, uses state estimates for real-time dispatch only in a limited fashion. However, state estimates are used extensively in calculating real-time locational marginal price (LMP) [36]. Thus, attacks that affect state estimates will affect the real-time LMP calculation [37–39]. The way that a topology attack affects LMP is significantly different from that of a state attack. We demonstrate that a topology attack has significant impact on real-time LMP.

1.3.2 Contributions

First, we characterize conditions under which undetectable attacks are possible, given a set of vulnerable meters that may be controlled by an adversary. To this end, we consider two attack regimes based on the *information set* available to the attacker. The more information the attacker has, the stronger its ability to launch a sophisticated attack that is hard to detect.

The *global information* regime is where the attacker can observe all meter and network data before altering the adversary-controlled part of them. Although it is unlikely in practice that an adversary is able to operate in such a regime, in analyzing the impact of attacks, it is typical to consider the worst case by granting the adversary additional power. We present a necessary and sufficient algebraic condition under which, given a set of adversary controlled meters, there exists an undetectable attack that misleads the control center with an incorrect "target" topology. This algebraic condition provides not only numerical ways to check if the grid is vulnerable to undetectable attacks but also insights into which meters to protect to defend against topology attacks. We also provide specific constructions of attacks and show certain optimality of the proposed attacks.

A more practically significant situation is the *local information* regime where the attacker has only local information from those meters it has gained control. We present that under certain conditions, undetectable attacks exist and can be implemented easily based on simple heuristics.

Secondly, we study conditions under which any topology attack can be made detectable. Such a condition, even if it may not be the tightest, provides insights into defense mechanisms against topology attacks. We show that if a set of meters satisfying a certain branch covering property are protected, then topology attacks can always be detected.

1.4   Framing Attack on State Estimation

In power system state estimation, it is well known that bad data identification rules may mistakenly identify good data entries as bad and remove them [40, 41].

We study how such an inherent weakness of the bad data test can be exploited by adversaries.

1.4.1 Related Works

We consider a man-in-the-middle attack, where an adversary can alter part of meter measurements such that the control center is misled with the partially corrupt measurements.

In [18], Liu, Ning, and Reiter presented perhaps the first man-in-the-middle (MiM) attack on the power system state estimation where an adversary replaces "normal" sensor data with "malicious data." It was shown that, if the adversary could gain control of a sufficient number of meters, it could perturb the state estimate by an arbitrary amount without being detected by the bad data detector employed at the control center. Such undetectable attacks are referred to as *covert data attacks*.

There is an extensive literature on covert data attacks, following the work of Liu, Ning, and Reiter [18]. While the data framing attack mechanism proposed here is fundamentally different, insights gained in existing work are particularly relevant. Here, we highlight some of these ideas in the literature.

The explicit link between covert attack on state estimation and system observability was made in [19, 23]. Consequently, classical observability conditions [25, 26, 42] can be modified for that for covert attacks and used to develop meter protection strategies [19, 24, 43–46]. A particularly important concept is the notion of critical set of meters (or critical measurements) [26, 47, 48]. In assessing the vulnerability of the grid, the minimum number of adversary meters necessary

for a covert attack was suggested as the security index for the grid [20, 24]. Subsequently, meter protection strategies were proposed in [21, 22] to maximize the security index under the protection resource constraint.

The framing attack strategy considered here relies on bad data identification and removal techniques that have long been subjects of study [40,41,47,49,50]. See [51, 52] and references therein. Typically, the residue vectors in normalized forms are widely used as statistics for the bad data test [40]. In particular, Mili *et al.* [50] proposed a hypothesis testing method, in which the set of suspect measurements are determined by the residue analysis in [40]. The use of non-quadratic cost functions in state estimation was also studied to enhance the bad data identification performance. Especially, the weighted least absolute value estimation [53–56] and the least median of squares regression [57,58] were considered as alternatives with comparably good performance. In this dissertation, we take the residue analysis in [40] as a representative bad data test and analyze the effect of the framing attack. However, the same analysis is applicable to general bad data tests.

Detection of data attacks on state estimation, referred to as *state attacks*, has been also studied in various frameworks. Kosut *et al.* [24] presented a generalized likelihood ratio test for detection. Morrow *et al.* [59] proposed the detection mechanism based on network parameter perturbation which deliberately modifies the line parameters and probes whether the measurements respond accordingly to the modification. Distributed detection and estimation of adversarial perturbation was also studied in [60]. In an effort to minimize the detection delay, the attack detection was also formulated as a quickest detection problem, and modified CUSUM algorithms were proposed [61–63].

1.4.2 Contributions

We propose a data framing attack on power system state estimation. Specifically, we formulate the design of optimal data framing attack as a quadratically constrained quadratic program (QCQP). To analyze the efficacy of the data framing attack, we present a sufficient condition under which the framing attack can achieve an arbitrary perturbation of the state estimate by controlling only half of the critical set of meters. We demonstrate with the IEEE 14-bus and 118-bus networks that the sufficient condition holds in critical sets associated with cuts.

The optimal design of framing attack is based on a linearized system. In practice, a nonlinear state estimator is often used. We demonstrate that, under the nonlinear measurement model, the framing attacks designed based on linearized system model successfully perturb the state estimate, and the adversary can control the degree of perturbation as desired.

1.5 Organization

In Chapter 2, we consider detection of an attack-associated information flow in communications networks. Specifically, the problem is formulated as detection of an information flow based on timing measurements. We first start with a simpler case of detection with a parametric flow model. Then, we present an unsupervised and nonparametric flow detector. The detector is proved to be consistent for a class of non-homogeneous Poisson traffic model. Lastly, the detector is tested and compared with other benchmark techniques using real-world TCP and VoIP traces.

From Chapter 3, we consider data attacks in power networks. A data attack

aimed at perturbing the topology estimate of the control center is studied. We first study the attack for an adversary with global information. A necessary and sufficient condition for an undetectable topology attack is presented, and the condition is used to construct a simple graph-theoretical meter protection strategy. Then, we consider an adversary with local information. An undetectable local attack is presented and tested with the IEEE 14-bus and 118-bus networks.

In Chapter 4, we study a data framing attack on power system state estimation. We present a new attack approach, which alters the adversary-controlled measurements deliberately such that the bad data detection and identification rule falsely removes measurements from normally operating meters while retaining adversarially altered measurements. We first present the main idea of the attack and then provide the optimization framework for the attack design. Controlling only half of a critical set of meters, the proposed attack is shown to be able to perturb the state estimate by an arbitrary degree. The numerical results with the IEEE benchmark networks are provided.

Finally, Chapter 5 provides concluding remarks and comments on future works.

DETECTION OF INFORMATION FLOWS

2.1 Introduction

We consider the problem of detecting information flows through a pair of monitored nodes as illustrated in Fig. 2.1. In particular, given the measurements of transmission timings from the monitored nodes, we are interested in determining whether the two monitored nodes are engaged in relaying packets of certain information flows (the alternative hypothesis), or they are merely transmitting independently (the null hypothesis). The network of our interest can be either wireless or wired as long as transmission timings can be measured.

The generic problem of flow detection arises from a number of practical applications, especially in the context of information forensics, network surveillance, and anonymous networking. For example, in the so-called stepping-stone attack [1] in a network, an adversary may attack a node by compromising a sequence of nodes that serve as stepping stones. When the attacker is involved in an interactive session (*e.g.*, SSH), a flow of packets travel through a chain of stepping stones. By
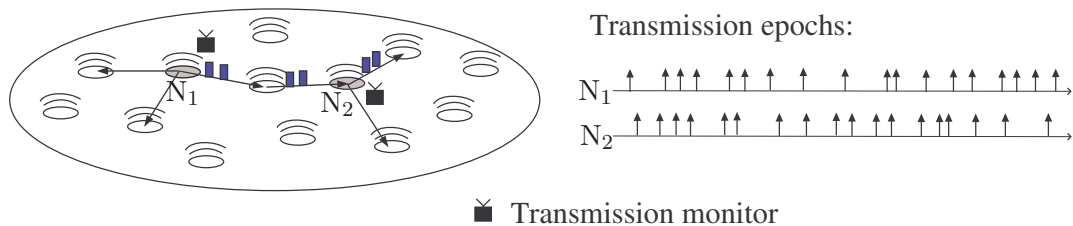
Figure 2.1: In the above wireless network, the transmission timings of two nodes, N_1 and N_2, are recorded. The horizontal axis is the time axis, and arrows represent packet transmissions at different time points. As illustrated, packets of certain information flows may travel through N_1 and N_2.

detecting the presence of unexpected flows through monitored nodes, the network owner can alert the possibility of an attack. Other applications include the detection of wormhole attack [64] in which a set of colluding nodes divert a valid network flow through a "wormhole tunnel." Understanding the problem of flow detection is also valuable for the design and assessment of anonymous networks [65, 66].

We restrict ourselves to the use of timing measurements only. Such a restriction is of course unnecessary because there are often other information available such as source-destination addresses, packet statistics, etc.; a detector should incorporate such side information. We choose to focus exclusively on the use of timing information for two reasons. First, timing can only be distorted but cannot be hidden by the transmitter, and its measurements can be obtained by simple devices. In contrast, source-destination addresses and packet characteristics can be masked using standard techniques in anonymous networking [66]. Second, timing is a fundamental traffic characteristic. It is therefore useful to understand the extent that timing reveals the presence of information flows. Furthermore, any side information, when incorporated properly, will enhance the performance of techniques based solely on timing information.

Even though transmission timings of nodes can be easily monitored, detecting information flows based on timing measurements is non-trivial, partly because of non-stationary traffic characteristics: transmission timings of nodes often have time-varying intensities, and they may be bursty when interactive users are involved. Moreover, in general, it is difficult to obtain an accurate parametric model for the monitored traffic, especially when there is no prior knowledge about the nature of the traffic and no training data available. The presence of noise-like epochs is another source of difficulty. When an information flow travels through

two nodes, the two nodes may have transmissions that do not belong to the flow. They may multiplex transmissions of other flows that go through only one of the two nodes, or intentionally superpose dummy transmissions to avoid detection. We refer to the epochs of such transmissions as *chaff* epochs.

It is easy to see that, if a node can arbitrarily delay packets in a flow, timing information is insufficient for detection. For latency-sensitive applications such as VoIP, multimedia streaming, etc., however, packets must satisfy certain end-to-end delay constraints, which make the presence of such flows detectable. For instance, VoIP applications require end-to-end delays to be bounded above by 150 msec [67]. We will consider the constraint that flow packets should satisfy the end-to-end delay constraint of Δ seconds.

2.1.1 Summary of Results and Organization

Our results include three parts: a nonparametric flow detection algorithm for unidirectional or bidirectional flows, the related performance analysis, and experiments with synthetic and real data. In developing an algorithm, our main contribution is a new nonparametric technique that does not rely on knowledge of traffic distribution; nor does it require a training data for either hypothesis. The key idea lies in a particular transformation of the measurements that leads to distinct statistical behaviors under two different hypotheses. The proposed detector does not assume stationarity of traffic and hence is applicable in time-varying traffic conditions. Furthermore, it is memory-efficient and has linear computational complexity with respect to the sample size thereby making real-time inference feasible.

In algorithm analysis, we aim to give theoretical justifications for the proposed

approach. To this end, we establish the consistency property of the proposed detector for a class of non-homogeneous Poisson traffic.

The performance of our detector is evaluated using synthetic Poisson traffic, LBL TCP traces [17], and real-world measurements from MSN VoIP sessions, and comparison with other benchmark passive detectors is provided. LBL TCP traces and MSN VoIP traces are of course not guaranteed to satisfy the assumptions made in our algorithm analysis, and our results indicate a level of robustness.

The rest of the chapter is organized as follows. Section 2.2 gives the notations and definitions employed throughout the chapter and formulates flow detection as a binary composite hypothesis testing problem. In Section 2.3, we consider the simpler case where the parametric model of the traffic is available. Then, Section 2.4 presents a nonparametric flow detection algorithm and its consistency property. In Section 2.5, the proposed detector is evaluated using synthetic Poisson traffic, LBL TCP traces, and MSN VoIP traffic. The proofs of theorems are given in Section 2.6.

2.2 Mathematical Formulation

This section introduces notations and definitions and formulates flow detection as one of binary composite hypothesis testing.

2.2.1 Notations and Flow Models

Transmission timings of each node are modeled as a point process on $[0, \infty)$, and detectors begin recording the timings at time 0. Bold upper-case letters (*e.g.*, $\mathbf{S}$) denote point processes, and bold lower-case letters (*e.g.*, $\mathbf{s}$) denote their realizations. $S(i)$ represents the ith epoch (*i.e.*, the time of the ith transmission) of $\mathbf{S}$, and $s(i)$ is its realization. The upper-case script letter $\mathcal{S}$ denotes the set of epochs in the realization $\mathbf{s}$: $\mathcal{S} \triangleq \{s(i), i \geq 1\}$. In addition, we define a *superposition* operator $\bigoplus$: given two increasing sequences $(a_i)_{i=1}^\infty$ and $(b_i)_{i=1}^\infty$, $(a_i)_{i=1}^\infty \oplus (b_i)_{i=1}^\infty = (c_i)_{i=1}^\infty$, where c_i is the ith element of the sequence of all the elements of $(a_i)_{i=1}^\infty$ and $(b_i)_{i=1}^\infty$ ordered in the increasing order.

First, we define a *unidirectional flow* as follow.

Definition 2.2.1 *An ordered pair of point processes* $(\mathbf{F}_1, \mathbf{F}_2)$ *forms a* unidirectional flow, *if for any realization* $(\mathbf{f}_1, \mathbf{f}_2)$ *there exists a bijection* $g : \mathcal{F}_1 \rightarrow \mathcal{F}_2$ *satisfying* $g(s) - s \in [0, \Delta]$ *for all* $s \in \mathcal{F}_1$.

As illustrated in Fig. 2.2, when packets of an information flow travel through node N_1 and node N_2, $\mathbf{F}_1$ and $\mathbf{F}_2$ can be interpreted as the transmission timings of the flow packets at N_1 and N_2 respectively. The bijection condition of g means packet conservation; every flow packet sent by N_1 is received and forwarded by N_2. The condition $g(s) - s \in [0, \Delta]$ means that every flow packet transmission satisfies causality and the delay constraint Δ. Based on the above definition, we define a *bidirectional flow* as a superposition of two unidirectional flows with opposite directions.

Definition 2.2.2 *A pair of point processes* $(\mathbf{F}_1, \mathbf{F}_2)$ *forms a* bidirectional flow, *if*
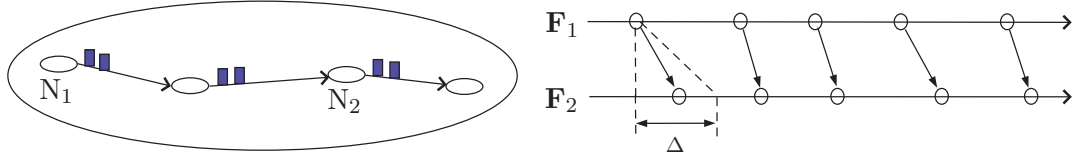
Figure 2.2: Every packet transmission of a unidirectional flow is assumed to satisfy packet conservation, causality, and the delay constraint Δ.

$\mathbf{F}_i$ *can be decomposed into* $\mathbf{F}_i^{12}$ *and* $\mathbf{F}_i^{21}$ *(i.e.,* $\mathbf{F}_i = \mathbf{F}_i^{12} \oplus \mathbf{F}_i^{21}$*) such that* $(\mathbf{F}_1^{12}, \mathbf{F}_2^{12})$ *and* $(\mathbf{F}_2^{21}, \mathbf{F}_1^{21})$ *are unidirectional flows.*

We allow $(\mathbf{F}_1^{12}, \mathbf{F}_2^{12})$ and $(\mathbf{F}_2^{21}, \mathbf{F}_1^{21})$ to have zero rate, so that a unidirectional flow is a special case of a bidirectional flow.

2.2.2 Problem Statement

We formulate detection of bidirectional flow as a binary composite hypothesis testing problem. Let $\mathbf{S}_1$ and $\mathbf{S}_2$ denote the transmission processes of N_1 and N_2, respectively. Given the measurements $(\mathbf{s}_i)_{i=1}^2$ in $[0, t]$, we test the following hypotheses:

$$\mathcal{H}_0 : \quad \mathbf{S}_1 \text{ and } \mathbf{S}_2 \text{ are independent;}$$
$$\mathcal{H}_1 : \quad \mathbf{S}_i = \mathbf{F}_i \oplus \mathbf{W}_i, \ i = 1, 2, \text{ and } (\mathbf{F}_1, \mathbf{F}_2) \text{ forms a bidirectional flow.}$$

(2.1)

We further assume that, under $\mathcal{H}_1$,

1. $\mathbf{F}_1$ and $\mathbf{F}_2$ are point processes with non-zero rates[1].

2. $\mathbf{F}_1$ and $\mathbf{F}_2$ are not independent.

---

[1] In other words, if $N_{\mathbf{F}_i}(t)$ denotes the number of epochs of $\mathbf{F}_i$ in $[0, t]$, there exists $\delta > 0$ such that $\liminf\limits_{t \to \infty} \dfrac{N_{\mathbf{F}_i}(t)}{t} \geq \delta$ almost surely, $i = 1, 2$.

3. $(\mathbf{F}_1, \mathbf{F}_2)$, $\mathbf{W}_1$, and $\mathbf{W}_2$ are independent.

$\mathcal{H}_0$ corresponds to the scenario that N_1 and N_2 have independent transmissions. $\mathcal{H}_1$ corresponds to the scenario that N_1 and N_2 relay packets of information flows in either or both directions: $(\mathbf{F}_i)_{i=1}^2$ and $(\mathbf{W}_i)_{i=1}^2$ represent the flow part and the chaff part, respectively. Note that under both hypotheses, no restriction is imposed on the marginal distributions of $\mathbf{S}_i$, $\mathbf{F}_i$, and $\mathbf{W}_i$.

The assumptions under $\mathcal{H}_1$ are imposed to make two hypotheses disjoint. The first assumption implies that the bidirectional flow should have positive rate. The second assumption means that the flow parts of N_1 and N_2 should not be independent, and this assumption is expected to hold in general due to the delay constraint Δ. The third assumption implies that the chaff parts of N_1 and N_2 are independent, and they are also independent of the flow part. We note here that the third assumption is more restrictive than that used in earlier works [13, 14].

We employ the notion of Chernoff consistency [68] to evaluate the asymptotic performance of detectors.

Definition 2.2.3 *For $j = 0$, 1, $\mathcal{P}_j$ denotes the set of all possible distributions of $(\mathbf{S}_i)_{i=1}^2$ under $\mathcal{H}_j$. A detector $\delta((\mathbf{s}_i)_{i=1}^2, t)$ is a function of the epochs of $(\mathbf{s}_i)_{i=1}^2$ in $[0, t]$, which is equal to j if the decision is $\mathcal{H}_j$. $\delta((\mathbf{s}_i)_{i=1}^2, t)$ is said to be* consistent *if*

1. $\forall\, Q_0 \in \mathcal{P}_0$, $\lim\limits_{t\to\infty} Q_0(\delta((\mathbf{S}_i)_{i=1}^2, t) = 1) = 0$, and

2. $\forall\, Q_1 \in \mathcal{P}_1$, $\lim\limits_{t\to\infty} Q_1(\delta((\mathbf{S}_i)_{i=1}^2, t) = 0) = 0$.

In other words, a detector is consistent if its false alarm and miss detection probabilities vanish as t grows, under all possible distributions in $\mathcal{P}_0$ and $\mathcal{P}_1$. In the following sections, we will reduce $\mathcal{P}_0$ and $\mathcal{P}_1$ to the sets of distributions satisfying certain additional conditions, and prove the consistency of our detection algorithms.

Intuitively, the greater the amount of chaff epochs, the harder the flow detection becomes. To measure the relative strength of the flow part with respect to the chaff part, we introduce the following definition of *flow fraction*.

Definition 2.2.4 *Under* $\mathcal{H}_1$, *suppose that* $(\mathbf{S}_i)_{i=1}^2$ *consists of the bidirectional flow* $(\mathbf{F}_i)_{i=1}^2$ *and the chaff part* $(\mathbf{W}_i)_{i=1}^2$. *Given a realization* $(\mathbf{s}_i)_{i=1}^2$, *where* $\mathbf{s}_i = \mathbf{f}_i \oplus \mathbf{w}_i$, $i = 1, 2$, *the* flow fraction *of* $(\mathbf{s}_i)_{i=1}^2$ *is defined as*

$$
\mathrm{R}(t) \triangleq \frac{\displaystyle\sum_{i=1}^2 |\mathcal{F}_i \cap [0, t]|}{\displaystyle\sum_{i=1}^2 |\mathcal{S}_i \cap [0, t]|}, \qquad \mathrm{R} \triangleq \liminf_{t \to \infty} \mathrm{R}(t) \tag{2.2}
$$

where $|\mathcal{F}_i \cap [0, t]|$ *is the number of flow packet transmissions at node* N_i *in* $[0, t]$, *and* $|\mathcal{S}_i \cap [0, t]|$ *is the number of total transmissions at node* N_i *in* $[0, t]$.

In other words, $\mathrm{R}(t)$ is the fraction of the flow epochs in the measurements up to time t, and R is its limiting value.

2.3   Parametric Flow Detection

We begin with an easier case where an accurate parametric model for traffic is available. The main result in this section is a simple algorithm that computes, for

measurements $(\mathbf{s}_i)_{i=1}^2$, the maximum schedulable flow fraction ($\bar{\mathrm{R}}$) as our decision statistic. The flow detection algorithm is a threshold decision rule based on $\bar{\mathrm{R}}$. The computation of the threshold, however, requires the knowledge of the traffic distribution under $\mathcal{H}_0$, which we assume available at the moment; this assumption is removed in Section 2.4.

2.3.1 Decision Statistic: Maximum Schedulable Flow Fraction

Under both hypotheses, given a realization $(\mathbf{s}_i)_{i=1}^2$ in $[0, t]$, its *maximum schedulable flow fraction* $\bar{\mathrm{R}}(t)$ is defined as

$$\bar{\mathrm{R}}(t) \triangleq \max_{\{(\mathbf{f}_i, \mathbf{w}_i)_{i=1}^2 \ : \ \mathbf{s}_i = \mathbf{f}_i \oplus \mathbf{w}_i \sim \mathcal{H}_1\}} \frac{\displaystyle\sum_{i=1}^2 |\mathcal{F}_i \cap [0, t]|}{\displaystyle\sum_{i=1}^2 |\mathcal{S}_i \cap [0, t]|}$$

where $\mathbf{s}_i = \mathbf{f}_i \oplus \mathbf{w}_i \sim \mathcal{H}_1$ denotes the constraint that $\mathbf{s}_i = \mathbf{f}_i \oplus \mathbf{w}_i$, $i = 1, 2$, and $(\mathbf{f}_1, \mathbf{f}_2)$ is a realization[2] of a bidirectional flow. In other words, we schedule a maximum number of bidirectional flow transmissions between $\mathbf{s}_1$ and $\mathbf{s}_2$ in $[0, t]$, and denote the fraction of the flow part by $\bar{\mathrm{R}}(t)$.

To effectively evaluate $\bar{\mathrm{R}}(t)$, we propose a matching algorithm called Bidirectional-Bounded-Greedy-Match (BiBGM). To achieve its goal, BiBGM starts with the first epoch in $\mathcal{S}_1 \cup \mathcal{S}_2$, and subsequently finds the earliest one-to-one matches satisfying causality and the delay constraint. We explain below the operation of BiBGM using an example in Fig. 2.3 accompanied by a pseudocode

---

[2] In other words, $\mathbf{f}_i$ can be partitioned into two subsequences $\mathbf{f}_i^{12}$ and $\mathbf{f}_i^{21}$ such that there exist bijections $g_1 : \mathcal{F}_1^{12} \to \mathcal{F}_2^{12}$ and $g_2 : \mathcal{F}_2^{21} \to \mathcal{F}_1^{21}$ satisfying $g_1(s) - s \in [0, \Delta]$, $\forall s \in \mathcal{F}_1^{12}$ and $g_2(s) - s \in [0, \Delta]$, $\forall s \in \mathcal{F}_2^{21}$.

Table 2.1: Bidirectional-Bounded-Greedy-Match

BiBGM($\mathbf{s}_1$, $\mathbf{s}_2$, Δ):

```
1:     m = n = 1;
2:     while m ≤ |S₁| and n ≤ |S₂|
3:         if s₂(n) < s₁(m) − Δ
4:             s₂(n) is chaff;   n ← n + 1;
5:         else if s₂(n) > s₁(m) + Δ
6:             s₁(m) is chaff;   m ← m + 1;
7:         else
8:             match s₁(m) with s₂(n);   m ← m + 1;   n ← n + 1;
9:         end
10:    end
```

11:    return $\dfrac{|\{\text{Matched epochs}\}|}{|S_1| + |S_2|}$

implementation in Table 2.1:

1. At the beginning, all the epochs in $S_1 \cup S_2$ are unmatched. Start with the earliest epoch in $S_1 \cup S_2$, and go to **MATCH** to find its match.

2. **MATCH**: Let t denote the epoch for which we want to find a match. For $i = 1, 2$, if $t \in S_i$, search for the earliest unmatched epoch in $[t, t+\Delta] \cap S_{(3-i)}$ and match it with t; if there is no unmatched epoch in the interval, label t as chaff (an epoch is said to be *checked* if it is either matched with another epoch or labeled as chaff). Go to **MOVE**.

3. **MOVE**: If every epoch in $S_1 \cup S_2$ is checked, terminate. Otherwise, move to the next unchecked epoch in $S_1 \cup S_2$ and go to **MATCH** to find its match.

For the example in Fig. 2.3, BiBGM starts with t_1. Since $t_1 \in S_1$, we search for the earliest unmatched epoch in $[t_1, t_1 + \Delta] \cap S_2$, which is t_2. Hence, t_1 is matched with t_2. Then, we move to the next unchecked epoch, t_3 of S_1. Because t_2 is the only epoch in $[t_3, t_3+\Delta] \cap S_2$ and it is already matched with t_1, we label t_3 as chaff.
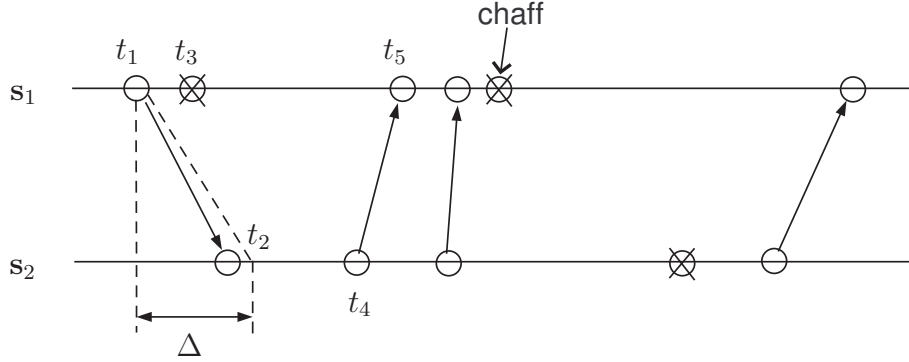
Figure 2.3: Bidirectional-Bounded-Greedy-Match

Next, we move to the next unchecked epoch (t_4 of $\mathcal{S}_2$) and searches for the earliest unmatched epoch in $[t_4, t_4 + \Delta] \cap \mathcal{S}_1$. BiBGM continues until the last epoch of $\mathcal{S}_1 \cup \mathcal{S}_2$ is checked.

From Table 2.1, it can be easily seen that BiBGM has linear computational complexity with respect to the sample size (*i.e.*, the total number of observed epochs). The following theorem states that BiBGM indeed achieves the optimal scheduling such that the flow part is maximized.

Theorem 2.3.1 *Suppose we run BiBGM on $(\mathbf{s}_i)_{i=1}^2$ in $[0, t]$. Then, the fraction of the matched epochs is equal to $\bar{R}(t)$.*

Proof: See Section 2.6. ∎

2.3.2   Parametric Flow Detection under Poisson Models

In this section, we assume the knowledge of the underlying parametric model for transmission processes and propose a detection algorithm called Bidirectional Flow

Detector (BFD). BFD is a threshold decision rule based on $\bar{R}(t)$. Specifically, BFD with a threshold τ takes the following form:

$$\begin{cases} \text{If } \bar{R}(t) \geq \tau, & \text{declare } \mathcal{H}_1; \\ \text{otherwise}, & \text{declare } \mathcal{H}_0. \end{cases}$$

If $(\mathbf{S}_i)_{i=1}^2$ contains a bidirectional flow, $\bar{R}(t)$ is, by definition, an upper bound on $R(t)$ and will tend to be greater compared to the case that $(\mathbf{S}_i)_{i=1}^2$ is an independent pair; this is the intuition behind declaring $\mathcal{H}_1$ when $\bar{R}(t)$ is greater than τ.

Under $\mathcal{H}_1$, since $\bar{R}(t) \geq R(t)$, BFD with τ can detect any flow with $R(t) \geq \tau$, and a smaller τ makes BFD capable of detecting a larger set of flows. However, a smaller τ results in a higher false alarm probability. Hence, there exists a trade-off between the detection ability of BFD and its false alarm probability, and we need to consult the parametric model for $(\mathbf{S}_i)_{i=1}^2$ under $\mathcal{H}_0$ to find out how small τ should be. Specifically, if under $\mathcal{H}_0$, as t increases $\bar{R}(t)$ converges to or stays close to a certain constant τ_0 with high probability, we can set τ slightly greater than τ_0 and make the false alarm probability become negligible as t grows. For homogeneous Poisson traffic, the following convergence result gives a guidance for setting τ.

Theorem 2.3.2 *Under $\mathcal{H}_0$, if $\mathbf{S}_1$ and $\mathbf{S}_2$ are homogeneous Poisson processes with rates λ_1 and λ_2 respectively, then as t grows to infinity, $\bar{R}(t)$ converges almost surely (a.s.) to*

$$\phi_{(\lambda_1,\lambda_2)} = \begin{cases} \dfrac{2\lambda_1\lambda_2(1 - e^{2\Delta(\lambda_1-\lambda_2)})}{(\lambda_1 + \lambda_2)(\lambda_2 - \lambda_1 e^{2\Delta(\lambda_1-\lambda_2)})} & \text{if } \lambda_1 \neq \lambda_2 \\[4mm] \dfrac{2\lambda\Delta}{1 + 2\lambda\Delta} & \text{if } \lambda_1 = \lambda_2 = \lambda. \end{cases}$$

Proof: See Section 2.6. ∎

25

Especially, if $(\mathbf{S}_i)_{i=1}^2$ under $\mathcal{H}_0$ and $(\mathbf{W}_i)_{i=1}^2$ under $\mathcal{H}_1$ are homogeneous Poisson processes, the following theorem states that any bidirectional flow with a positive rate is detectable regardless of the amount of chaff epochs.

Theorem 2.3.3 *Suppose that (i) under $\mathcal{H}_0$, $\mathbf{S}_1$ and $\mathbf{S}_2$ are homogeneous Poisson processes, (ii) under $\mathcal{H}_1$, $\mathbf{W}_1$ and $\mathbf{W}_2$ are homogeneous Poisson processes, and (iii) under both hypotheses, the rates[3] of $\mathbf{S}_1$ and $\mathbf{S}_2$ are λ_1 and λ_2, respectively. Then, for any $\eta \in (0, 1)$, there exists a proper threshold τ, such that BFD with τ can consistently detect[4] any bidirectional flow with $\mathrm{R} \geq \eta$ a.s., with the false alarm probability decaying exponentially fast as the sample size grows. Especially, for $\eta \in \left(0, \frac{2 \cdot \min\{\lambda_1, \lambda_2\}}{\lambda_1 + \lambda_2}\right)$, the following τ can be used:*

$$
\begin{cases}
\dfrac{2\lambda_1 - 2\lambda_2 \dfrac{\lambda_1(4-\eta) - \lambda_2\eta}{\lambda_2(4-\eta) - \lambda_1\eta} e^{2\Delta(\lambda_1 - \lambda_2)}}{(\lambda_2 + \lambda_1)\left(1 - \dfrac{\lambda_1(4-\eta) - \lambda_2\eta}{\lambda_2(4-\eta) - \lambda_1\eta} e^{2\Delta(\lambda_1 - \lambda_2)}\right)} & \text{if } \lambda_1 \neq \lambda_2, \\[4ex]
\dfrac{\eta + 2\lambda(2-\eta)\Delta}{2 + 2\lambda(2-\eta)\Delta} & \text{if } \lambda_1 = \lambda_2 = \lambda.
\end{cases}
$$

Proof: See Section 2.6. ■

It can be shown that the suggested τ in Theorem 2.3.3 is a strictly increasing function of η, and as η decreases to 0, it decreases to $\phi_{(\lambda_1, \lambda_2)}$ in Theorem 2.3.2. This means that to detect flows with smaller flow fraction, τ should be closer to the $\lim_{t \to \infty} \bar{\mathrm{R}}(t)$ value under $\mathcal{H}_0$.

Instead of the knowledge of the parametric model, training data can also be

---

[3]By rates, we mean that $\lim\limits_{t \to \infty} \dfrac{N_i(t)}{t} = \lambda_i$ a.s., where $N_i(t)$ denotes the number of epochs of $\mathbf{S}_i$ in $[0, t]$.

[4]In other words, if the distributions of $(\mathbf{S}_i)_{i=1}^2$ under $\mathcal{H}_0$ and $\mathcal{H}_1$ satisfy (i), (ii), (iii), and $\mathrm{R} \geq \eta$ a.s., then under all those distributions, the false alarm and miss detection probability vanish as t increases (as in Definition 2.2.3).

used to set τ. If a large set of different realizations of $\mathcal{H}_0$ traffic is available, we can run BiBGM over each realization in the training data set, estimate the statistical behavior of $\bar{\mathrm{R}}(t)$ under $\mathcal{H}_0$, and set τ such that the probability that $\bar{\mathrm{R}}(t) \geq \tau$ under $\mathcal{H}_0$ (*i.e.*, false alarm probability) becomes reasonably small as t grows. However, if neither a parametric model nor training data is available, it is non-trivial how to determine an appropriate τ; this is the case in many practical applications.

2.4 Nonparametric Flow Detection

In this section, we assume that neither a parametric model nor a training data set is available, and present a novel nonparametric flow detector.

2.4.1 Algorithm Structure

We begin by introducing the structure and the main intuition of our detection algorithm. Fig. 2.4 is describing its structure. A key component of our algorithm is a transformation of measurements $(\mathbf{s}_i)_{i=1}^2$, which we refer to as Independent Traffic Approximation (ITA). As the name suggests, ITA produces an approximately independent pair of transmission processes $(\bar{\mathbf{s}}_i)_{i=1}^2$ such that $\bar{\mathbf{s}}_i$ has similar traffic characteristics (*e.g.*, normalized intensity[5], interarrival distribution) with $\mathbf{s}_i$. After $(\bar{\mathbf{s}}_i)_{i=1}^2$ is generated, we compare the statistical characteristics of $(\mathbf{s}_i)_{i=1}^2$ and $(\bar{\mathbf{s}}_i)_{i=1}^2$. If the true hypothesis is $\mathcal{H}_0$, both $(\mathbf{s}_i)_{i=1}^2$ and $(\bar{\mathbf{s}}_i)_{i=1}^2$ are independent pairs with

---

[5]The normalized intensity of $\mathbf{S}_i$ for the $[0, t]$ interval represents the overall trend of its intensity change in $[0, t]$. Suppose that the local intensity of $\mathbf{S}_i$ is well defined in $[0, t]$: *i.e.*, $\lambda(x) \triangleq \lim_{\delta \to 0+} \dfrac{\mathbb{E}\{N_i[x, x + \delta)\}}{\delta}$ exists for all $x \in [0, t]$, where $N_i[a, b)$ denotes the number of $\mathbf{S}_i$ epochs in $[a, b)$. The normalized intensity $\bar{\lambda}^{(t)}$ of $\mathbf{S}_i$ for $[0, t]$ is defined as the time-scaled version of the intensity function: $\bar{\lambda}^{(t)}(x) \triangleq \lambda(tx), x \in [0, 1]$.
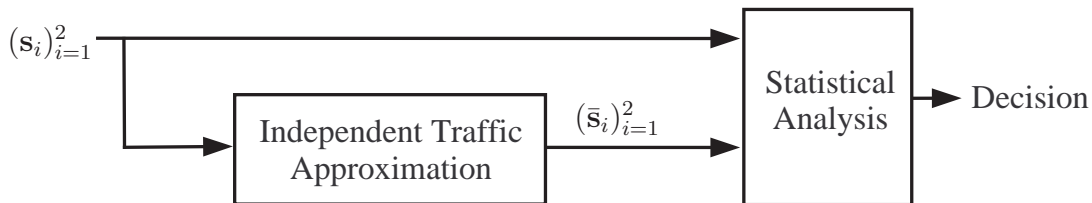
Figure 2.4: The structure of the nonparametric detection algorithm.

similar traffic characteristics. On the other hand, if $\mathcal{H}_1$ is true, $(\mathbf{s}_i)_{i=1}^2$ and $(\bar{\mathbf{s}}_i)_{i=1}^2$ have similar traffic characteristics, but $(\mathbf{s}_i)_{i=1}^2$ is a correlated pair containing a flow while $(\bar{\mathbf{s}}_i)_{i=1}^2$ approximates an independent pair. Thus, we attempt to infer the true hypothesis by exploiting the gap between the statistical characteristics of $(\mathbf{s}_i)_{i=1}^2$ and $(\bar{\mathbf{s}}_i)_{i=1}^2$: the larger the gap, the more probable $\mathcal{H}_1$ is.

2.4.2 Nonparametric Bidirectional Flow Detector

This section presents our detection algorithm, referred to as Nonparametric Bidirectional Flow Detector (NBFD). Here, we simply assume that ITA generates an output $(\bar{\mathbf{s}}_i)_{i=1}^2$ with desired properties: (i) $(\bar{\mathbf{s}}_i)_{i=1}^2$ approximates an independent pair of transmission processes, and (ii) its normalized intensities and interarrival distributions resemble that of $(\mathbf{s}_i)_{i=1}^2$. The detail about ITA is delayed to the next section, and here we focus on the operation of NBFD.

As described in Fig. 2.4, NBFD observes $(\mathbf{s}_i)_{i=1}^2$ in $[0, t]$ and first runs ITA to generate $(\bar{\mathbf{s}}_i)_{i=1}^2$. The next step is to compare the statistical characteristics of $(\mathbf{s}_i)_{i=1}^2$ and $(\bar{\mathbf{s}}_i)_{i=1}^2$. It was shown in Theorem 2.3.3, although stated under the homogeneous Poisson traffic assumption, that the maximum schedulable flow fraction $\bar{\mathrm{R}}(t)$ can be effectively used to distinguish whether the measurements are from a flow-containing pair or an independent pair. Moreover, $\bar{\mathrm{R}}(t)$ can be easily

28

evaluated by running BiBGM; hence, NBFD employs $\bar{\mathrm{R}}(t)$. NBFD runs BiBGM separately on $(\mathbf{s}_i)_{i=1}^2$ and $(\bar{\mathbf{s}}_i)_{i=1}^2$ and compares the fractions of the matched epochs in the two cases, denoted by $\bar{\mathrm{R}}(t)$ and $\bar{\tau}(t)$ respectively. If the true hypothesis is $\mathcal{H}_0$, both $(\mathbf{S}_i)_{i=1}^2$ and $(\bar{\mathbf{S}}_i)_{i=1}^2$ are independent pairs, and they have similar normalized intensities and interarrival distributions; this implies that $\bar{\mathrm{R}}(t)$ and $\bar{\tau}(t)$ are expected to be close under $\mathcal{H}_0$. On the other hand, when $\mathcal{H}_1$ is true, $(\mathbf{S}_i)_{i=1}^2$ and $(\bar{\mathbf{S}}_i)_{i=1}^2$ have similar normalized intensity functions and interarrival distributions, but $(\mathbf{S}_i)_{i=1}^2$ contains a flow while $(\bar{\mathbf{S}}_i)_{i=1}^2$ approximates an independent pair; hence, $\bar{\mathrm{R}}(t)$ is expected to be greater than $\bar{\tau}(t)$. Based on the above intuition, given $(\mathbf{s}_i)_{i=1}^2$ in $[0, t]$, NBFD with ϵ works as follows:

1. Run ITA on $(\mathbf{s}_i)_{i=1}^2$ in $[0, t]$ to generate $(\bar{\mathbf{s}}_i)_{i=1}^2$.

2. Run BiBGM on $(\mathbf{s}_i)_{i=1}^2$ and $(\bar{\mathbf{s}}_i)_{i=1}^2$: $\bar{\mathrm{R}}(t)$ and $\bar{\tau}(t)$ denote the fractions of the matched epochs for $(\mathbf{s}_i)_{i=1}^2$ and $(\bar{\mathbf{s}}_i)_{i=1}^2$ respectively.

3. If $\bar{\mathrm{R}}(t) \geq \bar{\tau}(t) + \epsilon$, declare $\mathcal{H}_1$; otherwise, declare $\mathcal{H}_0$.

where ϵ is a positive number added to $\bar{\tau}(t)$ to allow small difference between $\bar{\mathrm{R}}(t)$ and $\bar{\tau}(t)$ under $\mathcal{H}_0$. $\bar{\tau}(t)$ can also be seen as an estimate of what $\bar{\mathrm{R}}(t)$ would be under $\mathcal{H}_0$. Therefore, recalling the discussion of setting τ of BFD in Section 2.3.2, NBFD can be alternatively interpreted as BFD with a measurement-dependent threshold $\bar{\tau}(t) + \epsilon$.

It is evident from the form of NBFD that a smaller ϵ will lead to the decrease in the miss detection probability. However, the decrease in ϵ will increase the false alarm probability. Because of the trade-off associated with the choice of ϵ and the nonparametric characteristic of our problem, it is difficult to claim that certain
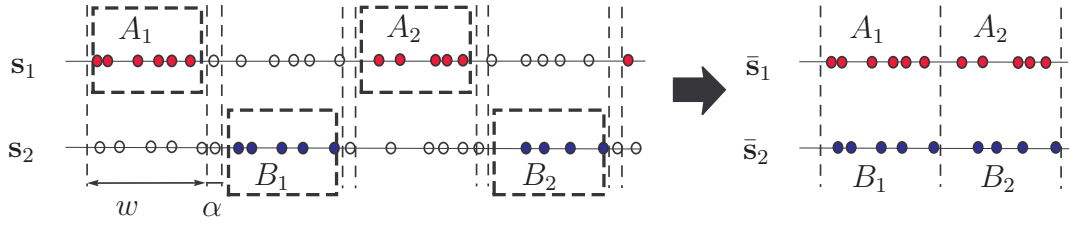
Figure 2.5: ITA samples w-second intervals $\{A_1, A_2, \ldots\}$ and $\{B_1, B_2, \ldots\}$ from $(\mathbf{s}_i)_{i=1}^2$ and assemble them to generate $(\bar{\mathbf{s}}_i)_{i=1}^2$.

ϵ value is the best choice. The experimental results in Section 2.5 suggest that setting $\epsilon \approx 0.05$ generally results in satisfactory performance.

2.4.3 Independent Traffic Approximation

In this section, we present how ITA approximates an independent pair of transmission processes that has the similar normalized intensity and interarrival distribution with $(\mathbf{S}_i)_{i=1}^2$.

Fig. 2.5 is illustrating the operation of ITA. ITA has two parameters: the sampling window width w and the gap α ($\alpha \geq \Delta$) between neighboring sampling windows. As described in Fig. 2.5, ITA samples the epochs in the w-second windows separated by α-second gaps, shifts them properly, and assembles them to approximate independent traffic. The intuition behind ITA is that if the gap α between two sampling windows is sufficiently large, the epochs in different windows will tend to be approximately uncorrelated. Note that when $(\mathbf{S}_i)_{i=1}^2$ contains a bidirectional flow, ITA disassembles the flow part and significantly reduces the flow-induced correlation. In addition, since we use a sequence of sampled intervals of $\mathbf{S}_i$ for generating $\bar{\mathbf{S}}_i$, $\bar{\mathbf{S}}_i$ and $\mathbf{S}_i$ are expected to share some common characteristics.

30
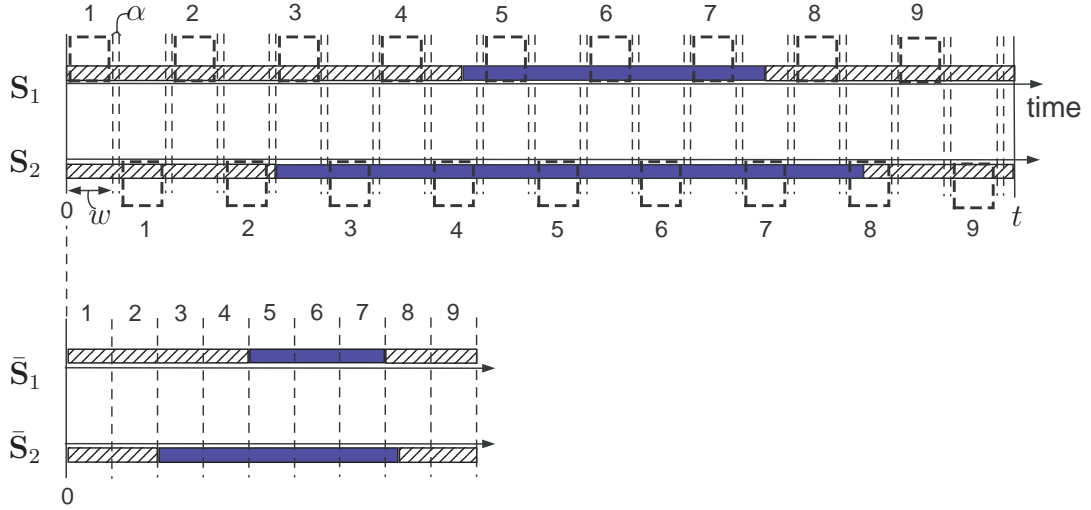
Figure 2.6: $\mathbf{S}_1$ and $\mathbf{S}_2$ are non-homogeneous Poisson processes, and $\lambda_1(x)$ and $\lambda_2(x)$ denote their local intensities at time x respectively. $\lambda_1(x)$ and $\lambda_2(x)$ can only take values from $\{\mu_1, \mu_2\}$ $(\mu_1 \neq \mu_2)$. The figure describes the intensity change of $\mathbf{S}_1$, $\mathbf{S}_2$, $\bar{\mathbf{S}}_1$, and $\bar{\mathbf{S}}_2$ using two types of bars. The bars filled with slant lines represent the intervals in which $\lambda_i(x) = \mu_1$, and the blue bars represent the intervals in which $\lambda_i(x) = \mu_2$. The numbers above or below the intervals describe the correspondence between the sampled intervals in $\mathbf{S}_i$ and the intervals in $\bar{\mathbf{S}}_i$.

To illustrate how the normalized intensities of $\mathbf{S}_i$ and $\bar{\mathbf{S}}_i$ are related, Fig. 2.6 describes the intensity change of $(\mathbf{S}_i)_{i=1}^2$ and $(\bar{\mathbf{S}}_i)_{i=1}^2$ for an example where $\mathbf{S}_1$ and $\mathbf{S}_2$ are non-homogeneous Poisson processes with two possible intensity levels. As observed in Fig. 2.6, if the average time that the intensity of $\mathbf{S}_i$ stays in one level is much longer than $2(w + \alpha)$ seconds, the normalized intensity function of $\bar{\mathbf{S}}_i$ is similar to that of $\mathbf{S}_i$. About interarrival distribution, if w is sufficiently large so that a w-second sampling window is likely to contain a large number of points, the interarrival distribution of $\bar{\mathbf{S}}_i$ will resemble that of $\mathbf{S}_i$. Moreover, if the interarrival distribution of $\mathbf{S}_i$ varies slowly over time, as in the example of Fig. 2.6, the interarrival distribution of $\bar{\mathbf{S}}_i$ will also change over time with the similar trend, even though the time scale is different due to the sampling procedure of ITA. Note that resampling from the empirical interarrival distributions (*i.e.*, generating i.i.d. in-

terarrival times of $\bar{\mathbf{S}}_i$ from the empirical interarrival distribution of $\mathbf{S}_i$, for $i = 1, 2$) can also produce an independent pair of point processes. However, unlike $(\bar{\mathbf{S}}_i)_{i=1}^2$ of ITA, when $(\mathbf{S}_i)_{i=1}^2$ is non-stationary, the results of such resampling approaches may have a totally different dynamics from $(\mathbf{S}_i)_{i=1}^2$; they may not capture the patterns of intensity change or interarrival distribution change in $(\mathbf{S}_i)_{i=1}^2$.

Now, we will check whether $(\bar{\mathbf{S}}_i)_{i=1}^2$ can approximate an independent pair. When $\mathbf{S}_1$ and $\mathbf{S}_2$ are independent, it directly follows that $\bar{\mathbf{S}}_1$ and $\bar{\mathbf{S}}_2$ are independent. On the other hand, if $(\mathbf{S}_i)_{i=1}^2$ contains a bidirectional flow, $\bar{\mathbf{S}}_1$ and $\bar{\mathbf{S}}_2$ are not necessarily independent. However, assuming that correlation across time is weak and the gap α is much larger than Δ, the epochs in different windows are expected to be approximately uncorrelated: *i.e.*, in Fig. 2.5, the epochs in each A_i will be approximately uncorrelated with the epochs in $\bigcup_{j \geq 1} B_j$. This implies that when temporal correlation is weak, $(\bar{\mathbf{S}}_1, \bar{\mathbf{S}}_2)$ is expected to approximate an independent pair. The following example illustrates a case where $(\mathbf{S}_i)_{i=1}^2$ has weak temporal correlation. Suppose $\mathbf{S}_1$ is a Poisson process and $\mathbf{S}_2$ is such that $S_2(i) = S_1(i) + D_i, \forall i$, where D_is are independent random delays bounded by Δ a.s.: *i.e.*, $(\mathbf{S}_i)_{i=1}^2$ is a unidirectional flow with a delay constraint Δ. The memoryless property of Poisson processes implies that epochs in an interval are correlated with epochs in another disjoint interval only if the gap between the two intervals is less than Δ seconds. Hence, if $\alpha \geq \Delta$, epochs in different sampling windows of ITA are independent, implying that $(\bar{\mathbf{S}}_i)_{i=1}^2$ is an independent pair.

Under $\mathcal{H}_0$, NBFD requires $(\bar{\mathbf{S}}_i)_{i=1}^2$ to be an independent pair having the similar traffic characteristics with $(\mathbf{S}_i)_{i=1}^2$, because $\bar{\mathrm{R}}(t)$ and $\bar{\tau}(t)$ have to be close under $\mathcal{H}_0$. However, under $\mathcal{H}_1$, NBFD does not necessitate the independence of $\bar{\mathbf{S}}_1$ and $\bar{\mathbf{S}}_2$, even though the independent case is ideal. Under $\mathcal{H}_1$, NBFD wants $\bar{\tau}(t)$ to

be less than $\bar{\mathrm{R}}(t)$, and this can be achieved by making $(\bar{\mathbf{S}}_i)_{i=1}^2$ very unlikely to contain a flow. Because, as can be inferred from the discussion in Section 2.3.2, the maximum schedulable flow fraction (*e.g.*, $\bar{\mathrm{R}}(t)$ and $\bar{\tau}(t)$ of NBFD) tends to be higher when the measurements come from a flow-containing pair. Note that ITA does make $(\bar{\mathbf{S}}_i)_{i=1}^2$ unlikely to contain a flow by tearing apart the flow part of $(\mathbf{S}_i)_{i=1}^2$ in its sampling procedure.

Given the measurements $(\mathbf{s}_i)_{i=1}^2$ in $[0,\,t]$, ITA with $(w,\,\alpha)$ generates $(\bar{\mathbf{s}}_i)_{i=1}^2$ as follows:

1. Initially, $\bar{\mathbf{s}}_1$ and $\bar{\mathbf{s}}_2$ contain no epoch.

2. For $i = 0,\,1,\,\ldots,\,\lfloor\frac{t}{2(w+\alpha)}\rfloor - 1$:

    (a) Take the epochs of $\mathbf{s}_1$ in $[2i(w+\alpha),\,2i(w+\alpha)+w]$, subtract $i(w+2\alpha)$ from the epochs, and add them to $\bar{\mathbf{s}}_1$.

    (b) Take the epochs of $\mathbf{s}_2$ in $[(2i+1)(w+\alpha),\,(2i+1)(w+\alpha)+w]$, subtract $i(w+2\alpha)+(w+\alpha)$ from the epochs, and add them to $\bar{\mathbf{s}}_2$.

The implementation of ITA is given in Table 2.2. As can be seen from Table 2.2, ITA has linear computational complexity with respect to the sample size.

One drawback of ITA is that it throws away more than a half of the measurements during the sampling procedure, thereby restricting the sample size of $(\bar{\mathbf{s}}_i)_{i=1}^2$ to be at most a half of that of $(\mathbf{s}_i)_{i=1}^2$. $(\bar{\mathbf{s}}_i)_{i=1}^2$, together with $(\mathbf{s}_i)_{i=1}^2$, is used to calculate the decision statistic of NBFD, so a large sample size is desirable. Therefore, we suggest a modification of ITA, referred to as ITA-double (ITAd), to double the sample size of $(\bar{\mathbf{s}}_i)_{i=1}^2$.

33

Table 2.2: Independent Traffic Approximation

ITA($\mathbf{s}_1$, $\mathbf{s}_2$, t, w, α):

```
1:    s̄₁ ← ();  s̄₂ ← ();  a₁ ← ();  a₂ ← ();  j = 1;  k = 1;
2:    for i = 0 : 1 : ⌊ t/(2(w+α)) ⌋ − 1
3:        while s₁(j) < 2i(w + α)
4:            j ← j + 1;
5:        end
6:        while s₁(j) ≤ 2i(w + α) + w
7:            a₁ ← a₁ ⊕ s₁(j);   j ← j + 1;
8:        end
9:        while s₂(k) < (2i + 1)(w + α)
10:           k ← k + 1;
11:       end
12:       while s₂(k) ≤ (2i + 1)(w + α) + w
13:           a₂ ← a₂ ⊕ s₂(k);   k ← k + 1;
14:       end
15:       a₁ ← a₁ − i(w + 2α);   s̄₁ ← s̄₁ ⊕ a₁;
16:       a₂ ← a₂ − (i(w + 2α) + w + α);   s̄₂ ← s̄₂ ⊕ a₂;
17:       a₁ ← ();   a₂ ← ();
18:   end
19:   return (s̄ᵢ)²ᵢ₌₁.
```

* For a sequence $(x_i)_{i \geq 1}$ and a real number r, $(x_i)_{i \geq 1} - r \triangleq (y_i)_{i \geq 1}$ where $y_i = x_i - r$, $\forall i$.

Figure 2.7: The sample size of $(\bar{\mathbf{s}}_i)_{i=1}^2$ doubles compared to ITA. Unlike ITA, ITAd does not throw away $\{A_2, A_4, \ldots\}$ or $\{B_2, B_4, \ldots\}$; it assembles all of $\{A_1, A_2, \ldots\}$ and $\{B_1, B_2, \ldots\}$ to generate $(\bar{\mathbf{s}}_i)_{i=1}^2$.

The operation of ITAd is illustrated in Fig. 2.7. In ITAd, when $\mathbf{S}_1$ and $\mathbf{S}_2$ are independent, so are $\bar{\mathbf{S}}_1$ and $\bar{\mathbf{S}}_2$. However, if $(\mathbf{S}_i)_{i=1}^2$ contains a flow, $(\bar{\mathbf{S}}_i)_{i=1}^2$ is not an independent pair, because the epochs in A_{i+1} and those in B_i are correlated due to the presence of the flow. However, $(\bar{\mathbf{S}}_i)_{i=1}^2$ is a concatenation of w-second

intervals, where, in each interval, the epochs of $\bar{\mathbf{S}}_1$ and $\bar{\mathbf{S}}_2$ are approximately uncorrelated. We believe that this property is enough for NBFD to sense the difference in statistical characteristics between $(\mathbf{S}_i)_{i=1}^2$ and $(\bar{\mathbf{S}}_i)_{i=1}^2$ under $\mathcal{H}_1$, especially when w is large. Although we have no analytical proof for the superiority of ITAd over ITA, the use of ITAd in NBFD instead of ITA consistently resulted in a better performance in all our simulations and experiments in Section 2.5.

2.4.4 Performance Analysis

This section provides the analysis of algorithmic efficiency and consistency of NBFD.

NBFD is efficient in terms of computation and memory requirement. Because its main components, ITA and BiBGM, have linear complexity, NBFD also has linear computational complexity with respect to the sample size. In addition, assuming that NBFD with (w, α, ϵ) is executed in real-time over transmission processes of two nodes, it only requires to save the most recent BiBGM matches of $(\mathbf{s}_i)_{i=1}^2$ and $(\bar{\mathbf{s}}_i)_{i=1}^2$ and the timing measurements in the most recent $2(w+\alpha)$-second interval; they are all the information needed to continue running ITA and BiBGM over the future timing measurements.

For a class of non-homogeneous Poisson traffic, NBFD has a consistency property as stated in the following theorem.

Theorem 2.4.1 *Assume that w and α are any positive numbers with $\alpha \geq \Delta$. For any $\eta \in (0, 1)$, there exists an $\bar{\epsilon} \in (0, 1)$ such that, for any $\epsilon \in (0, \bar{\epsilon}]$, NBFD with (w, α, ϵ) consistently detects any bidirectional flow with $\mathrm{R} \geq \eta$ a.s., if the*

distributions of $(\mathbf{S}_i)_{i=1}^2$ under $\mathcal{H}_0$ and $\mathcal{H}_1$ satisfy the following assumptions[6]:

1. *Under both hypotheses, $\mathbf{S}_1$ and $\mathbf{S}_2$ are non-homogeneous Poisson processes. In addition, under $\mathcal{H}_1$, $\mathbf{S}_i = (\mathbf{F}_i^{12} \oplus \mathbf{F}_i^{21}) \oplus \mathbf{W}_i$. $\mathbf{F}_1^{12}$, $\mathbf{F}_2^{21}$, $\mathbf{W}_1$, and $\mathbf{W}_2$ are independent non-homogeneous Poisson processes, $\mathbf{F}_2^{12}$ is[7] $\mathbf{sort}\{\mathbf{F}_1^{12}(i) + \alpha_i, i \geq 1\}$, and $\mathbf{F}_1^{21}$ is $\mathbf{sort}\{\mathbf{F}_2^{21}(i) + \beta_i, i \geq 1\}$ where $\{\alpha_i, i \geq 1\}$ and $\{\beta_i, i \geq 1\}$ are random variables satisfying $\alpha_i, \beta_i \in [0, \Delta]$ almost surely. Furthermore, $\{\alpha_i, i \geq 1\} \perp\!\!\!\perp \mathbf{W}_1$, $\{\beta_i, i \geq 1\} \perp\!\!\!\perp \mathbf{W}_2$, and[8] $\perp\!\!\!\perp \{\alpha_i, i \geq 1\}$, $\{\beta_i, i \geq 1\}$, $\mathbf{F}_1^{12}$, $\mathbf{F}_2^{21}$.*

2. *Let $\lambda_1(t)$, $\lambda_2(t)$, $\lambda_{f1}(t)$, and $\lambda_{f2}(t)$ denote the local intensities of $\mathbf{S}_1$, $\mathbf{S}_2$, $\mathbf{F}_1^{12}$, and $\mathbf{F}_2^{21}$ respectively. There exist two finite sets $\Lambda_0 \triangleq \{\vec{\mu}^{(j)} \triangleq (\mu_1^{(j)}, \mu_2^{(j)}), 1 \leq j \leq M_0\}$ and $\Lambda_1 \triangleq \{\vec{\lambda}^{(k)} \triangleq (\lambda_1^{(k)}, \lambda_2^{(k)}, \lambda_{f1}^{(k)}, \lambda_{f2}^{(k)}), 1 \leq k \leq M_1\}$ with $\mu_i^{(j)} > 0$, $\lambda_i^{(k)} > 0$, $i = 1, 2$, $\forall j, k$. Under $\mathcal{H}_0$, $(\lambda_1(t), \lambda_2(t))$ can only take values in Λ_0. Under $\mathcal{H}_1$, $\vec{\lambda}(t) \triangleq (\lambda_1(t), \lambda_2(t), \lambda_{f1}(t), \lambda_{f2}(t))$ can only take values in Λ_1.*

3. *Under $\mathcal{H}_0$, if $c(t)$ denotes the number of times that $(\lambda_1(t), \lambda_2(t))$ changes its value in $[0, t]$, then $\lim_{t \to \infty} \frac{c(t)}{t} = 0$. Similarly, under $\mathcal{H}_1$, if $c(t)$ denotes the number of times that $\vec{\lambda}(t)$ changes its value in $[0, t]$, then $\lim_{t \to \infty} \frac{c(t)}{t} = 0$.*

4. *Under $\mathcal{H}_0$, if $\rho_k(t)$ $(1 \leq k \leq M_0)$ denotes the fraction of the time in $[0, t]$ that $(\lambda_1(t), \lambda_2(t)) = \vec{\mu}^{(k)}$, then as t increases, each $\rho_k(t)$ converges. Similarly, under $\mathcal{H}_1$, if $\rho_k(t)$ $(1 \leq k \leq M_1)$ denotes the fraction of the time in $[0, t]$ that $\vec{\lambda}(t) = \vec{\lambda}^{(k)}$, then as t increases, each $\rho_k(t)$ converges.*

---

[6]In other words, if the distributions of $(\mathbf{S}_i)_{i=1}^2$ under $\mathcal{H}_0$ and $\mathcal{H}_1$ satisfy the listed assumptions (including $R \geq \eta$ a.s. under $\mathcal{H}_1$), then under all those distributions, the false alarm and miss detection probabilities of NBFD with (w, α, ϵ) vanish as t grows (as in Definition 2.2.3).

[7]For a countable set A of real numbers, $\mathbf{sort}\{A\}$ is the sequence of the elements of A ordered in the increasing order.

[8]For random processes $\mathbf{A}_i$s, $\mathbf{A}_1 \perp\!\!\!\perp \mathbf{A}_2$ means $\mathbf{A}_1$ and $\mathbf{A}_2$ are independent, and $\perp\!\!\!\perp \mathbf{A}_1, \ldots, \mathbf{A}_n$ means $\mathbf{A}_1, \ldots, \mathbf{A}_n$ are independent.

Proof: See Section 2.6. ∎

The first assumption means that under $\mathcal{H}_1$, $(\mathbf{S}_i)_{i=1}^2$ is a superposition of three independent parts: the unidirectional flow from $\mathbf{S}_1$ to $\mathbf{S}_2$, the unidirectional flow from $\mathbf{S}_2$ to $\mathbf{S}_1$, and the chaff parts. α_i and β_i represent packet delays of the two unidirectional flows, and they satisfy certain independence relationships involving the flow parts and the chaff parts. The first assumption is sufficient to guarantee that the output of ITA, $(\bar{\mathbf{S}}_i)_{i=1}^2$, is an independent pair under $\mathcal{H}_1$. The second assumption implies that the local intensities of the total traffic and flows can only take a finite number of different values. The third assumption says that the number of intensity changes in $[0, t]$ grows as $o(t)$. Finally, the last assumption means that the fraction of the time that the intensity vector assumes a specific value converges as the observation time increases. Under these assumptions, Theorem 2.4.1 states that a bidirectional flow with any positive rate can be consistently detected by NBFD if ϵ is properly set. Note that the assumptions do not restrict traffic to be stationary.

As pointed out by Paxson and Floyd [17], a Poisson process is not always a good model for network arrival processes. Several network traces (*e.g.*, Ethernet and World Wide Web traffic) have been experimentally proved to display self-similarity [69–71], which Poisson processes do not show. To test the performance of NBFD over non-Poisson traffic, we will evaluate NBFD in the following section using LBL TCP traces, which were used in [17] to invalidate Poisson modeling, and real-world measurements from MSN VoIP sessions.

2.5 Numerical Results

NBFD was tested using the synthetic Poisson traffic, LBL TCP traces, and the real-world measurements from MSN VoIP sessions. Comparison with other passive flow detectors is also provided: the wavelet analysis in [1], Detect-Attack-Chaff (DAC) in [13], and the random projection method in [16].

The wavelet analysis [1] calculates the wavelet coefficients of $N_1(t)$ and $N_2(t)$ using the mother Haar wavelet with a sufficiently large scale, where $N_i(t)$ is the number of epochs of $\mathbf{S}_i$ in $[0, t]$. Then, it calculates the Peason's correlation coefficient between the wavelet coefficients of $N_1(t)$ and that of $N_2(t)$, and declares $\mathcal{H}_1$ if the correlation coefficient is greater than a predetermined threshold κ; otherwise, it declares $\mathcal{H}_0$. The intuition of the algorithm is based on their analysis under the Poisson traffic assumption: the correlation coefficient converges to a positive constant as the scale[9] grows to infinity if $(\mathbf{S}_i)_{i=1}^2$ contains a flow.

DAC [13] is based on the intuition that as t increases $|N_1(t) - N_2(t)|$ tends to grow large when $\mathbf{S}_1$ and $\mathbf{S}_2$ are independent, whereas it tends to stay small if $(\mathbf{S}_i)_{i=1}^2$ contains a flow with a much higher rate than the chaff part. DAC with a parameter[10] p_Δ monitors $|N_1(t) - N_2(t)|$. At every $8(p_\Delta + 1)^2$ packet transmissions, both N_1 and N_2 are set to be zero and new counting begins. It declares $\mathcal{H}_0$ if $|N_1(t) - N_2(t)|$ grows larger than a threshold $2p_\Delta$. If $|N_1(t) - N_2(t)|$ stays less than $2p_\Delta$ during the whole observation duration, DAC declares $\mathcal{H}_1$. Note that

---

[9]Since the wavelet analysis relies on the convergence of the correlation coefficient as the scale grows, a large scale is desired. However, given a fixed observation duration, using too large scale can cause the sample size of the correlation coefficient estimation (*i.e.*, the number of wavelet coefficients) to be very small. To prevent this, in our experiments, the sample size is fixed to be 100, and the scale is set to be (the observation duration)/100.

[10]In [13], p_Δ is defined to be a uniform upper bound on the number of epochs of a node ($\mathbf{S}_1$ or $\mathbf{S}_2$) in any Δ-sec interval. However, none of our test traces guarantees such a uniform upper bound. Hence, we tried DAC with various p_Δ values, which include large enough numbers to bound the number of epochs in any Δ-sec interval with high probability.

under $\mathcal{H}_1$, if bursty chaff transmissions occur in either node, $|N_1(t) - N_2(t)|$ may suddenly grow larger than $2p_\Delta$ thereby resulting in a miss detection. Hence, DAC is vulnerable to bursty chaff insertion.

The random projection method in [16], which we denote by RP, is based on the idea of measuring the distance between $\mathbf{S}_1$ and $\mathbf{S}_2$ after random projection. It first partitions the observation interval into the time slots with length L_{TS}, and counts the number of epochs in each time slot. The number of epochs of $\mathbf{S}_i$ in the jth time slot is denoted by $V_i(j)$, $i = 1, 2, 1 \le j \le T$. Then, RP generates a set of K random basis vectors $\{B_k \in \{-1, 1\}^T, 1 \le k \le K\}$, where each $B_k(j)$ $(1 \le j \le T)$ is either 1 or -1 with an equal probability[11]. After that, V_i is projected on $\{B_k, 1 \le k \le K\}$: $C_i(k) \triangleq \sum_j V_i(j)B_k(j)$, $i = 1, 2, 1 \le k \le K$. Finally, RP obtains a K-dimensional binary vector $\bar{C}_i$, where $\bar{C}_i(k) \triangleq 1_{\{C_i(k)>0\}}$, referred to as the *binary sketch* of $\mathbf{S}_i$. The decision statistic of RP is the Hamming distance between $\bar{C}_1$ and $\bar{C}_2$. If the distance is less than a threshold th, RP declares $\mathcal{H}_1$; otherwise, $\mathcal{H}_0$ is declared.

2.5.1 Simulation Results: Poisson Traffic and LBL traces

We first performed Monte Carlo simulations using the synthetic non-homogeneous Poisson traffic. In the simulations, $\mathbf{S}_1$ and $\mathbf{S}_2$ are Poisson processes with intensity functions $\lambda_1(t)$ and $\lambda_2(t)$ respectively. Under $\mathcal{H}_1$, $(\mathbf{S}_i)_{i=1}^2$ is a superposition of two independent parts, the unidirectional flow $(\mathbf{F}_i)_{i=1}^2$ and the chaff part $(\mathbf{W}_i)_{i=1}^2$. $\mathbf{F}_1$ is a Poisson process with intensity function $\lambda_f(t)$, and $\mathbf{F}_2$ is generated by adding a random delay to each epoch of $\mathbf{F}_1$. Random delays are independent and identically

---

[11]About the parameters of RP, we used $L_{TS} = 0.5s$, as recommended in [16]. As explained in [16], large K is desired since it will allow us to extract more information from $\mathbf{S}_i$. We used $K = 4096$, which we believe is sufficiently large (four times the maximum K used in [16]).

distributed (i.i.d.) and uniformly distributed in $[0, \Delta]$, where $\Delta = 0.1s$. $\mathbf{W}_1$ and $\mathbf{W}_2$ are independent Poisson processes with intensity functions $\lambda_1(t) - \lambda_f(t)$ and $\lambda_2(t) - \lambda_f(t)$, respectively. In each run of the simulation, $(\lambda_1(t), \lambda_2(t), \lambda_f(t))$ is piecewise constant, and it takes different values in the first third, the second third, and the last third of the observation duration. Specifically, it follows one of the below change scenarios with equal probability:

1. $(15, 15, 5) \rightarrow (15, 15, 12) \rightarrow (15, 15, 7)$.

2. $(25, 10, 8) \rightarrow (10, 10, 8) \rightarrow (10, 25, 8)$.

3. $(25, 25, 20) \rightarrow (12, 12, 7) \rightarrow (8, 8, 3)$.

4. $(21, 15, 14) \rightarrow (12, 6, 5) \rightarrow (12, 24, 5)$.

Under $\mathcal{H}_0$, $\mathbf{S}_1$ and $\mathbf{S}_2$ are independent Poisson processes, and in each run of the simulation, $(\lambda_1(t), \lambda_2(t))$ follows one of the above change scenarios (with no λ_f part) with equal probability. In real world, such changes in intensity may correspond to the beginning of new sessions, the end of old sessions, the rate change of existing sessions, and so on. All change scenarios have the same average rates, but each scenario displays a different dynamics. By this simulation setting, we aimed at testing the performance of detectors over the non-stationary traffic displaying possibly a different dynamics at each observation interval.

Fig. 2.8 shows the ROC curves of NBFD (with ITAd), NBFD (with ITA), the wavelet analysis, DAC, and RP. To obtain the ROC curves, we increased ϵ of NBFD and κ of the wavelet analysis from 0 to 1 with an increment of 0.01, p_Δ of DAC from 4 to 100 with an increment of 2, and th of RP from 0 to K ($K = 4096$) with an increment of 1 while plotting $(P_F, 1 - P_M)$ of each case, where P_F and P_M denote the false alarm probability and miss detection probability respectively.
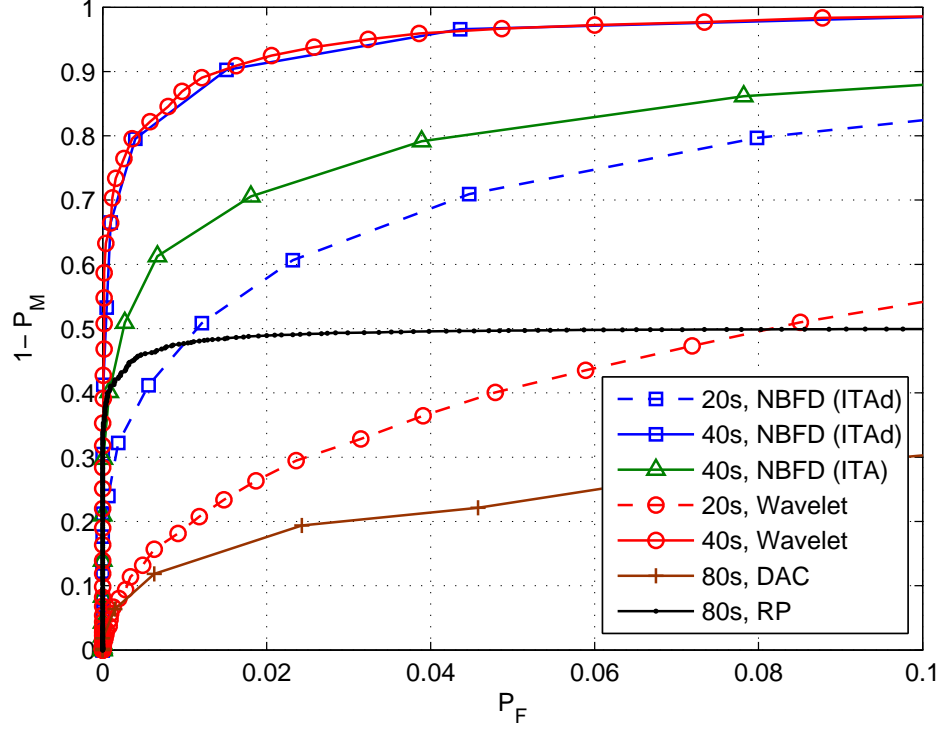
Figure 2.8: ROC curves of NBFD (ITAd), NBFD (ITA), the wavelet analysis, DAC, and RP for different observation durations: NBFD parameters are $w = 2s$ and $\alpha = \Delta = 0.1s$, and the number of Monte Carlo runs is 10000.

When we further increased the sample size, the ROC curves of NBFD (ITAd), NBFD (ITA), and the wavelet analysis approached the upper left corner implying that perfect detection is possible if the thresholds are properly set. On the other hand, DAC and RP resulted in non-negligible error probabilities in every case, and their ROC curves did not improve much from the curves in Fig. 2.8, even when we further increased the observation duration to 160s. By comparing the ROC curves of NBFD (ITAd) and NBFD (ITA), we can observe that ITAd, the heuristic to double the sample size of $(\bar{\mathbf{s}}_i)_{i=1}^2$, resulted in a better detection performance than ITA. In all our simulations and experiments, ITAd consistently showed better results than ITA. In the rest of this section, NBFD is assumed to employ ITAd and will be compared with other detectors.

To test the performance of detectors over non-Poisson traffic, we generated synthetic traffic based on the TCP packet timestamps in LBL-PKT-3 (2 hours), LBL-PKT-4 (1 hour), and LBL-PKT-5 (1 hour) in [17]. These traces were measured at the Lawrence Berkeley Laboratory's wide-area Internet gateway, and each trace was gathered at a different date in January 1994. For the detail, refer to [17]. From each dataset, we extracted timestamps of TCP packets that originated from specific users, and used them for traffic generation. For the flow part of $\mathcal{H}_1$ traffic, timestamps of one user in LBL-PKT-3 were used as $\mathbf{F}_1$, and $\mathbf{F}_2$ was generated by adding a delay to each epoch in $\mathbf{F}_1$. The delays are i.i.d. and uniformly distributed in $[0, \Delta]$, where $\Delta = 0.1s$. For the chaff part, timestamps of one user in LBL-PKT-4 were used as $\mathbf{W}_1$, and those of one user in LBL-PKT-5 were used as $\mathbf{W}_2$. For $\mathcal{H}_0$ traffic, $\mathbf{S}_1$ is generated by superposing traces of two users in LBL-PKT-4, and $\mathbf{S}_2$ is similarly generated with two users in LBL-PKT-5. Using different sets of users for the traffic generation, we were able to create the four-hour long test traffic.

We tested DAC with various p_Δ ranging from 10 to 400, but its miss detection probability was higher than 0.38 in every case. This is not surprising because DAC is vulnerable to bursty chaff transmissions and LBL TCP traces were shown to be bursty in [17]. Table 2.3 shows the error probabilities of NBFD, the wavelet analysis, and RP. For NBFD, we used $\epsilon = 0.05$. For the wavelet analysis and RP, assuming the absence of a parametric model and training data, we have no clear standard to set their thresholds. Hence, we tried all values from 0 to 1 with an increment of 0.01 for κ of the wavelet analysis and all values from 0 to 4096 with an increment of 1 for th of RP, and found their crossover error rates and the corresponding thresholds, which are listed in Table 2.3. NBFD and the wavelet analysis outperformed RP, and for long observation durations (160s and 320s), NBFD performed better than the wavelet analysis.

Table 2.3: Performance on LBL TCP traces. NBFD parameters are $w = 2s$, $\alpha = \Delta = 0.1s$, and $\epsilon = 0.05$. The numbers of experiments are 180, 90, and 45 for observation duration 80s, 160s, and 320s, respectively. Under $\mathcal{H}_0$, the average traffic rate is $(\lambda_1, \lambda_2) = (36.4, 36.1)$. Under $\mathcal{H}_1$, $(\lambda_1, \lambda_2) = (36.1, 36.8)$. The fraction of chaff in $\mathcal{H}_1$ traffic is 0.37.

	NBFD		Wavelet			RP		
Time	P_F	P_M	κ	P_F	P_M	th	P_F	P_M
80s	0	0.100	0.19	0.034	0.056	762	0.101	0.144
160s	0	0.057	0.20	0.034	0.056	793	0.112	0.133
320s	0	0.022	0.19	0.023	0.067	774	0	0.089

2.5.2 Experimental Results: MSN VoIP Traffic

We tested the detectors using three-and-a-half-hour long real-world traffic involving the MSN VoIP application[12], which is a representative example of latency-sensitive applications. Fig. 2.9 is illustrating the experimental setup. The laptop P_1 is located in the place covered by the wireless access point A_1, and two other laptops, P_2 and P_3, are located in the different place covered by the wireless access point A_2, which is controlled to serve only P_2 and P_3. Suppose it is known that P_1 is engaged in a VoIP conversation. By measuring the wireless transmission epochs of P_1 and A_2, our objective is to detect whether P_1 is having a VoIP conversation with any device served by the access point A_2. In practice, there may be additional information available: packet sizes, protocol types (TCP or UDP), destination addresses, and so on. However, here we assume that we have no access to such information due to encryption or other countermeasures employed by the network administrator, and only the timing measurements are available.

Let $\mathbf{S}_1$ and $\mathbf{S}_2$ denote the transmission processes of P_1 and A_2 respectively.

---

[12]Windows Live Messenger 2009 (14.0.8089.726) was used for MSN VoIP calls, and Wireshark network protocol analyzer (ver 1.2.6.) with the AirPcap classic adaptor was used to record the timings of wireless transmissions.
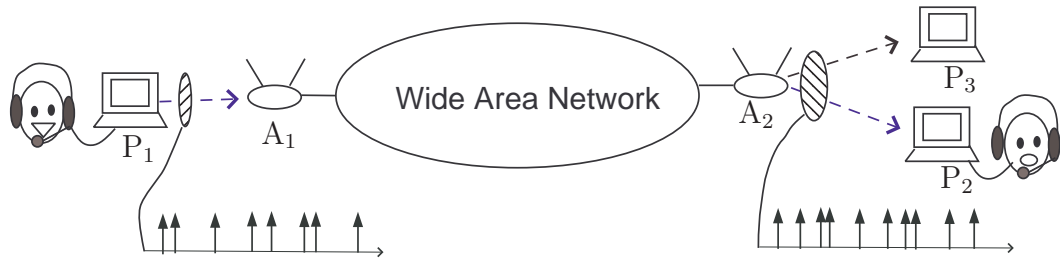
Figure 2.9: If P_1 has a VoIP conversation with either P_2 or P_3, the VoIP packets should depart from P_1 and travel through A_2.

Under $\mathcal{H}_1$, P_1 has a VoIP conversation with P_2, and P_3 downloads a file from a distant FTP server with 20kB/s rate. Since A_2 transmits packets for both P_2 and P_3, its transmission timings of FTP packets, destined for P_3, form the chaff part of S_2. Under $\mathcal{H}_0$, P_1 and P_2 engage in independent VoIP conversations while P_3 does the same job as in $\mathcal{H}_1$. Hence, VoIP packet timings in S_1 and those in S_2 are independent under $\mathcal{H}_0$. Under both hypotheses, the timings of network control/management packets from P_1 and A_2 (except beacon frames of A_2) are also included in S_1 and S_2.

We used $\Delta = 150$ms in the detection algorithms, because 150ms is the upper bound of acceptable end-to-end delays of VoIP packets recommended by ITU-T recommendation G.114 [67]. We first tested DAC with various p_Δ ranging from 10 to 400. Similar to the result on LBL TCP traces, the miss detection probability was higher than 0.55 in every case due to the bursty chaff transmissions (*i.e.*, bursty FTP transmissions from A_2 to P_3). Table 2.4 shows the error probabilities of NBFD, the wavelet analysis, and RP. As in the test using LBL traces, we used $\epsilon = 0.05$ for NBFD; for the wavelet analysis and RP, the crossover error rates and the corresponding thresholds are listed in the table. NBFD and the wavelet analysis outperformed RP, and they displayed vanishing error probabilities as the

44

Table 2.4: Performance on MSN VoIP data. NBFD parameters are $w = 2s$, $\alpha = \Delta = 0.15s$, and $\epsilon = 0.05$. The numbers of experiments are 162, 81, and 40 for observation duration 80s, 160s, and 320s, respectively. Under $\mathcal{H}_0$ and $\mathcal{H}_1$, the average rate is $(\lambda_1, \lambda_2) = (26.8, 34.9)$. The fraction of chaff in $\mathcal{H}_1$ traffic is 0.18.

	NBFD		Wavelet			RP		
Time	P_F	P_M	κ	P_F	P_M	th	P_F	P_M
80s	0.086	0.056	0.14	0.093	0.093	949	0.086	0.099
160s	0	0.049	0.17	0.012	0.012	989	0.049	0.074
320s	0	0	0.23	0	0	1005	0.075	0.050

observation duration increases.

In all the tests we executed, NBFD and the wavelet analysis consistently outperformed DAC and RP. Even though the wavelet analysis performed well over most traces, we need to recall that the results in Table 2.3 and Table 2.4 were possible because its threshold κ was set *a posteriori* to minimize its error probabilities. If neither a training data set nor a parametric model is available, we have no clear standard to set κ. For the further comparison of NBFD and the wavelet analysis, Fig. 2.10 shows P_F and P_M of NBFD and the wavelet analysis with various thresholds. We can observe that the optimal κ of the wavelet analysis varies significantly for different observation durations and different test traces. For instance, in the test result for synthetic Poisson traffic, $\kappa \approx 0.25$ gave the best performance when the observation duration is 80s, but it resulted in $P_M \approx 0.85$ for the 20s case. In addition, for the fixed observation duration of 80s, the optimal κ for the Poisson traffic (≈ 0.25) and that for the VoIP traffic (≈ 0.15) are quite different. In contrast, for NBFD, it can be observed that $\epsilon = 0.05$ results in almost optimal performance in every case. Especially, in every test, its false alarm probability vanished as the observation duration increases. This suggests that under $\mathcal{H}_0$, the difference between $\bar{R}(t)$ and $\bar{\tau}(t)$ of NBFD is well bounded by $\epsilon = 0.05$.
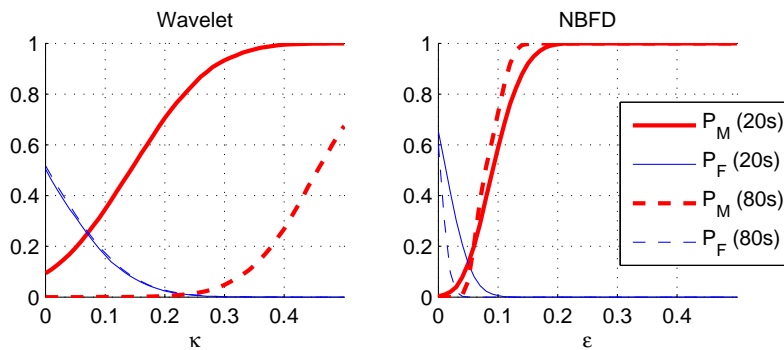
(a) Synthetic Poisson traffic
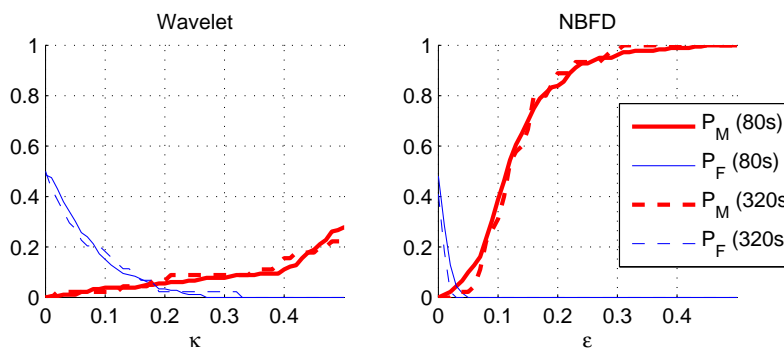
(b) LBL TCP traces
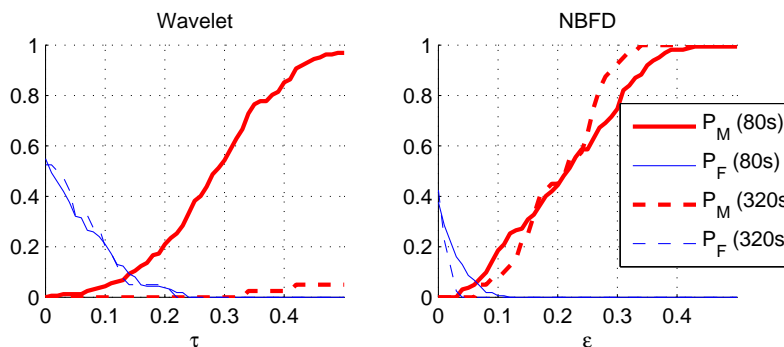
(c) MSN VoIP Experiment

Figure 2.10: False alarm and miss detection probabilities of the wavelet analysis and NBFD with various thresholds.

2.6 Proofs

2.6.1 Proof of Theorem 2.3.1

We use the following lemma about the relation between BiBGM with Δ and Bounded-Greedy-Match (BGM) [13] with 2Δ (For the detail of BGM, refer to

Section 4.A of [14]).

Lemma 2.6.0.1 *Running BiBGM on $(\mathbf{s}_i)_{i=1}^2$ with Δ is equivalent to the following:*

1. *Increase all the epochs of $\mathbf{s}_2$ by Δ.*

2. *Apply BGM with the delay constraint 2Δ to the modified measurements.*

Proof of Lemma 2.6.0.1: Let $\hat{\mathbf{s}}_2$ be a sequence generated by increasing every epoch in $\mathbf{s}_2$ by Δ (*i.e.*, $\hat{s}_2(i) = s_2(i) + \Delta$, $1 \le i \le |\mathcal{S}_2|$). Then, replacing $s_2(n)$ with $\hat{s}_2(n) - \Delta$ in Table 2.1 results in exactly the same pseudocode with BGM with 2Δ on $(\mathbf{s}_1, \hat{\mathbf{s}}_2)$ (see Table 3 in [14] for the pseudocode). □

Note that $(a, b) \in \mathcal{S}_1 \times \mathcal{S}_2$ and $|a - b| < \Delta$ if and only if $(a, b + \Delta) \in \mathcal{S}_1 \times \hat{\mathcal{S}}_2$ and $b + \Delta \in [a, a + 2\Delta]$. Hence, the optimal partitioning of $(\mathbf{s}_i)_{i=1}^2$ is equivalent to partitioning $(\mathbf{s}_1, \hat{\mathbf{s}}_2)$ into the unidirectional flow part (with the delay constraint 2Δ) and the chaff part such that the flow part is maximized; BGM with 2Δ was proved in [13] to achieve the optimal partitioning of $(\mathbf{s}_1, \hat{\mathbf{s}}_2)$. Thus, Lemma 2.6.0.1 implies the result. ■

2.6.2 Proof of Theorem 2.3.2

Let $\hat{\mathbf{S}}_2$ denote the point process with $\hat{S}_2(i) = S_2(i) + \Delta$, $i \ge 1$. Theorem 4.2 in [14] showed that if we run BGM with 2Δ on $(\mathbf{S}_1, \hat{\mathbf{S}}_2)$, the fraction of the matched epochs in total epochs converges a.s. to the following:

$$
\begin{cases}
\dfrac{2\lambda_1\lambda_2(1 - e^{2\Delta(\lambda_1 - \lambda_2)})}{(\lambda_1 + \lambda_2)(\lambda_2 - \lambda_1 e^{2\Delta(\lambda_1 - \lambda_2)})} & \text{if } \lambda_1 \ne \lambda_2 \\
\dfrac{2\lambda\Delta}{1 + 2\lambda\Delta} & \text{if } \lambda_1 = \lambda_2 = \lambda
\end{cases}
$$

Therefore, Lemma 2.6.0.1 implies the result. ■

2.6.3 Proof of Theorem 2.3.3

We first introduce the following lemma about the statistical behavior of $\bar{\mathrm{R}}(t)$ under $\mathcal{H}_1$.

Lemma 2.6.0.2 *Suppose that the distributions of $(\mathbf{S}_i)_{i=1}^2$ under $\mathcal{H}_1$ satisfy the conditions that (i) $\mathbf{S}_1$ and $\mathbf{S}_2$ have rates λ_1 and λ_2 respectively, (ii) $(\mathbf{F}_1, \mathbf{F}_2)$ is a bidirectional flow with rate[13] λ_f, and (iii) $\mathbf{W}_1$ and $\mathbf{W}_2$ are homogeneous Poisson processes. Then, under every distribution in $\mathcal{H}_1$, $\liminf_{t\to\infty} \bar{\mathrm{R}}(t) \geq \theta_{(\lambda_1, \lambda_2, \lambda_f)}$ a.s., where $\theta_{(\lambda_1, \lambda_2, \lambda_f)}$ is defined as*

$$
\begin{cases}
\dfrac{2\lambda_1 - 2\lambda_2(\frac{\lambda_1-\lambda_f}{\lambda_2-\lambda_f})e^{2\Delta(\lambda_1-\lambda_2)}}{(\lambda_2+\lambda_1)(1-(\frac{\lambda_1-\lambda_f}{\lambda_2-\lambda_f})e^{2\Delta(\lambda_1-\lambda_2)})} & \text{if } \lambda_1 \neq \lambda_2 \\[4mm]
\dfrac{\lambda_f + 2\lambda(\lambda-\lambda_f)\Delta}{\lambda(1+2(\lambda-\lambda_f)\Delta)} & \text{if } \lambda_1 = \lambda_2 = \lambda
\end{cases}
$$

Proof of Lemma 2.6.0.2: Let $N(t)$, $N_f(t)$, and $N_c(t)$ denote the number of epochs of $(\mathbf{S}_i)_{i=1}^2$, $(\mathbf{F}_i)_{i=1}^2$, and $(\mathbf{W}_i)_{i=1}^2$ in $[0, t]$, respectively. $M(t)$ denotes the number of the matched epochs found by running BiBGM over $(\mathbf{S}_i)_{i=1}^2$ in $[0, t]$.

Consider running BiBGM on $(\mathbf{F}_i)_{i=1}^2$ and $(\mathbf{W}_i)_{i=1}^2$ separately in $[0, t]$: $\hat{M}(t)$ denotes the sum of the number of the matched epochs in $(\mathbf{F}_i)_{i=1}^2$ and that in $(\mathbf{W}_i)_{i=1}^2$, and $\bar{\mathrm{R}}_w(t)$ denotes the fraction of the matched epochs in $(\mathbf{W}_i)_{i=1}^2$. Theorem 2.3.1 implies that running BiBGM on $(\mathbf{S}_i)_{i=1}^2$ results in a greater or an equal number of matched epochs than running BiBGM on $(\mathbf{F}_i)_{i=1}^2$ and $(\mathbf{W}_i)_{i=1}^2$ separately. Therefore,

$$M(t) \geq \hat{M}(t) = N_f(t) + N_c(t)\bar{\mathrm{R}}_w(t),$$
$$\frac{M(t)}{N(t)} \geq \frac{N_f(t)}{N(t)} + \frac{N_c(t)}{N(t)}\bar{\mathrm{R}}_w(t) = \frac{N_f(t)/t}{N(t)/t} + \frac{N_c(t)/t}{N(t)/t}\bar{\mathrm{R}}_w(t).$$

---

[13]If $N_1(t)$, $N_2(t)$, and $N_F(t)$ denote the number of epochs of $\mathbf{S}_1$, $\mathbf{S}_2$, and $\mathbf{F}_1$ in $[0, t]$, respectively, then $\lim_{t\to\infty} \frac{N_i(t)}{t} = \lambda_i$ a.s. for $i = 1, 2$, and $\lim_{t\to\infty} \frac{N_F(t)}{t} = \lambda_f$ a.s..

We have $\frac{M(t)}{N(t)} = \bar{R}(t)$, $\lim_{t\to\infty} \frac{N_f(t)/t}{N(t)/t} = \frac{2\lambda_f}{\lambda_1+\lambda_2}$ a.s., $\lim_{t\to\infty} \frac{N_c(t)/t}{N(t)/t} = \frac{\lambda_1+\lambda_2-2\lambda_f}{\lambda_1+\lambda_2}$ a.s., and $\lim_{t\to\infty} \bar{R}_w(t) = \phi_{(\lambda_1-\lambda_f,\lambda_2-\lambda_f)}$ a.s., where ϕ is defined as in Theorem 2.3.2. Thus,

$$\liminf_{t\to\infty} \bar{R}(t) \geq \frac{2\lambda_f}{\lambda_1+\lambda_2} + \frac{\lambda_1+\lambda_2-2\lambda_f}{\lambda_1+\lambda_2} \phi_{(\lambda_1-\lambda_f,\lambda_2-\lambda_f)} \quad \text{a.s..}$$

It can be shown that the right hand side is $\theta_{(\lambda_1,\lambda_2,\lambda_f)}$. $\qquad\square$

Let η be any fixed number in $(0, \frac{2\min\{\lambda_1,\lambda_2\}}{\lambda_1+\lambda_2})$ and τ be the suggested threshold for η. Then, there exists a positive $\hat{\lambda}_f$ such that $\frac{\hat{\lambda}_f}{\lambda_1+\lambda_2} = \frac{\eta}{4}$. Let $h(x) \triangleq \theta_{(\lambda_1,\lambda_2,x)}$. It can be checked that $h(x)$ is strictly increasing in $[0, \min\{\lambda_1,\lambda_2\}]$, and $h(\hat{\lambda}_f)$ is equal to τ.

(i) Miss detection probability: Suppose $\mathcal{H}_1$ is true and $R \geq \eta$ a.s.. Then, $R = \frac{2\lambda_f}{(\lambda_1+\lambda_2)}$ and $\lambda_f = \frac{(\lambda_1+\lambda_2)}{2}R > \hat{\lambda}_f$, because $R \geq \eta > \frac{\eta}{2}$. Lemma 2.6.0.2 and the monotonicity of h give

$$\liminf_{t\to\infty} \bar{R}(t) \geq \theta_{(\lambda_1,\lambda_2,\lambda_f)} = h(\lambda_f) > h(\hat{\lambda}_f) = \tau \quad \text{a.s.}$$

Hence, $\lim_{t\to\infty} \Pr(\bar{R}(t) < \tau) = 0$.

(ii) False alarm probability: Note that $h(0) = \phi_{(\lambda_1,\lambda_2)}$. Under $\mathcal{H}_0$,

$$\lim_{t\to\infty} \bar{R}(t) = \phi_{(\lambda_1,\lambda_2)} = h(0) < h(\hat{\lambda}_f) = \tau \quad \text{a.s.,}$$

and thus $\lim_{t\to\infty} \Pr(\bar{R}(t) \geq \tau) = 0$. Furthermore, Lemma 2.6.0.1 and Theorem 6.4 in [14] imply the exponential decay of the false alarm probability. $\qquad\blacksquare$

2.6.4 Proof of Theorem 2.4.1

We first introduce a useful lemma.
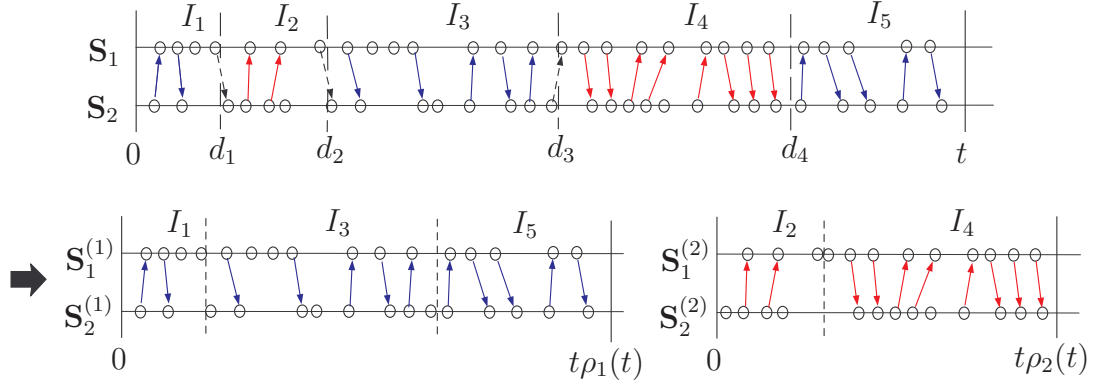
Figure 2.11: In this example, $M = 2$, $a_{1;1} = 1$, $a_{1;2} = 3$, $a_{1;3} = 5$, $a_{2;1} = 2$, and $a_{2;2} = 4$. We ran BiBGM on $(\mathbf{S}_i)_{i=1}^2$ and marked the matches by the arrows. Some matches consist of epochs in two different partitions, and they are marked by the dashed arrows. The matches consisting of epochs in a single partition are marked by the solid arrows. One can observe that each solid arrow in $(\mathbf{S}_i)_{i=1}^2$ can be found either in $(\mathbf{S}_i^{(1)})_{i=1}^2$ or $(\mathbf{S}_i^{(2)})_{i=1}^2$.

Lemma 2.6.0.3 *Suppose that $\mathbf{S}_1$ and $\mathbf{S}_2$ are non-homogeneous Poisson processes, and their local intensities always stay in $[\lambda_{min}, \lambda_{max}]$, where $\lambda_{min} > 0$. As illustrated in Fig. 2.11, we partition $[0, \infty)$ into a countable number of subintervals: I_i denotes the ith subinterval, T_i is the length of I_i, and $d(t)$ denotes the number of I_is with $I_i \subset [0, t]$. Suppose $\frac{d(t)}{t}$ decreases to 0 as t grows.*

Let M be a finite natural number and suppose we partition the set $\{I_i, i \geq 1\}$ into M subsets $\{I_{a_{k;i}}, i \geq 1\}$, $1 \leq k \leq M$, where $(a_{k;i})_{i\geq1}$, $1 \leq k \leq M$, are subsequences of $(1, 2, 3, \ldots)$. For $1 \leq k \leq M$, we use the epochs of $(\mathbf{S}_i)_{i=1}^2$ in $(I_{a_{k;i}})_{i\geq1}$ to generate point processes $(\mathbf{S}_i^{(k)})_{i=1}^2$, as described in Fig. 2.11:

1. *Initially, $\mathbf{S}_1^{(k)}$ and $\mathbf{S}_2^{(k)}$ have no epoch.*

2. *For $n \geq 1$, for $i = 1, 2$, subtract $\sum_{j=1}^{a_{k;n}-1} T_j$ from all the epochs of $\mathbf{S}_i$ in the interval $I_{a_{k;n}}$, add $\sum_{j=1}^{n-1} T_{a_{k;j}}$ to them, and add these epochs to $\mathbf{S}_i^{(k)}$.*

Let $N(t)$ denote the number of epochs of $(\mathbf{S}_i)_{i=1}^2$ in $[0, t]$ and $N^{(k)}(t)$ denote

the number of epochs of $(\mathbf{S}_i^{(k)})_{i=1}^2$, whose original epoch in $(\mathbf{S}_i)_{i=1}^2$ is in $[0, t]$; by definition, $N(t) = \sum_{k=1}^M N^{(k)}(t)$. We run BiBGM on $(\mathbf{S}_i)_{i=1}^2$ and let $\bar{R}(t)$ denote the fraction of the matched epochs in the total epochs in $[0, t]$. In addition, we run BiBGM on $(\mathbf{S}_i^{(k)})_{i=1}^2$ separately for each k, and $N_f^{(k)}(t)$ denotes the number of the matched epochs among the earliest $N^{(k)}(t)$ epochs of $(\mathbf{S}_i^{(k)})_{i=1}^2$. And, we define $\hat{R}(t)$ as $\dfrac{\sum_{k=1}^M N_f^{(k)}(t)}{N(t)}$.

Then, $\lim_{t \to \infty} \bar{R}(t) - \hat{R}(t) = 0$ almost surely.

Proof of Lemma 2.6.0.3: Let $N_f(t)$ denote the number of BiBGM-matched epochs of $(\mathbf{S}_i)_{i=1}^2$ in $[0, t]$. Then, by definition, $\bar{R}(t) = \frac{N_f(t)}{N(t)}$. Let d_i denote the time that the ith division occurs; in other words, d_i is the time that the ith jump of $d(t)$ occurs. Formally, we say that a BiBGM match (t_1, t_2), where t_i is an epoch of $\mathbf{S}_i$, is *broken* if $t_1 \in I_a$, $t_2 \in I_b$, and $a \neq b$. Let $\tilde{N}_f(t)$ denote the number of epochs of the unbroken BiBGM matches in $[0, t]$. As described in Fig. 2.11, if an unbroken BiBGM match (t_1, t_2) in $[0, t]$ is such that t_1 and t_2 are included in a single partition $I_{a_{k;\,i}}$, then its shifted version can be found in the $[0, t\rho_k(t)]$ interval of $(\mathbf{S}_i^{(k)})_{i=1}^2$, where $\rho_k(t)$ is the fraction of $(\bigcup_{i \geq 1} I_{a_{k;\,i}}) \cap [0, t]$ in $[0, t]$. In addition, Theorem 2.3.1 implies that $N_f^{(k)}(t)$ is no less than the number of epochs belonging to the shifted unbroken matches in $[0, t\rho_k(t)]$ of $(\mathbf{S}_i^{(k)})_{i=1}^2$ (i.e., solid arrows in Fig. 2.11). Therefore, $\sum_{k=1}^M N_f^{(k)}(t) \geq \tilde{N}_f(t)$.

For $j = 1, 2$, let $X_j(i)$ denote the number of epochs of $\mathbf{S}_j$ in $[\max\{\frac{d_{i-1}+d_i}{2}, d_i - \Delta\}, \min\{\frac{d_i+d_{i+1}}{2}, d_i + \Delta\})$, where $d_0 \triangleq -d_1$. The number of epochs of the broken matches in $[0, t]$ is bounded above by $\sum_{i=1}^{d(t)} X_1(i) + \sum_{i=1}^{d(t)} X_2(i)$. Hence,

$$\tilde{N}_f(t) \geq N_f(t) - \sum_{i=1}^{d(t)} X_1(i) - \sum_{i=1}^{d(t)} X_2(i)$$

51

There exist sequences of i.i.d. Poisson random variables $(\bar{X}_1(i))_{i\geq 1}$ and $(\bar{X}_2(i))_{i\geq 1}$ with mean $2\lambda_{max}\Delta$ such that $X_j(i) \leq \bar{X}_j(i)$ a.s. for $i \geq 1$, $j = 1, 2$. Hence,

$$\sum_{k=1}^{M} N_f^{(k)}(t) \geq N_f(t) - \sum_{i=1}^{d(t)} \bar{X}_1(i) - \sum_{i=1}^{d(t)} \bar{X}_2(i),$$
$$\bar{R}(t) - \hat{R}(t) \leq \frac{\sum_{i=1}^{d(t)} \bar{X}_1(i)}{N(t)} + \frac{\sum_{i=1}^{d(t)} \bar{X}_2(i)}{N(t)}.$$

For $j = 1, 2$, we have

$$\limsup_{t\to\infty} \frac{d(t)/t}{N(t)/t} \frac{\sum_{i=1}^{d(t)} \bar{X}_j(i)}{d(t)} = 0 \quad \text{a.s..}$$

Hence,

$$\limsup_{t\to\infty}(\bar{R}(t) - \hat{R}(t)) \leq 0 \quad \text{a.s..}$$

Similarly, we can partition $(\mathbf{S}_i^{(k)})_{i=1}^2$ at time points $(d_{k;i})_{i\geq 1}$, where $d_{k;i} \triangleq \sum_{j=1}^{i} T_{a_{k;j}}$, and use the number of unbroken BiBGM matches of $(\mathbf{S}_i^{(k)})_{i=1}^2$ in $[0, t\rho_k(t)]$, $1 \leq k \leq M$, to obtain a lower bound on the number of BiBGM matches of $(\mathbf{S}_i)_{i=1}^2$ in $[0, t]$. Then, based on the similar argument, we can derive $\liminf_{t\to\infty}(\bar{R}(t) - \hat{R}(t)) \geq 0$ a.s.. Hence, we have $\lim_{t\to\infty}(\bar{R}(t) - \hat{R}(t)) = 0$ a.s., and the proof is complete. $\qquad\square$

The proof of Theorem 2.4.1 consists of two parts: one for proving the vanishing false alarm probability under $\mathcal{H}_0$, and the other for proving the vanishing miss detection probability under $\mathcal{H}_1$.

False Alarm Probability

Suppose that $\mathcal{H}_0$ is true and the distribution of $(\mathbf{S}_i)_{i=1}^2$ satisfies the assumptions of the theorem. $\mathbf{S}_1$ and $\mathbf{S}_2$ are independent non-homogeneous Poisson processes, and so are the output of ITA, $\bar{\mathbf{S}}_1$ and $\bar{\mathbf{S}}_2$. Suppose we run BiBGM on $(\mathbf{S}_i)_{i=1}^2$ and let $\bar{R}(t)$ denote the fraction of the matched epochs in the total epochs in $[0, t]$.

We also run BiBGM on $(\bar{\mathbf{S}}_i)_{i=1}^2$ and let $\bar{\mathrm{T}}(t)$ denote the fraction of the matched epochs in the total epochs in $[0, \lfloor \frac{t}{2(w+\alpha)} \rfloor w]$. In the following, we will show that $\bar{\mathrm{R}}(t) - \bar{\mathrm{T}}(t)$ converges a.s. to 0.

Because $\lim_{t \to \infty} \frac{c(t)}{t} = 0$, there are at most a countable number of intensity changes. Let $(c_i)_{i \geq 1}$ denote the increasing sequence of the time points at which $(\lambda_1(t), \lambda_2(t))$ changes. We partition $[0, \infty)$ into a countable number of subintervals $\{I_i \triangleq [c_{i-1}, c_i), i \geq 1\}$. For $1 \leq k \leq M_0$, $(a_{k;i})_{i \geq 1}$ denotes the increasing sequence of all the indices of I_is in which $(\lambda_1(t), \lambda_2(t)) = \vec{\mu}^{(k)}$. For each k, we use the epochs of $(\mathbf{S}_i)_{i=1}^2$ in $(I_{a_{k;i}})_{i \geq 1}$ to generate a pair of point processes $(\mathbf{S}_i^{(k)})_{i=1}^2$, as described in Lemma 2.6.0.3.

Let $N(t)$ denote the number of epochs of $(\mathbf{S}_i)_{i=1}^2$ in $[0, t]$. Suppose we run BiBGM on $(\mathbf{S}_i^{(k)})_{i=1}^2$ separately for $1 \leq k \leq M_0$. $N^{(k)}(t)$ denotes the number of epochs of $(\mathbf{S}_i^{(k)})_{i=1}^2$ in $[0, t\rho_k(t)]$, and $N_f^{(k)}(t)$ denotes the number of BiBGM-matched epochs among those $N^{(k)}(t)$ epochs. Then, Lemma 2.6.0.3 implies $\lim_{t \to \infty} \left(\bar{\mathrm{R}}(t) - \frac{\sum_{k=1}^M N_f^{(k)}(t)}{N(t)} \right) = 0$ a.s.. And,

$$\frac{\sum_{k=1}^{M_0} N_f^{(k)}(t)}{N(t)} = \frac{t}{N(t)} \sum_{k=1}^{M_0} \rho_k(t) \frac{N^{(k)}(t)}{t\rho_k(t)} \frac{N_f^{(k)}(t)}{N^{(k)}(t)}.$$

By analyzing the limiting behaviors of $\frac{t}{N(t)}$, $\rho_k(t) \frac{N^{(k)}(t)}{t\rho_k(t)}$, and $\frac{N_f^{(k)}(t)}{N^{(k)}(t)}$ (use Theorem 2.3.2), we have

$$\lim_{t \to \infty} \frac{t}{N(t)} \sum_{k=1}^{M_0} \rho_k(t) \frac{N^{(k)}(t)}{t\rho_k(t)} \frac{N_f^{(k)}(t)}{N^{(k)}(t)} = \frac{\sum_{k=1}^{M_0} \rho_k(\mu_1^{(k)} + \mu_2^{(k)}) \phi_{(\mu_1^{(k)}, \mu_2^{(k)})}}{\sum_{k=1}^{M_0} \rho_k(\mu_1^{(k)} + \mu_2^{(k)})} \quad \text{a.s.}$$

where $\rho_k \triangleq \lim_{t \to \infty} \rho_k(t)$ and ϕ is defined in Theorem 2.3.2. Then, by Lemma 2.6.0.3,

$$\lim_{t \to \infty} \bar{\mathrm{R}}(t) = \frac{\sum_{k=1}^{M_0} \rho_k(\mu_1^{(k)} + \mu_2^{(k)}) \phi_{(\mu_1^{(k)}, \mu_2^{(k)})}}{\sum_{k=1}^{M_0} \rho_k(\mu_1^{(k)} + \mu_2^{(k)})} \quad \text{a.s.} \tag{2.3}$$
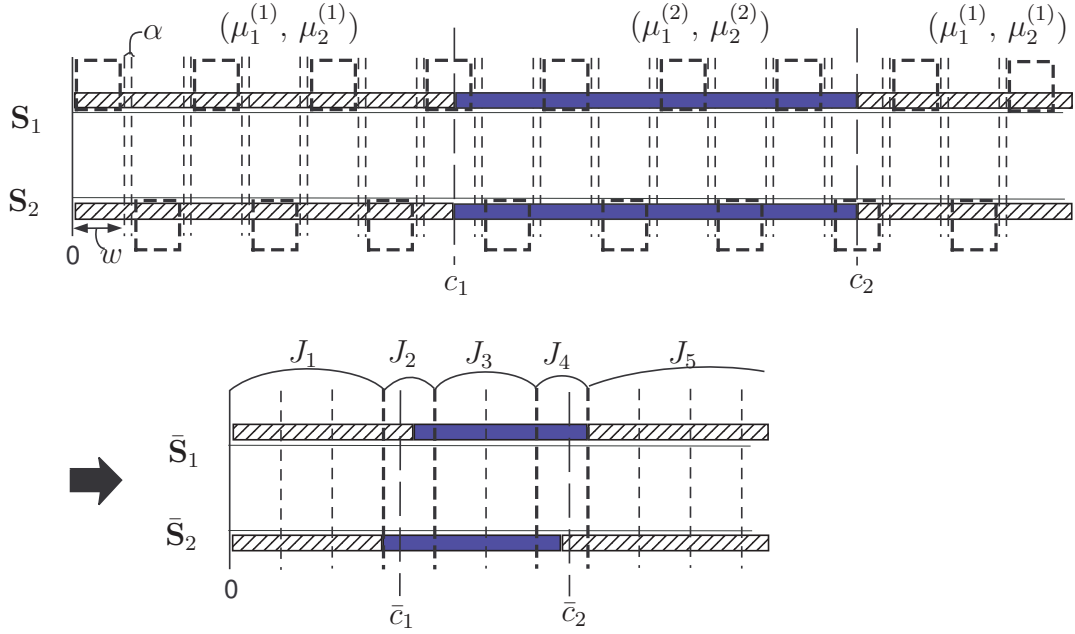
Figure 2.12: This figure illustrates a simple case that $(\lambda_1(t), \lambda_2(t))$ can only take either $(\mu_1^{(1)}, \mu_2^{(1)})$ or $(\mu_1^{(2)}, \mu_2^{(2)})$. The bars filled with slant lines represent the intervals in which $\lambda_i(t) = \mu_i^{(1)}$, and the blue bars represent the intervals in which $\lambda_i(t) = \mu_i^{(2)}$. In this example, J_2 and J_4 are in C.

Now, we will prove that $\bar{T}(t)$ also converges almost surely to the same constant. Let $\bar{c}_i \triangleq \frac{w}{2(w+\alpha)} c_i$, $\forall i$. As depicted in Fig. 2.12, the local intensities of $\bar{S}_1$ and $\bar{S}_2$, denoted by $(\bar{\lambda}_1(t), \bar{\lambda}_2(t))$, may be equal to $(\mu_1^{(j)}, \mu_2^{(k)})$ with $j \neq k$, and it happens only if any $\bar{c}_i$ is in $[w\lfloor \frac{t}{w} \rfloor, w(\lfloor \frac{t}{w} \rfloor + 1))$. Define C as a set

$$\{[w(k-1), wk) : k \in \mathbb{N}, \exists\, i \text{ s.t. } \bar{c}_i \in [w(k-1), wk)\}$$

As illustrated in Fig. 2.12, we partition $[0, \infty)$ of $(\bar{S}_i)_{i=1}^2$ into the intervals in C and the gap intervals between two adjacent intervals in C, and $(J_i)_{i\geq 1}$ denotes the sequence of these intervals arranged in a time order. $\{\bar{a}_{0;i}, i \geq 1\}$ denotes the increasing sequence of the indices of J_is satisfying $J_i \in C$. For $1 \leq k \leq M_0$, $\{\bar{a}_{k;i}, i \geq 1\}$ denotes the increasing sequence of the indices of J_is satisfying $(\bar{\lambda}_1(t), \bar{\lambda}_2(t)) = \vec{\mu}^{(k)}$, $\forall t \in J_i$. Then, $\{J_i, i \geq 1\}$ can be partitioned into the

54

$(M_0 + 1)$ sets, $\{J_{\bar{a}_{k;\,i}},\ i \geq 1\}$, $0 \leq k \leq M_0$. For $0 \leq k \leq M_0$, we use the epochs of $(\bar{\mathbf{S}}_i)_{i=1}^2$ in $(J_{\bar{a}_{k;\,i}})_{i\geq 1}$ to generate $(\bar{\mathbf{S}}_i^{(k)})_{i=1}^2$, in the same manner as we generate $(\mathbf{S}_i^{(k)})_{i=1}^2$ based on $(I_{a_{k;\,i}})_{i\geq 1}$ in Lemma 2.6.0.3. Then, based on Lemma 2.6.0.3 and $(\bar{\mathbf{S}}_i^{(k)})_{i=1}^2$ $(0 \leq k \leq M_0)$, we can use the similar argument as in obtaining $\lim_{t\to\infty} \bar{R}(t)$ and show

$$\lim_{t\to\infty} \bar{T}(t) = \frac{\sum_{k=1}^{M_0} \rho_k(\mu_1^{(k)} + \mu_2^{(k)})\phi_{(\mu_1^{(k)}, \mu_2^{(k)})}}{\sum_{k=1}^{M_0} \rho_k(\mu_1^{(k)} + \mu_2^{(k)})} \quad \text{a.s.} \tag{2.4}$$

From (2.3) and (2.4), we can see that $\bar{R}(t) - \bar{T}(t)$ converges almost surely to 0 as t grows. Hence, for any positive ϵ, the false alarm probability vanishes as t grows:

$$\lim_{t\to\infty} P_F(t) = \lim_{t\to\infty} \Pr(\bar{R}(t) - \bar{T}(t) > \epsilon) = 0$$

Miss Detection Probability

Suppose that $\mathcal{H}_1$ is true and the distribution of $(\mathbf{S}_i)_{i=1}^2$ satisfies the assumptions of the theorem including $R \geq \eta$ a.s. Due to the almost sure convergence of $R(t)$, '$R = \liminf_{t\to\infty} R(t) \geq \eta$ a.s.' is equivalent to

$$\frac{\sum_{k=1}^{M_1} \rho_k(\lambda_{f1}^{(k)} + \lambda_{f2}^{(k)})}{\sum_{k=1}^{M_1} \rho_k(\lambda_1^{(k)} + \lambda_2^{(k)})} \geq \eta, \tag{2.5}$$

where $\rho_k \triangleq \lim_{t\to\infty} \rho_k(t)$. In addition, the first assumption of the theorem guarantees that $\bar{\mathbf{S}}_1$ and $\bar{\mathbf{S}}_2$ are independent non-homogeneous Poisson processes. We run BiBGM on $(\mathbf{S}_i)_{i=1}^2$ and let $\bar{R}(t)$ denote the fraction of the matched epochs in the total epochs in $[0, t]$. We also run BiBGM on $(\bar{\mathbf{S}}_i)_{i=1}^2$ and let $\bar{T}(t)$ denote the fraction of the matched epochs in the total epochs in $[0, \lfloor \frac{t}{2(w+\alpha)} \rfloor w]$. First of all, by following exactly the same steps as in the proof of vanishing false alarm

probability, we can derive

$$\lim_{t \to \infty} \bar{\mathrm{T}}(t) = \frac{\sum_{k=1}^{M_1} \rho_k(\lambda_1^{(k)} + \lambda_2^{(k)})\phi_{(\lambda_1^{(k)}, \lambda_2^{(k)})}}{\sum_{k=1}^{M_1} \rho_k(\lambda_1^{(k)} + \lambda_2^{(k)})} \quad \text{a.s.} \tag{2.6}$$

Let $(c_i)_{i \geq 1}$ denote the increasing sequence of the time points at which $\vec{\lambda}(t)$ changes, and we partition $[0, \infty)$ into a countable number of subintervals $\{I_i \triangleq [c_{i-1}, c_i), i \geq 1\}$. For $1 \leq k \leq M_1$, let $(a_{k;i})_{i \geq 1}$ denote the increasing sequence of all the indices of I_is satisfying $\vec{\lambda}(t) = \vec{\lambda}^{(k)}, \forall t \in I_i$. We use the epochs of $(\mathbf{S}_i)_{i=1}^2$ in $(I_{a_{k;i}})_{i \geq 1}$ to generate a pair of point processes $(\mathbf{S}_i^{(k)})_{i=1}^2$, as in Lemma 2.6.0.3. Then, based on Lemma 2.6.0.2, Lemma 2.6.0.3, and $(\mathbf{S}_i^{(k)})_{i=1}^2$ $(1 \leq k \leq M_1)$, we can use the similar argument as in obtaining $\lim_{t \to \infty} \bar{\mathrm{R}}(t)$ in the proof of vanishing false alarm probability and derive

$$\liminf_{t \to \infty} \bar{\mathrm{R}}(t) \geq \frac{\sum_{k=1}^{M_1} \rho_k(\lambda_1^{(k)} + \lambda_2^{(k)})\theta_{(\lambda_1^{(k)}, \lambda_2^{(k)}, \lambda_{f1}^{(k)} + \lambda_{f2}^{(k)})}}{\sum_{k=1}^{M_1} \rho_k(\lambda_1^{(k)} + \lambda_2^{(k)})} \quad \text{a.s.}$$

where θ is defined in Lemma 2.6.0.2. For fixed λ_1 and λ_2, $\theta_{(\lambda_1, \lambda_2, \lambda_f)}$ is a strictly increasing function of λ_f, and it decreases to $\phi_{(\lambda_1, \lambda_2)}$ as λ_f decays to 0. Hence, if we define γ as

$$\min_{(\rho_k)_{k=1}^{M_1}} \frac{\sum_{k=1}^{M_1} \rho_k(\lambda_1^{(k)} + \lambda_2^{(k)})\theta_{(\lambda_1^{(k)}, \lambda_2^{(k)}, \lambda_{f1}^{(k)} + \lambda_{f2}^{(k)})}}{\sum_{k=1}^{M_1} \rho_k(\lambda_1^{(k)} + \lambda_2^{(k)})} - \frac{\sum_{k=1}^{M_1} \rho_k(\lambda_1^{(k)} + \lambda_2^{(k)})\phi_{(\lambda_1^{(k)}, \lambda_2^{(k)})}}{\sum_{k=1}^{M_1} \rho_k(\lambda_1^{(k)} + \lambda_2^{(k)})},$$

where the minimization is over $\{(\rho_k)_{k=1}^{M_1} : \ (2.5) \text{ holds}\}$, then it can be easily seen that γ is strictly greater than 0. Set $\bar{\epsilon} = \frac{1}{3}\gamma$, and let ϵ be an arbitrary number in $(0, \bar{\epsilon}]$. Then, if the condition (2.5) holds,

$$\liminf_{t \to \infty}(\bar{\mathrm{R}}(t) - \bar{\mathrm{T}}(t)) \geq \gamma > 2\epsilon \quad \text{a.s..}$$

Therefore, as long as the condition (2.5) holds, the miss detection probability vanishes as t grows:

$$\lim_{t \to \infty} P_M(t) = \lim_{t \to \infty} \Pr(\bar{\mathrm{R}}(t) - \bar{\mathrm{T}}(t) < \epsilon) = 0.$$

∎

CHAPTER 3

TOPOLOGY ATTACK OF A POWER GRID

3.1 Introduction

A defining feature of a smart grid is its abilities to monitor the state of a large power grid, to adapt to changing operating conditions, and to react intelligently to contingencies, all of which depend critically on a reliable and secure cyber-infrastructure. It has been widely recognized that the heavy reliance on a wide area communications network for grid monitoring and real-time operation comes with increasing security risks of cyber-attacks. See [72] for a vulnerability analysis of energy delivery control systems.

While *information* security has been a major focus of research for over half a century, the mechanisms and the impacts of attack on *cyber physical systems* such as the power grid are not yet well understood, and effective countermeasures are still lacking.

We consider a form of "man-in-the-middle" (MiM) attack [73] on the topology of a power grid. An MiM attack exploits the lack of authentication in a system, which allows an adversary to impersonate a legitimate participant. In the context of monitoring a transmission grid, sophisticated authentications are typically not implemented due to the need of reducing communication delay and the presence of legacy communication equipment. If an adversary is able to gain access to remote terminal units (RTUs) or local data concentrators, it is possible for the adversary to replace actual data packets with carefully constructed malicious data packets and impersonate a valid data source.

MiM attacks on a power grid may have severe consequences. The adversary can mislead the control center that the grid is operating under a topology different from that in reality. Such an attack, if launched successfully and undetected by the control center, will have serious implications: a grid that is under stress may appear to be normal to the operator thereby delaying the deployment of necessary measures to ensure stability. Similarly, a grid operating normally may appear to be under stress to the operator, potentially causing load shedding and other costly remedial actions by the operator.

Launching a topology attack, fortunately, is not easy; a modern energy management system is equipped with relatively sophisticated bad data and topology error detectors, which alerts the operator that either the data in use are suspicious or there may indeed be changes in the network topology. When there are inconsistencies between the estimated network topology (estimated mostly using switch and breaker states) and the meter data (*e.g.*, there is significant amount of power flow on a line disconnected in the estimated topology,) the operator takes actions to validate the data in use. Only if data and the estimated topology pass the bad data test, will the topology change be accepted and updates be made for subsequent actions.

The attacks that are perhaps the most dangerous are those that pass the bad data detection so that the control center accepts the change (or the lack of change) of network topology. To launch such attacks, the adversary needs to modify simultaneously the meter data and the network data (switch and breaker states) in such a way that the estimated topology is consistent with the data. Such attacks are referred to as *undetectable attacks*; they are the main focus of this study.

3.1.1 Summary of Results and Organization

We aim to achieve two objectives. First, we characterize conditions under which undetectable attacks are possible, given a set of vulnerable meters that may be controlled by an adversary. To this end, we consider two attack regimes based on the *information set* available to the attacker. The more information the attacker has, the stronger its ability to launch a sophisticated attack that is hard to detect.

The *global information* regime is where the attacker can observe all meter and network data before altering the adversary-controlled part of them. Although it is unlikely in practice that an adversary is able to operate in such a regime, in analyzing the impact of attacks, it is typical to consider the worst case by granting the adversary additional power. In Section 3.3, we present a necessary and sufficient algebraic condition under which, given a set of adversary controlled meters, there exists an undetectable attack that misleads the control center with an incorrect "target" topology. This algebraic condition provides not only numerical ways to check if the grid is vulnerable to undetectable attacks but also insights into which meters to protect to defend against topology attacks. We also provide specific constructions of attacks and show certain optimality of the proposed attacks.

A more practically significant situation is the *local information* regime where the attacker has only local information from those meters it has gained control. Under certain conditions, undetectable attacks exist and can be implemented easily based on simple heuristics. We present in Section 3.4 intuitions behind such simple attacks and implementation details.

The second objective is to provide conditions under which topology attack cannot be made undetectable. Such a condition, even if it may not be the tightest,

provides insights into defense mechanisms against topology attacks. In Section 3.5, we show that if a set of meters satisfying a certain branch covering property are protected, then topology attacks can always be detected. In practice, protecting a meter may be carried out at multiple levels, from physical protection measures to software protection schemes using more sophisticated authentication protocols.

The rest of the chapter is organized as follows. Section 3.2 presents mathematical models of state estimation, bad data test, and topology attacks. In Section 3.3, we study topology attacks in the global information regime. The algebraic condition for an undetectable attack is presented, and construction of a cost-effective undetectable attack is provided. Section 3.4 presents a heuristic attack for the attacker with local information. Based on the algebraic condition presented in Section 3.3, Section 3.5 provides a graph theoretical strategy to add protection to a subset of meters to prevent undetectable attacks. Section 3.6 presents simulation results to demonstrate practical uses of our analysis and feasibility of the proposed attacks.

3.2 Preliminaries

In this section, we present models for the power network, measurements, and adversary attacks. We also summarize essential operations such as state estimation and bad data detection that are targets of data attacks.

3.2.1 Network and Measurement Models

The control center receives two types of data from meters and sensors deployed throughout the grid. One is the digital network data $\mathbf{s} \in \{0,1\}^d$, which can be represented as a string of binary bits indicating the on and off states of various switches and line breakers. The second type is the analog meter data $\mathbf{z}$, which is a vector of bus injection and line flow measurements.

Without an attack or a sensing error, $\mathbf{s}$ gives the true breaker states. Each $\mathbf{s} \in \{0,1\}^d$ corresponds to a system topology, which is represented by a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of buses and $\mathcal{E}$ is the set of *connected* transmission lines. For each physical transmission line between two buses (*e.g.*, i and j), we assign an arbitrary direction for the line (*e.g.*, (i,j)), and (i,j) is in $\mathcal{E}$ if and only if the line is connected. In addition, $\mathcal{E}_0$ denotes the set of all lines (with the assigned directions), both connected and disconnected. Assigning arbitrary directions for lines is not intended to deliver any physical meaning, but only for ease of presentation.

The state of a power system is defined as the vector $\mathbf{x}$ of voltage phasors on all buses. In the absence of attacks and measurement noise, the meter data $\mathbf{z}$ collected by the SCADA system are related to the system state $\mathbf{x}$ and the system topology $\mathcal{G}$ via the AC power flow model [51]:

$$\mathbf{z} = h(\mathbf{x}, \mathcal{G}) + \mathbf{e} \qquad (3.1)$$

where $\mathbf{z}$ typically includes real and reactive parts of bus injection and line flow measurements, h is the nonlinear measurement function of $\mathbf{x}$ and $\mathcal{G}$, and $\mathbf{e}$ the additive noise.

A simplified model, one that is often used in real-time operations such as the

computation of real-time LMP, is the so-called DC model [51] where the nonlinear function h is linearized near the operating point. In particular, the DC model is given by

$$\mathbf{z} = H\mathbf{x} + \mathbf{e} \qquad (3.2)$$

where $\mathbf{z} \in \mathbb{R}^m$ consists of only the real parts of injection and line flow measurements, $H \in \mathbb{R}^{m \times n}$ is the measurement matrix, $\mathbf{x} \in \mathbb{R}^n$ is the state vector consisting of voltage phase angles at all buses except the slack bus, and $\mathbf{e} \in \mathbb{R}^m$ is the Gaussian measurement noise with a diagonal covariance matrix Σ.

The fact that the measurement matrix H depends on the network topology $\mathcal{G}$ is important, although we use the notation H without explicit association with its topology $\mathcal{G}$ for notational convenience. For ease of presentation, consider the noiseless measurement $\mathbf{z} = H\mathbf{x}$. If an entry z_k of $\mathbf{z}$ is the measurement of the line flow from i to j of a *connected* line in $\mathcal{G}$, z_k is $B_{ij}(x_i - x_j)$ where B_{ij} is the line susceptance and x_i is the voltage phase angle at bus i. The corresponding row of H is equal to

$$\mathbf{h}_{(i,j)} \triangleq [0 \cdots 0 \quad \underbrace{B_{ij}}_{i\text{th entry}} \quad 0 \cdots 0 \quad \underbrace{-B_{ij}}_{j\text{th entry}} \quad 0 \cdots 0]. \qquad (3.3)$$

On the other hand, if z_k is the measurement of the line flow through a *disconnected* line in $\mathcal{G}$, z_k is zero, and the corresponding row of H consists of all zero entries. If z_k is the measurement of bus injection at i, it is the sum of all the outgoing line flows from i, and the corresponding row of H is the sum of the row vectors corresponding to all the outgoing line flows.

We consider both AC and DC power flow models. The DC model allows us to obtain a succinct characterization of undetectable attacks as described in Section 3.3. However, these results hold only locally around the operating point,

62

because the results are obtained from the linearized model. General results for the more realistic (nonlinear) AC model are difficult to obtain. We present in Section 3.4 a heuristic attack that are undetectable for both AC and DC models.

It was shown in [39] that using the DC model and linear state estimator in numerical analysis of an attack tends to exaggerate the impact of the attack. Hence, for accurate analysis, we use the AC model and nonlinear state estimator in the numerical simulations presented in Section 3.6.

3.2.2 Adversary Model

The adversary aims at modifying the topology estimate from $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to a different "target" topology $\bar{\mathcal{G}} = (\mathcal{V}, \bar{\mathcal{E}})$. Note that $\mathcal{G}$ and $\bar{\mathcal{G}}$ have the same set of vertices. In other words, we only consider the attacks aimed at perturbing transmission line connectivities[1]. In addition, we assume that the power system is observable regardless an attack is present or not: *i.e.*, the measurement matrix in the DC model always has full rank. This means that the adversary avoids misleading the control center with drastic system changes (*e.g.*, division into two diconnected parts) that may draw too much attention of the control center[2]. We call the lines not common to both $\bar{\mathcal{E}}$ and $\mathcal{E}$ (*i.e.*, lines in $\bar{\mathcal{E}} \triangle \mathcal{E} \triangleq (\bar{\mathcal{E}} \setminus \mathcal{E}) \cup (\mathcal{E} \setminus \bar{\mathcal{E}})$) *target lines* and the buses at the ends of the target lines *target buses*.

To alter the network topology, the adversary launches a man-in-the-middle

---

[1]The attacks aiming to split or combine buses are out of scope of this chapter. Such attacks require modifying the measurements of breaker states *inside* substations. If the control center employs generalized state estimation [74], such modification invokes substation-level state estimation which leads to a robust bad data test. Hence, such attacks are harder to avoid detection.

[2]In fact, the results to be presented in this chapter also hold for the general case where the target topology can be anything (*e.g.*, the system may be divided into several disconnected parts), if the control center employs the same bad data test even when the network is unobservable.
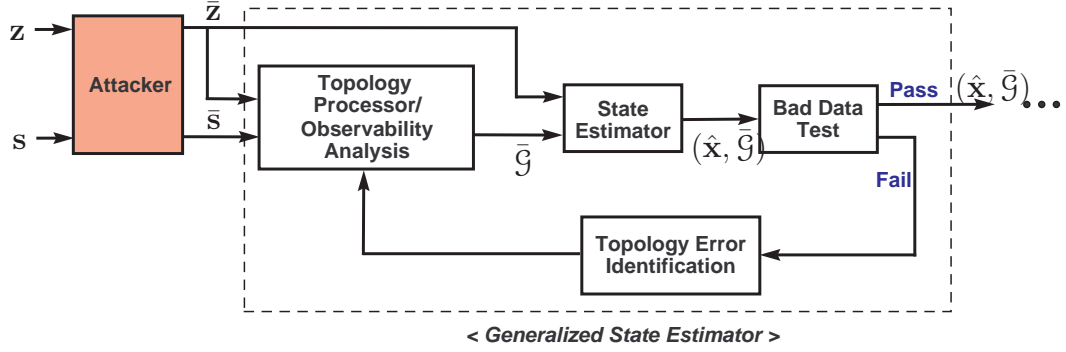
Figure 3.1: Attack Model with Generalized State Estimation

attack as described in Fig. 3.1: it intercepts $(\mathbf{s}, \mathbf{z})$ from RTUs, modifies part of them, and forwards the modified version $(\bar{\mathbf{s}}, \bar{\mathbf{z}})$ to the control center.

Throughout this chapter, except in Section 3.4, we assume that the adversary has global information, $i.e.$, it knows network parameters and observes all entries of $(\mathbf{s}, \mathbf{z})$ before launching the attack, although it may modify only the entries it gained control of. Such an unlimited access to network parameters and data is a huge advantage to the attacker. In Section 3.5, countermeasures are designed under this assumption so that they can be robust to such worst case attacks.

The mathematical model of an attack to modify $\mathcal{G}$ to $\bar{\mathcal{G}}$ is as follows (the notation that a bar is on a variable denotes the value modified by the adversary):

$$
\begin{aligned}
\bar{\mathbf{s}} &= \mathbf{s} + \mathbf{b} \pmod 2, \\
\bar{\mathbf{z}} &= \mathbf{z} + \mathbf{a}(\mathbf{z}), \quad \mathbf{a}(\mathbf{z}) \in \mathcal{A},
\end{aligned}
\tag{3.4}
$$

where $\bar{\mathbf{s}}$ is the modified network data corresponding to $\bar{\mathcal{G}}$, $\mathbf{b} \in \{0, 1\}^d$ represents the modifications on the network data $\mathbf{s}$, $\mathbf{a}(\mathbf{z}) \in \mathbb{R}^m$ denotes the attack vector added to the meter data $\mathbf{z}$, and $\mathcal{A} \subset \mathbb{R}^m$ denotes the subspace of feasible attack vectors.

We assume that the adversary can modify the network data accordingly for any

64

target topology that deems to be valid to the control center. This is the opposite of the assumption employed by most existing studies on *state* attacks where network data that specify the topology are not under attack.

For the attack on analog meter data, we use the notation $\mathbf{a}(\mathbf{z})$ to emphasize that the adversary can design the attack vector based on the whole meter data $\mathbf{z}$. This assumption will be relaxed in Section 3.4 to study an attack with local information. In addition, $\mathcal{A}$ has a form of $\{\mathbf{c} \in \mathbb{R}^m : c_i = 0, \ i \in \mathcal{I}_S\}$ where $\mathcal{I}_S$ is the set of indices of secure meter data entries that the adversary cannot alter and $\{1, \ldots, m\} \setminus \mathcal{I}_S$ represents the adversary-controlled entries. Note that $\mathcal{A}$ fully characterizes the power of the adversary, and the mapping $\mathbf{a} : \mathbb{R}^m \to \mathcal{A}$ fully defines the attack strategy.

3.2.3 State Estimation, Bad Data Test, and Undetectable Attacks

As illustrated in Fig. 3.1, the control center executes generalized state estimation (GSE) [74] with network and meter data as inputs; the inputs are $(\mathbf{s}, \mathbf{z})$ in the absence of an attack and $(\bar{\mathbf{s}}, \bar{\mathbf{z}})$ if there is an attack. GSE regards both network and meter data as possibly erroneous. Once the bad data test detects inconsistency among data and estimates, GSE filters out the outliers from the data and searches for a new *pair* of topology and state estimates that fit the data best. Our focus is on the attacks that can pass the bad data test such that no alarm is raised by GSE.

Under the general AC model (3.1), if $(\mathbf{s}, \mathbf{z})$ is the input to GSE, and $\hat{\mathcal{G}}$ is the topology corresponding to $\mathbf{s}$, the control center obtains the weighted least squares

(WLS) estimate of the state $\mathbf{x}$:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{y}}(\mathbf{z} - h(\mathbf{y}, \hat{\mathcal{G}}))^t \Sigma^{-1}(\mathbf{z} - h(\mathbf{y}, \hat{\mathcal{G}})).$$

Note that $\hat{\mathcal{G}} = \mathcal{G}$ in the absence of an attack while $\hat{\mathcal{G}} = \bar{\mathcal{G}}$ in the presence of an attack. In practice, nonlinear WLS estimation is implemented numerically [51].

Under the DC model (3.2), the WLS state estimator is a linear estimator with a closed form expression

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{y}}(\mathbf{z} - \hat{H}\mathbf{y})^t \Sigma^{-1}(\mathbf{z} - \hat{H}\mathbf{y})$$
$$= (\hat{H}^t \Sigma^{-1}\hat{H})^{-1}\hat{H}^t \Sigma^{-1}\mathbf{z},$$

where $\hat{H}$ is the measurement matrix for $\hat{\mathcal{G}}$. The linear estimator is sometimes used as part of an iterative procedure to obtain the nonlinear WLS solution.

The residue error is often used at the control center for bad data detection [51]. In the so-called $J(\hat{\mathbf{x}})$ test [40], the weighted least squares error

$$J(\hat{\mathbf{x}}) = (\mathbf{z} - h(\hat{\mathbf{x}}, \hat{\mathcal{G}}))^t \Sigma^{-1}(\mathbf{z} - h(\hat{\mathbf{x}}, \hat{\mathcal{G}}))$$

is used in a threshold test:

$$\begin{cases} \text{bad data} & \text{if } J(\hat{\mathbf{x}}) > \tau, \\ \text{good data} & \text{if } J(\hat{\mathbf{x}}) \leq \tau, \end{cases} \qquad (3.5)$$

where τ is the detection threshold, and it is determined to satisfy a certain false alarm constraint α.

We define that an attack is *undetectable* if its detection probability is as low as the false alarm rate of the detector. We assume that the $J(\hat{\mathbf{x}})$ test is used as the bad data detector.

Definition 3.2.1 *An attack* **a** *to modify* $\mathcal{G}$ *to* $\bar{\mathcal{G}}$ *is said to be* undetectable *if, for any true state* **x***, the* $J(\hat{\mathbf{x}})$*-test with any false alarm constraint detects the attack with the detection probability no greater than its false alarm rate.*

In the absence of noise, the only source of bad data is, presumably, an attack. In this case, the probabilistic statement of undetectability becomes a deterministic one. A data attack $(\mathbf{z} + a(\mathbf{z}), \bar{\mathbf{s}})$ that modifies the topology from $\mathcal{G}$ to $\bar{\mathcal{G}}$ is undetectable if for every noiseless measurement **z**, there exists a state vector $\bar{\mathbf{x}}$ such that $\mathbf{z} + a(\mathbf{z}) = h(\bar{\mathbf{x}}, \bar{\mathcal{G}})$. Unfortunately, such a nonlinear condition is difficult to check.

Under the DC model, however, the undetectability condition has a simple algebraic form. Let $(\mathbf{s}, \mathbf{z})$ be the input to GSE and H is the measurement matrix for the topology corresponding to **s**. In the presence of an attack, GSE receives $(\bar{\mathbf{s}}, \bar{\mathbf{z}})$ instead of $(\mathbf{s}, \mathbf{z})$, and $\bar{H}$–the measurement matrix for the target topology $\bar{\mathcal{G}}$–replaces H. In the absence of noise, the $J(\hat{\mathbf{x}})$-detector is equivalent to checking whether the received meter data is in the column space of the valid measurement matrix. Thus, the equivalent undetectable topology attack can be defined by the following easily checkable form:

Definition 3.2.2 *An attack to modify* $\mathcal{G}$ *to* $\bar{\mathcal{G}}$ *with the attack vector* **a** *is said to be* undetectable *if*

$$\mathbf{z} + \mathbf{a}(\mathbf{z}) \in Col(\bar{H}), \quad \forall \mathbf{z} \in Col(H), \tag{3.6}$$

where H and $\bar{H}$ are the measurement matrices for $\mathcal{G}$ and $\bar{\mathcal{G}}$ respectively, and $Col(H)$ is the column space of H and $Col(\bar{H})$ the column space of $\bar{H}$.

3.3 Topology Attack with Global Information

We assume the DC model (3.2) and present the result for the existence of undetectable topology attacks.

3.3.1 Condition for an Undetectable Attack

We first derive a necessary and sufficient algebraic condition for existence of an undetectable attack that modifies $\mathcal{G}$ to $\bar{\mathcal{G}}$ with the subspace $\mathcal{A}$ of feasible attack vectors. To motivate the general result, consider first the noiseless case.

Noiseless Measurement Case

Suppose there is an undetectable attack $\mathbf{a}$ with $\mathbf{a}(\mathbf{z}) \in \mathcal{A}$, $\forall \mathbf{z} \in \mathrm{Col}(H)$. Then, undetectability implies that $\mathbf{z} + \mathbf{a}(\mathbf{z}) \in \mathrm{Col}(\bar{H})$, $\forall \mathbf{z} \in \mathrm{Col}(H)$, and thus, $\mathrm{Col}(H) \subset \mathrm{Col}(\bar{H}, \mathcal{A})$.[3]

Now suppose $\mathrm{Col}(H) \subset \mathrm{Col}(\bar{H}, \mathcal{A})$. There exists a basis $\{\mathbf{c}_1, \ldots, \mathbf{c}_p, \mathbf{d}_1, \ldots, \mathbf{d}_q\}$ of $\mathrm{Col}(\bar{H}, \mathcal{A})$ such that $\{\mathbf{c}_1, \ldots, \mathbf{c}_p\}$ is a subset of columns of $\bar{H}$ and $\{\mathbf{d}_1, \ldots, \mathbf{d}_q\}$ is a set of linearly independent vectors in $\mathcal{A}$. For any $\mathbf{z} \in \mathrm{Col}(H)$, since $\mathrm{Col}(H) \subset \mathrm{Col}(\bar{H}, \mathcal{A})$, there exist unique $(\alpha_i)_{i=1}^p \in \mathbb{R}^p$ and $(\beta_j)_{j=1}^q$ such that $\mathbf{z} = \sum_{i=1}^p \alpha_i \mathbf{c}_i + \sum_{j=1}^q \beta_j \mathbf{d}_j$. If we set $\mathbf{a}(\mathbf{z}) = -\sum_{j=1}^q \beta_j \mathbf{d}_j$, $\mathbf{z} + \mathbf{a}(\mathbf{z}) = \sum_{i=1}^p \alpha_i \mathbf{c}_i \in \mathrm{Col}(\bar{H})$. In addition, $\mathbf{a}(\mathbf{z}) \in \mathcal{A}$ for all $\mathbf{z}$. Hence, there exists an undetectable attack with the subspace $\mathcal{A}$ of feasible attack vectors.

The above arguments lead to the following theorem.

---

[3]$\mathrm{Col}(\bar{H}, \mathcal{A})$ denotes the space spanned by the columns of $\bar{H}$ and a basis of $\mathcal{A}$.

Theorem 3.3.1 *There exists an undetectable attack to modify $\mathcal{G}$ to $\bar{\mathcal{G}}$ with the subspace $\mathcal{A}$ of feasible attack vectors if and only if $Col(H) \subset Col(\bar{H}, \mathcal{A})$.*

Noisy Measurement Case

The following theorem states that the algebraic condition in Theorem 3.3.1 can also be used in the noisy measurement case.

Theorem 3.3.2 *There exists an undetectable attack to modify $\mathcal{G}$ to $\bar{\mathcal{G}}$ with the subspace $\mathcal{A}$ of feasible attack vectors if and only if $Col(H) \subset Col(\bar{H}, \mathcal{A})$.*

In addition, if an attack $\mathbf{a}$ is such that $Col(H) \nsubseteq Col(\bar{H}, \mathcal{A})$, then for almost every[4] $\mathbf{x} \in \mathbb{R}^n$, when $\mathbf{x}$ is the true state, the detection probability for the attack approaches 1 as the noise variances uniformly decrease to 0 (i.e., $\max_i(\Sigma_{ii})$, where Σ_{ii} is the (i,i) entry of Σ, decays to 0).

Proof: See Section 3.7. ∎

Note that when the algebraic condition is not met, the attack can be detected with high probability if the noise variances are sufficiently small. With this algebraic condition, we can check whether the adversary can launch an undetectable attack with $\mathcal{A}$ for the target $\bar{\mathcal{G}}$. The condition will be used in Section 3.5 to construct a meter protection strategy to disable undetectable attacks for any target topology.

By finding the smallest dimension of $\mathcal{A}$ satisfying the condition, we can also characterize the minimum cost of undetectable attacks for $\bar{\mathcal{G}}$; in the adversary's

---

[4]This means "for all $\mathbf{x} \in \mathbb{R}^n \setminus \mathcal{S}$, for some $\mathcal{S} \subset \mathbb{R}^n$ with a zero Lebesgue measure".

point of view, a smaller dimension of $\mathcal{A}$ is preferred, because increasing the dimension of $\mathcal{A}$ necessitates compromising more RTUs or communication devices. In the following section, we present an undetectable attack requiring a small number of data modifications and prove its optimality for a class of targets by utilizing the algebraic condition.

3.3.2   State-preserving Attack

This section presents a simple undetectable attack, referred to as *state-preserving attack*. As the name suggests, the attack intentionally preserves the state in order to have a sparse attack vector. We again motivate our result by considering first the noiseless case.

Noiseless Measurement Case

Given $\mathbf{z} = H\mathbf{x} \in \mathrm{Col}(H)$, the state-preserving attack sets $\mathbf{a}(\mathbf{z})$ equal to $(\bar{H} - H)\mathbf{x}$. Then, $\mathbf{z} + \mathbf{a}(\mathbf{z}) = \bar{H}\mathbf{x} \in \mathrm{Col}(\bar{H})$; the attack is *undetectable*. Note that the state $\mathbf{x}$ remains the same after the attack. Since H has full column rank, $\mathbf{a}(\mathbf{z})$ can be simply calculated as

$$\mathbf{a}(\mathbf{z}) = (\bar{H} - H)\mathbf{x} = (\bar{H} - H)(H^t H)^{-1} H^t \mathbf{z}. \tag{3.7}$$

For $\mathbf{a}(\mathbf{z})$ above to be a valid attack vector, it is necessary to be a sparse vector constrained by the meters, the data of which can be altered by the adversary.

To see an intuitive reason why $\bar{H}\mathbf{x} - H\mathbf{x}$ is sparse, consider the simple case that a line is removed from the topology while the state is *preserved*. In this case, the line flows through all the lines, except the removed line, stay the same. Because,

the line flow from i to j is determined by (i) (x_i, x_j) and (ii) whether i and j are connected, and for most lines, these two factors remain the same. Hence, only few entries are different between $\bar{H}\mathbf{x}$ and $H\mathbf{x}$. Below, we will show that, for all state $\mathbf{x} \in \mathbb{R}^n$, all entries of $(\bar{H} - H)\mathbf{x}$ are zeros except those associated with the target lines.

As noted in [26], H can be decomposed as $H = MBA^t$, where $M \in \mathbb{R}^{m \times l}$ is the measurement-to-line incidence matrix with $l \triangleq |\mathcal{E}_0|$, $B \in \mathbb{R}^{l \times l}$ is a diagonal matrix with the line susceptances in the diagonal entries, and $A^t \in \mathbb{R}^{l \times n}$ is the line-to-bus incidence matrix. Each column of M (each row of A^t) corresponds to a distinct line in $\mathcal{E}_0$. For $1 \leq j \leq l$, if the jth column of M corresponds to $(a, b) \in \mathcal{E}_0$, let $v_j^+ \triangleq a$ and $v_j^- \triangleq b$. Then, M is defined such that $M_{ij} = \pm 1$ if the ith meter (the meter corresponding to the ith row of M) measures (i) the line flow from $v_j^{\pm}$ to $v_j^{\mp}$ or (ii) the injection at bus $v_j^{\pm}$; otherwise, $M_{ij} = 0$. For A^t, $(A^t)_{ji} = \pm 1$ if $v_j^{\pm} = i$, and the line corresponding to the jth row of A^t (or equivalently the jth column of M) is *connected* in $\mathcal{G}$; otherwise, $(A^t)_{ji} = 0$. Note that M and B are independent of the topology, but A^t does depend on $\mathcal{G}$. Fig. 3.2 provides an example to illustrate the structures of M, B, and A^t. Similarly, $\bar{H}$ is decomposed as $\bar{H} = MB\bar{A}^t$.

As illustrated in Fig. 3.2, the entries of $BA^t\mathbf{x} \in \mathbb{R}^{l \times 1}$ correspond to the line flows of all the lines in $\mathcal{E}_0$ when the state is $\mathbf{x}$ and the topology is $\mathcal{G}$. Similarly, $B\bar{A}^t\mathbf{x}$ is the vector of line flows when the state is $\mathbf{x}$ and the topology is $\bar{\mathcal{G}}$. Since the states are the same, the kth entry of $BA^t\mathbf{x}$ and that of $B\bar{A}^t\mathbf{x}$ are different only if the corresponding line is connected in one of $\mathcal{G}$ and $\bar{\mathcal{G}}$ while disconnected in the other. Therefore, $(B\bar{A}^t - BA^t)\mathbf{x}$ has all zero entries except the entries corresponding to the lines in $\bar{\mathcal{E}}\triangle\mathcal{E}$. Specifically, the entry corresponding to $(i, j) \in \bar{\mathcal{E}} \setminus \mathcal{E}$ assumes $f_{ij}(\mathbf{x}) \triangleq B_{ij}(x_i - x_j)$, and the entry corresponding to $(i, j) \in \mathcal{E} \setminus \bar{\mathcal{E}}$ assumes $-f_{ij}(\mathbf{x})$.

$M = \begin{bmatrix} 0 & 1 & 1 & -1 & 0 \\ 0 & 0 & -1 & 0 & -1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{matrix} \bar{2} \\ \bar{4} \\ \overline{(1,3)} \\ \overline{(2,1)} \\ \overline{(2,4)} \\ \overline{(3,2)} \\ \overline{(3,4)} \end{matrix}$

(columns labeled: (1,3) (2,1) (2,4) (3,2) (3,4))

$A^t = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & -1 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ 0 & 1 & -1 \end{bmatrix} \begin{matrix} (1,3) \\ (2,1) \\ (2,4) \\ (3,2) \\ (3,4) \end{matrix}$

(columns labeled: 2 3 4)

$B = \mathrm{diag}(B_{13}, B_{21}, B_{24}, B_{32}, B_{34})$

$BA^t\mathbf{x} = \begin{bmatrix} B_{13}(-x_3) \\ B_{21}(x_2) \\ B_{24}(x_2 - x_4) \\ \mathbf{0} \\ B_{34}(x_3-x_4) \end{bmatrix}$

$B\bar{A}^t\mathbf{x} = \begin{bmatrix} B_{13}(-x_3) \\ B_{21}(x_2) \\ B_{24}(x_2 - x_4) \\ \mathbf{B_{32}(x_3-x_2)} \\ B_{34}(x_3-x_4) \end{bmatrix}$
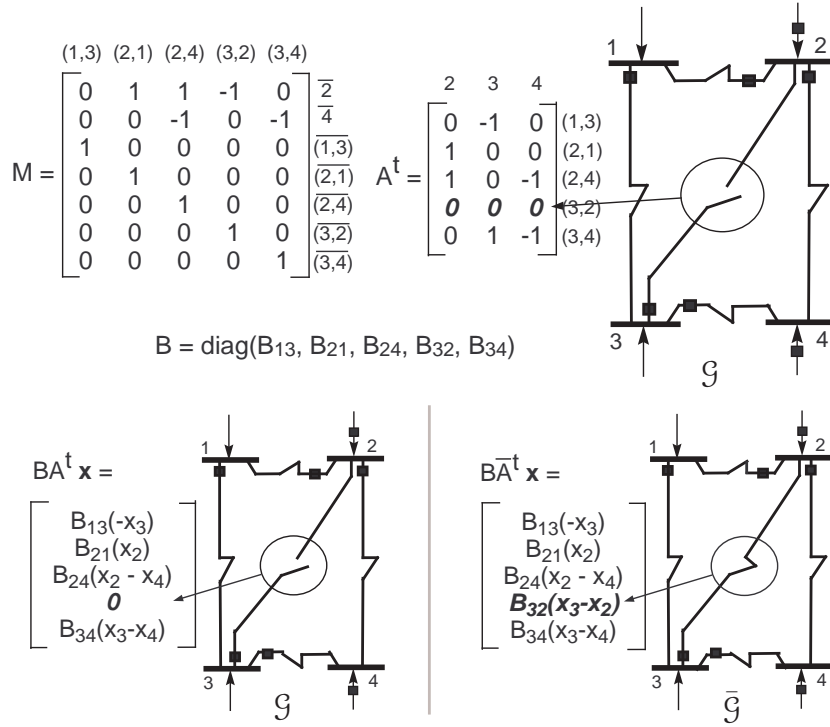
Figure 3.2: The measurement, line, or bus corresponding to each row or column is labled. Bus 1 is the slack bus. For the rows of M, $\bar{i}$ denotes the injection meter at bus i, and $\overline{(i,j)}$ the meter for the line flow from i to j.

Hence, $(\bar{H} - H)\mathbf{x} = M(B\bar{A}^t - BA^t)\mathbf{x}$ is equal to

$$\sum_{(i,j)\in\bar{\mathcal{E}}\setminus\mathcal{E}} f_{ij}(\mathbf{x})\mathbf{m}_{(i,j)} - \sum_{(i,j)\in\mathcal{E}\setminus\bar{\mathcal{E}}} f_{ij}(\mathbf{x})\mathbf{m}_{(i,j)} \tag{3.8}$$

where $\mathbf{m}_{(i,j)}$ is the column vector of M corresponding to (i,j). Note that $\mathbf{m}_{(i,j)}$ is a sparse vector that has nonzero entries only at the rows corresponding to the line flow meters on the line (i,j) and the injection meters at i and j.

From (3.8), for any state $\mathbf{x} \in \mathbb{R}^n$, $(\bar{H} - H)\mathbf{x}$ is a linear combination of elements in $\{\mathbf{m}_{(i,j)} : (i,j) \in \bar{\mathcal{E}}\triangle\mathcal{E}\}$. Hence, the state-preserving attack, which sets $\mathbf{a(z)} = (\bar{H}-H)\mathbf{x}$, modifies at most the line flow meters on the target lines and the injection meters at the target buses.

We now show in the next two theorems that, under certain conditions, the state-

preserving attack has the least cost in the sense that it requires the adversary to modify the smallest number of meter data (*i.e.*, the smallest dimension for $\mathcal{A}$).

Theorem 3.3.3 *Assume that (i) the actual and target topologies differ by only one line, i.e., $|\bar{\mathcal{E}} \triangle \mathcal{E}| = 1$, and (ii) every line in $\bar{\mathcal{E}}$, incident[5] from or to any target bus with an injection meter, has at least one line flow meter on it. Then, among all undetectable attacks, the state-preserving attack modifies the smallest number of meters, which is the total number of line flow and injection meters located on the target line and target buses.*

Proof: *See Section 3.7.* ∎

Another scenario that the state-preserving attack has the minimum cost is when the adversary aims to delete lines from the actual topology.

Theorem 3.3.4 *Let $\mathcal{G}^*$ and $\bar{\mathcal{G}}^*$ denote the undirected versions of $\mathcal{G}$ and $\bar{\mathcal{G}}$ respectively. Suppose that the adversary aims to remove lines from $\mathcal{G}$, i.e., $\bar{\mathcal{E}} \subsetneq \mathcal{E}$, and the following hold:*

> *Every line in $\bar{\mathcal{E}}$, incident from or to a target bus with an injection meter, has at least one line flow meter on it.*
>
> *In $\mathcal{G}^*$, target lines do not form a closed path.*
>
> *$\bar{\mathcal{G}}^*$ does* not *include a tree $\mathcal{T}$ satisfying the following:*
>
> *1) (number of nodes in $\mathcal{T}$) ≥ 4, and*
>
> *2) every node in $\mathcal{T}$ is a target bus with an injection meter.*

---

[5]A line (i, j) is said to be incident from i and incident to j.

Then, among all undetectable attacks, the state-preserving attack modifies the smallest number of meters, which is the total number of line flow and injection meters located on the target lines and target buses.

Proof: *See Section 3.7.* ∎

Roughly speaking, the assumptions in Theorem 3.3.4 hold when target lines are far from each other such that there is no big tree in $\bar{\mathcal{G}}$ consisting solely of target buses.

The main advantage of the state-preserving attack is that by preserving the system state during the attack, the attack can be launched by perturbing only *local* meters around the target lines; hence, only few data entries need to be modified. Theorem 3.3.3 and Theorem 3.3.4 supports the claim by stating the optimality of the state-preserving attack under the mild assumptions. The theorems also imply that the minimum cost of an undetectable attack can be easily characterized if the target topology satisfies the theorem assumptions.

Noisy Measurement Case

Following the intuition behind the state-preserving attack in the noiseless case, we will construct its counterpart for the noisy measurement case. Recall the relation (3.8):

$$(\bar{H} - H)\mathbf{x} = \sum_{(i,j) \in \bar{\mathcal{E}} \setminus \mathcal{E}} f_{ij}(\mathbf{x})\mathbf{m}_{(i,j)} - \sum_{(i,j) \in \mathcal{E} \setminus \bar{\mathcal{E}}} f_{ij}(\mathbf{x})\mathbf{m}_{(i,j)}.$$

The above implies that

$$(\bar{H} - H)\mathbf{x} \in \mathcal{M} \triangleq \text{span}\{\mathbf{m}_{(i,j)} : (i,j) \in \bar{\mathcal{E}} \triangle \mathcal{E}\} \tag{3.9}$$

We set $\mathbf{a}(\mathbf{z})$ as a minimizer of the $J(\hat{\mathbf{x}})$-test statistic[6]:

$$\mathbf{a}(\mathbf{z}) \triangleq \arg\min_{\mathbf{d}\in\mathcal{M}} \|(\mathbf{z}+\mathbf{d}) - \bar{H}\hat{\mathbf{x}}_{\mathrm{WLS}}[\mathbf{z}+\mathbf{d}]\|^2_{\Sigma^{-1}} \tag{3.10}$$

where $\hat{\mathbf{x}}_{\mathrm{WLS}}[\mathbf{z}+\mathbf{d}]$ denotes the WLS state estimate when the topology estimate is $\bar{\mathcal{G}}$, and $\mathbf{z}+\mathbf{d}$ is observed at the control center. Note that, since $\mathbf{a}(\mathbf{z}) \in \mathcal{M}$, the attack with $\mathbf{a}$ modifies at most the line flow measurements of the target lines and the injection measurements of the target buses.

Now, suppose that the adversary modifies breaker state measurements such that the topology estimate becomes $\bar{\mathcal{G}}$ and simultaneously modifies the meter data with $\mathbf{a}(\mathbf{z})$. Then, the $J(\hat{\mathbf{x}})$-test statistic at the control center is upper bounded as

$$\|(\mathbf{z}+\mathbf{a}(\mathbf{z})) - \bar{H}\hat{\mathbf{x}}_{\mathrm{WLS}}[\mathbf{z}+\mathbf{a}(\mathbf{z})]\|^2_{\Sigma^{-1}}$$

$$\leq \|(\bar{H}\mathbf{x}+\mathbf{e}) - \bar{H}\hat{\mathbf{x}}_{\mathrm{WLS}}[\bar{H}\mathbf{x}+\mathbf{e}]\|^2_{\Sigma^{-1}},$$

because $(\bar{H}-H)\mathbf{x}$ is an element of $\mathcal{M}$. Note that the right hand side is the $J(\hat{\mathbf{x}})$-test statistic when the meter data are consistent with the topology estimate $\bar{\mathcal{G}}$. Hence, it has χ^2_{m-n} distribution, the same as the distribution of the $J(\hat{\mathbf{x}})$-test statistic under the absence of bad data [40]. This argument leads to the following theorem stating that this attack is undetectable.

Theorem 3.3.5 *The state-preserving attack* $\mathbf{a}$, *defined in (3.10), is undetectable.*

Note that $\hat{\mathbf{x}}_{\mathrm{WLS}}[\mathbf{z}+\mathbf{d}]$ in (3.10) is a linear function of $\mathbf{z}+\mathbf{d}$, so $\mathbf{a}(\mathbf{z})$ can be obtained as a linear weighted least squares solution. Specifically, $\mathbf{a}(\mathbf{z})$ has a form of $\mathbf{a}(\mathbf{z}) = D\mathbf{z}$ where $D \in \mathbb{R}^{m\times m}$ depends on $\mathcal{G}$, $\bar{\mathcal{G}}$, and Σ, but not on $\mathbf{z}$. Hence, D can be obtained off-line before observing $\mathbf{z}$.

---

[6]We use $\|\mathbf{r}\|^2_{\Sigma^{-1}}$ to denote the quadratic form $\mathbf{r}^t\Sigma^{-1}\mathbf{r}$.

Note also that the state-preserving attacks in the noiseless and noisy cases modify the same set of meters. In addition, recall that the condition for existence of an undetectable attack is the same for both noiseless and noisy cases. The optimality statements for the state-preserving attack in Theorem 3.3.3 and Theorem 3.3.4 were derived purely based on the condition for undetectability. Hence, the same optimality statements hold for the noisy measurement case, as stated in the following corollary, and the same interpretation can be made.

Corollary 3.3.5.1 *For the noisy measurement DC model, suppose that the condition in Theorem 3.3.3 or the condition in Theorem 3.3.4 hold. Then, among all undetectable attacks, the state-preserving attack modifies the smallest number of meters, which is the total number of line flow and injection meters located on the target lines and target buses.*

3.4 Topology Attack with Local Information

In this section, we consider the more realistic scenario of a weak attacker who does not have the measurement data of the entire network; it only has access to a few meters. The information available to the adversary is local. We also generalize the linear (DC) measurement model to the nonlinear (AC) model. The resulting undetectable attacks, however, are limited to line removal attacks, *i.e.*, the adversary only tries to remove lines from the actual network topology.

We first consider the noiseless measurement case under the DC model. Since we are restricted to line-removal attacks, $\bar{\mathcal{E}}$ is a strict subset of $\mathcal{E}$. Therefore, recalling

(3.8), we have

$$(\bar{H} - H)\mathbf{x} = -\sum_{(i,j)\in\mathcal{E}\setminus\bar{\mathcal{E}}} f_{ij}(\mathbf{x})\mathbf{m}_{(i,j)} \tag{3.11}$$

where $f_{ij}(\mathbf{x})$, as defined in Section 3.3, denotes the line flow from i to j when the line is connected, and the state is $\mathbf{x}$.

Let z_{ij} denote the measurement of the line flow from i to j. Due to the absence of noise, $z_{ij} = f_{ij}(\mathbf{x}) = -f_{ji}(\mathbf{x}) = -z_{ji}$. With this observation and (3.11), we have

$$(\bar{H} - H)\mathbf{x} = -\sum_{(i,j)\in\mathcal{E}\setminus\bar{\mathcal{E}}} z_{ij}\mathbf{m}_{(i,j)} \tag{3.12}$$

Therefore, setting $\mathbf{a}(\mathbf{z}) = (\bar{H} - H)\mathbf{x}$, which is the state-preserving attack, is *equivalent* to setting

$$\mathbf{a}(\mathbf{z}) = -\sum_{(i,j)\in\mathcal{E}\setminus\bar{\mathcal{E}}} z_{ij}\mathbf{m}_{(i,j)} \tag{3.13}$$

From (3.13), one can see that adding the above $\mathbf{a}(\mathbf{z})$ to $\mathbf{z}$ is equivalent to the following heuristic described in Fig. 3.3:

1. For every target line (i, j), subtract z_{ij} and z_{ji} from the injection measurements at i and j respectively.

2. For every target line (i, j), modify z_{ij} and z_{ji} to 0.

This heuristic simply forces the line flows through the target lines, which are disconnected in $\bar{\mathcal{G}}$, to be zeros, while adjusting the injections at the target buses to satisfy the power balance equations [51]. If a target line (i, j) has only one line flow meter (*e.g.*, z_{ji}), we can use $-z_{ji}$ in the place of z_{ij}. But, if some target line has no line flow meter, this heuristic is not applicable. Note that the heuristic only requires the ability to observe and modify the line flow measurements of the target lines and the injection measurements at the target buses. The adversary
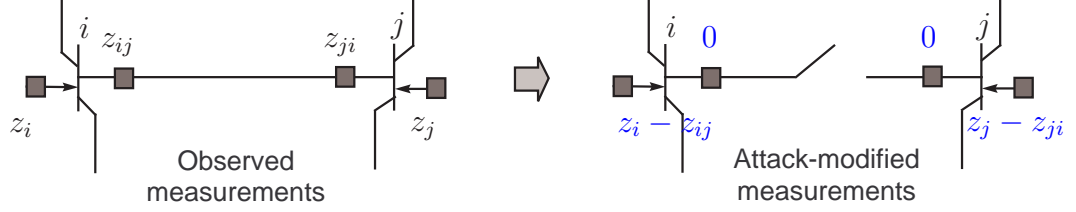
Figure 3.3: Heuristic Operations Around the Target Line (i, j)

can launch it without knowing the topology or network parameters (*i.e.*, H and $\bar{H}$ are not necessary). Since the heuristic is equivalent to the state-preserving attack, it is undetectable.

The same heuristic is applicable to the noisy measurements $\mathbf{z} = H\mathbf{x} + \mathbf{e}$. To avoid detection, the adversary can make $\mathbf{a}(\mathbf{z})$ approximate $\bar{H}\mathbf{x} - H\mathbf{x}$ such that $\mathbf{z} + \mathbf{a}(\mathbf{z})$ is close to $\bar{H}\mathbf{x} + \mathbf{e}$. Because $z_{ij} = f_{ij}(\mathbf{x}) + e_{ij}$, z_{ij} is an unbiased estimate of $f_{ij}(\mathbf{x})$. Similarly, $-\sum_{(i,j)\in\mathcal{E}\backslash\bar{\mathcal{E}}} z_{ij}\mathbf{m}_{(i,j)}$ is an unbiased estimate of $-\sum_{(i,j)\in\mathcal{E}\backslash\bar{\mathcal{E}}} f_{ij}(\mathbf{x})\mathbf{m}_{(i,j)}$, which is equal to $\bar{H}\mathbf{x} - H\mathbf{x}$. Hence, it is reasonable to set $\mathbf{a}(\mathbf{z}) = -\sum_{(i,j)\in\mathcal{E}\backslash\bar{\mathcal{E}}} z_{ij}\mathbf{m}_{(i,j)}$ even in the noisy measurement case.

The same idea is applicable to the AC power flow model with the nonlinear state estimator. Suppose that $\mathbf{z}$ is the real power measurement from the AC power flow model: $\mathbf{z} = h(\mathbf{x}) + \mathbf{e}$, where $\mathbf{x}$ is the vector of the voltage phasors at all buses, and h is the nonlinear measurement function for $\mathcal{G}$. Let $\bar{h}$ denote the measurement function for $\bar{\mathcal{G}}$. If $\mathbf{a}(\mathbf{z})$ is equal to $\bar{h}(\mathbf{x}) - h(\mathbf{x})$,

$$\bar{\mathbf{z}} = (h(\mathbf{x}) + \mathbf{e}) + \mathbf{a}(\mathbf{z}) = \bar{h}(\mathbf{x}) + \mathbf{e}, \tag{3.14}$$

which is consistent with $\bar{\mathcal{G}}$, so the attack cannot be detected. We will show that the attack vector of the heuristic approximates $\bar{h}(\mathbf{x}) - h(\mathbf{x})$.

For simplicity, assume that the attacker aims at removing a single line (i, j)

from $\mathcal{G}$. Then, $h(\mathbf{x})$ and $\bar{h}(\mathbf{x})$ are different only in the entries corresponding to the injections at i and j and the line flows through (i, j). Specifically, $\bar{h}(\mathbf{x}) - h(\mathbf{x})$ has all zero entries except $-h_{ij}(\mathbf{x})$ at the rows corresponding to the injection at i and the line flow from i to j, and $-h_{ji}(\mathbf{x})$ at the rows corresponding to the injection at j and the line flow from j to i, where $h_{ij}(\mathbf{x})$ denotes the entry of $h(\mathbf{x})$ corresponding to the line flow from i to j. Since $z_{ij} = h_{ij}(\mathbf{x}) + e_{ij}$ and $z_{ji} = h_{ji}(\mathbf{x}) + e_{ji}$, z_{ij} and z_{ji} can be considered as unbiased estimates of $h_{ij}(\mathbf{x})$ and $h_{ji}(\mathbf{x})$ respectively. Hence, the attacker can use z_{ij} and z_{ji} to construct an unbiased estimate of $\bar{h}(\mathbf{x}) - h(\mathbf{x})$. Adding this estimate to $\mathbf{z}$ is equivalent to the heuristic operation of Fig. 3.3, which subtracts z_{ij} and z_{ji} from z_i and z_j respectively, and sets z_{ij} and z_{ji} to zeros. The same argument holds for the reactive measurement part and multiple-line removal attacks. In practice, the heuristic attack should be executed twice separately, once for real measurements and second for reactive measurements. In Section 3.6, numerical simulations demonstrate that the heuristic attack on the AC power flow model with the nonlinear state estimation has a very low detection probability.

3.5 Countermeasure for Topology Attacks

In this section, we consider countermeasures that prevent attacks by a strong adversary with global information. In particular, we assume that a subset of meters can be secured so that the adversary cannot modify data from these meters. In practice, this can be accomplished by implementing more sophisticated authentication protocols. We present a so-called cover-up protection that identifies the set of meters that need to be secured.

The algebraic condition in Theorems 3.3.1-3.3.2 provides a way to check

whether a set of adversary-controlled meters is enough to launch an undetectable attack. Restating the algebraic condition, there exists an undetectable attack with the subspace $\mathcal{A}$ of feasible attack vectors, if and only if $\mathrm{Col}(H) \subset \mathrm{Col}(\bar{H}, \mathcal{A})$ for some $\bar{\mathcal{G}}$ (different from $\mathcal{G}$).

Let $\mathcal{I}_S$ denote the set of indices for the entries of $\mathbf{z}$ corresponding to the protected meters. Then, $\mathcal{A}$ is $\{\mathbf{c} \in \mathbb{R}^m : c_i = 0, \ i \in \mathcal{I}_S\}$. The objective of the control center is to make any undetectable attack infeasible while minimizing the cost of protection (*i.e.*, minimizing $|\mathcal{I}_S|$ or equivalently, maximizing the dimension of $\mathcal{A}$).

To achieve the protection goal, $\mathcal{A}$ should satisfy that for any target topology $\bar{\mathcal{G}}$, $\mathrm{Col}(H) \not\subset \mathrm{Col}(\bar{H}, \mathcal{A})$. However, finding such $\mathcal{A}$ by checking the conditions for all possible targets is computationally infeasible. To avoid computational burden, the following theorem gives a simple graph-theoretical strategy.

Theorem 3.5.1 (Cover-up strategy) *Let $\tilde{\mathcal{E}}$ and $\tilde{\mathcal{E}}_0$ denote the undirected counterparts of $\mathcal{E}$ and $\mathcal{E}_0$ respectively. For $i \in \mathcal{V}$, let $\mathcal{L}_i$ denote the set of edges in $(\mathcal{V}, \tilde{\mathcal{E}}_0)$ that are incident to i.*

Suppose there is a spanning tree $\mathcal{T} = (\mathcal{V}, \mathcal{E}_{\mathcal{T}})$ of $(\mathcal{V}, \tilde{\mathcal{E}})$ (the current topology) and a vertex subset $\mathcal{B}$ ($\mathcal{B} \subset \mathcal{V}$) that satisfies

$$\mathcal{E}_{\mathcal{T}} \cup (\cup_{b \in \mathcal{B}} \mathcal{L}_b) = \tilde{\mathcal{E}}_0. \tag{3.15}$$

Then, if we protect (i) one line flow meter for each line in $\mathcal{E}_{\mathcal{T}}$ and (ii) the injection meters at all buses in $\mathcal{B}$, an undetectable attack does not exist for any target topology.

Proof: *See Section 3.7.* ∎

The condition (3.15) means that the edges of $\mathcal{T}$ and the edges incident to vertices in $\mathcal{B}$ can cover all the lines (both connected and disconnected) of the grid. One can easily find such $\mathcal{T}$ and $\mathcal{B}$ using available graph algorithms.

Fig. 3.4 describes a cover-up strategy for IEEE 14-bus system. The strategy used the spanning tree $\mathcal{T}$ marked by red dash lines, and $\mathcal{B} = \{1, 4, 13\}$. The unprotected meters and protected meters are marked by black rectangles and blue circles respectively. In this example, the strategy requires protection of 30% of meters. In addition, numerically checking the algebraic condition showed that if the control center removes *any* of the protections, the grid becomes vulnerable to undetectable topology attacks. This suggests that the strategy does not require protection of an excessive number of meters. For IEEE 118-bus system, a cover-up strategy required protection of 31% of meters.

The cover-up strategy also prevents undetectable state attacks [18]. It follows from Theorem 1 in [24], which states that an undetectable state attack does not exist if and only if the secure meters, protected by the control center, make the system state observable. Because the strategy protects one line meter for each line in the spanning tree $\mathcal{T}$, the system state is always observable with the protected meters [26].

3.6   Numerical Results

We first present practical uses of the algebraic condition for undetectable attacks. Then, we test the proposed attacks with IEEE 14-bus and 118-bus systems, and present their effect on real-time LMPs.
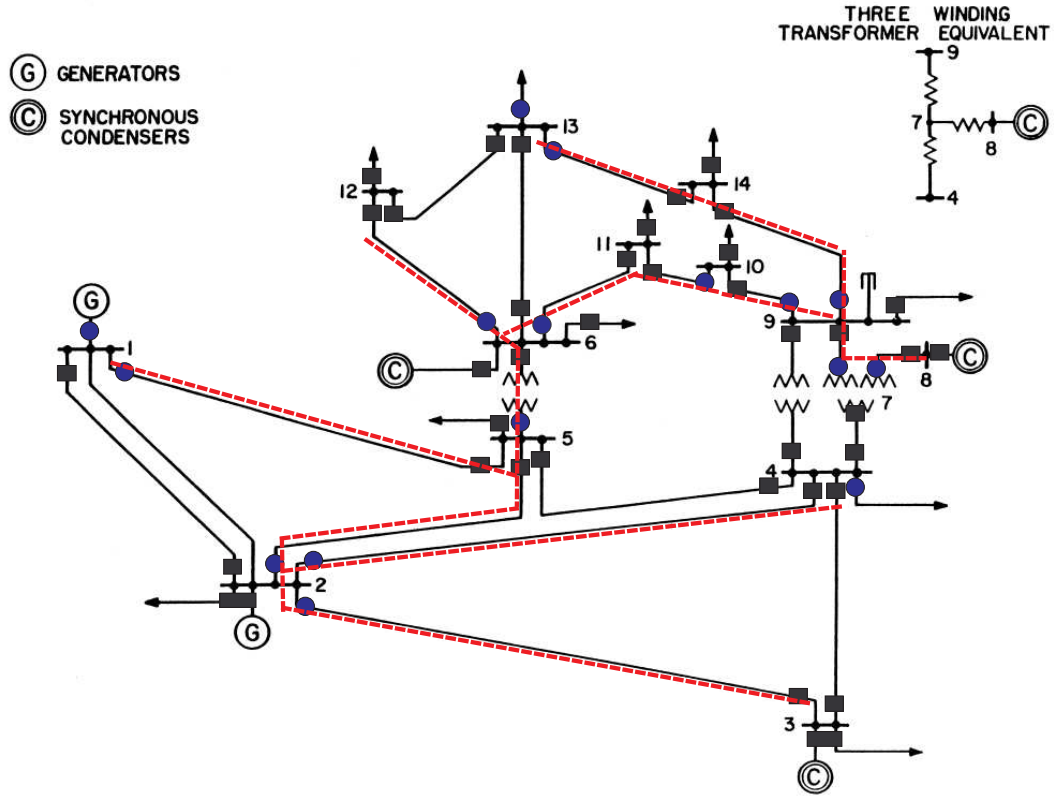
Figure 3.4: Rectangles (or circles) on buses and lines represent injection meters and line flow meters respectively. We assume that $\mathcal{E} = \mathcal{E}_0$. The attacker may attempt to remove lines from $\mathcal{G}$.

3.6.1   Application of Undetectability Condition

In Section 3.3.1, the necessary and sufficient algebraic condition is given to check whether an adversary can launch an undetectable attack for a target $\bar{\mathcal{G}}$ with a subspace $\mathcal{A}$ of feasible attack vectors. Here, we provide examples of how the condition can be used by both attackers and the control center.

Suppose that an attacker with global information aims to remove a specific set of lines from the topology. In Section 3.3.1, we have shown that the state-preserving attack requires the smallest dimension of $\mathcal{A}$ among undetectable attacks

Table 3.1: The adversary-controlled meters for the attacks to remove lines $(2, 4)$ and $(12, 13)$: $i \to j$ denotes the meter for the line flow from bus i to bus j. i denotes the injection meter at bus i.

	Adversary-controlled meters
State-preserving attack	$2 \to 4$, $4 \to 2$, $12 \to 13$, $13 \to 12$, 2, 4, 12, 13
Alternative 1 (not modifying 12)	$2 \to 4$, $4 \to 2$, $12 \to 13$, $13 \to 12$, $6 \to 12$, $12 \to 6$, 2, 4, 6, 13
Alternative 2 (not modifying 4)	$2 \to 4$, $4 \to 2$, $12 \to 13$, $13 \to 12$, $2 \to 3$, $3 \to 2$, $3 \to 4$, $4 \to 3$, 2, 3, 12, 13

under mild conditions. If the conditions are met and the attacker can perform the necessary meter modifications, the state-preserving attack can be launched with the guaranteed optimality. However, if the attacker cannot perform some meter modification required by the state-preserving attack, it should search for an undetectable alternative with a reasonably small dimension for $\mathcal{A}$. The algebraic condition can be used to find such an alternative[7]. For instance, for a line-removal attack on the IEEE 14-bus network in Fig. 3.4, Table 3.1 shows some alternatives to the state-preserving attack when the attacker cannot modify some injection meter.

When the set of adversary-controlled meters is fixed, the algebraic condition can be exploited to find the target topologies, for which the attacker can launch undetectable attacks. For instance, in the IEEE 14-bus network in Fig. 3.4, assume that the attacker can modify the data from the injection meters at 11, 12, and

---

[7]One heuristic way to find an alternative, which we employed, is to begin with a large set $\mathcal{K}$ of adversary-controlled meters that satisfies the algebraic condition and the constraint (*e.g.*, exclude a certain injection meter) and remove meters from $\mathcal{K}$ one by one such that after each removal of a meter, $\mathcal{K}$ still satisfies the algebraic condition. If no more meter can be removed, we take $\mathcal{K}$ as an alternative. The final set depends on the initial $\mathcal{K}$ and the sequence of removed elements. One can try this procedure multiple times with different initial $\mathcal{K}$s and removal sequences, and pick the one with the smallest size.

Table 3.2: The Sets of Lines Undetectable Attacks Can Remove

$|\bar{\mathcal{E}}\triangle\mathcal{E}|$	$\bar{\mathcal{E}}\triangle\mathcal{E}$ (lines to be removed by the attack)
1	$\{(6,12)\}$, $\{(6,11)\}$, $\{(10,11)\}$, $\{(9,10)\}$, $\{(9,14)\}$, $\{(13,14)\}$, $\{(12,13)\}$
2	$\{(10,11),(13,14)\}$, $\{(9,14),(12,13)\}$, $\{(9,10),(13,14)\}$, $\{(6,12),(13,14)\}$, $\{(6,12),(10,11)\}$, $\{(6,12),(9,10)\}$, $\{(6,11),(12,13)\}$, $\{(6,11),(9,14)\}$
3	$\{(6,11),(9,14),(12,13)\}$, $\{(6,12),(9,10),(13,14)\}$, $\{(6,12),(10,11),(13,14)\}$

14, and all the line flow meters on $(6,12)$, $(6,11)$, $(10,11)$, $(9,10)$, $(9,14)$, and $(13,14)$. Then, numerically checking the algebraic condition show that the attacker cannot launch an undetectable attack for any target. However, if the attacker can additionally control the line flow meters on $(12,13)$, it can launch an undetectable attack to remove any set of lines listed in Table 3.2 from the current topology.

The control center can also utilize the algebraic condition to decide which meters to put more security measures on. For instance, in the IEEE 14-bus network, suppose that the control center protects all the injection meter. In the worst case, the attacker may be able to modify all the line flow measurements. In this case, checking the algebraic condition shows that the attacker can launch an undetectable line-removal attack for any target topology, as long as the system with the target topology is observable. However, checking the algebraic condition also shows that if the control center can additionally protect any line flow meter, an undetectable attack does not exist for any target. Therefore, it is worthwhile for the control center to make an effort to secure one more line flow meter.

3.6.2 Undetectability and Effects on Real-time LMP

We tested the state-preserving attack with global information and the heuristic with local information on IEEE 14-bus and IEEE 118-bus system, and investigated their effect on real-time LMPs. The AC power flow model and nonlinear state estimation were used to emulate the real-world power grid.

For simulations, we first assigned the line capacities, generation limits, and estimated loads, and obtained the day-ahead dispatch. Then, we modeled the voltage magnitudes and phases of buses as Gaussian random variables centered at the system state for the day-ahead dispatch, with small variances. In each Monte Carlo run, we generated a state vector from the distribution and used the nonlinear AC power flow model[8] with Gaussian measurement noise to generate the noisy measurements. The attacker observed the noisy measurements, added the corresponding attack vector to them, and passed the corrupt measurements to the control center. The control center employed the nonlinear state estimator to obtain the residue and performed the $J(\hat{\mathbf{x}})$-test with the residue. If $J(\hat{\mathbf{x}})$-test failed to detect the attack, the real-time LMPs were calculated based on the state estimate.

In simulations, we assumed that the attacker aims to remove a single line from the topology. Fig. 3.5 presents the detection probability of the proposed attacks on IEEE 14-bus system, for different target lines. The attacks on most target lines succeeded with low detection probabilities, close to the false alarm constraint 0.1. Table 3.3 shows the detection probability averaged over all possible single-line

---

[8]In simulations, we have reactive measurements, which were not considered in our analysis of the state-preserving attack. We simply applied the same analysis for the reactive components of the linearlized decoupled model [51] and derived the reactive counterpart of the state-preserving attack.
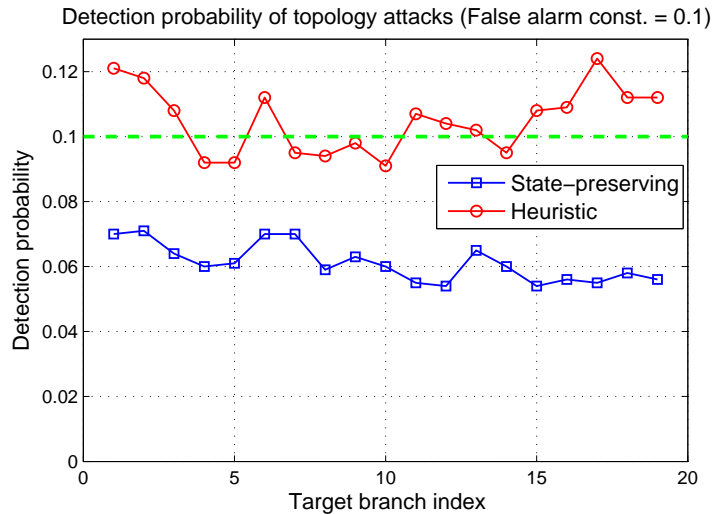
Figure 3.5: The x-axis is for the index of the target line. Measurement noise standard deviation is 0.5 p.u., and 1000 Monte Carlo runs are used.

Table 3.3: 1000 Monte Carlo runs are used.

	14-bus		118-bus	
false alarm const. α	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.01$
state-preserving	0.061	0.009	0.075	0.005
heuristic	0.105	0.019	0.095	0.009

removal attacks. In both IEEE 14-bus and 118-bus systems, the proposed attacks were hardly detected. In most cases, detection probabilities were as low as the false alarm rates. The performance of the heuristic was remarkably good, considering that it only requires to observe and control few local data.

We also examined the absolute perturbation of the real-time LMPs (see [36] for real-time LMP). The parameters in the real-time LMP calculation include the estimated set of congested lines and the shift-factor matrix; both depend on the topology estimate. Hence, we expect that topology attacks would disturb the real-time LMP calculation. In our simulations, both the state-preserving attack and the heuristic perturbed the real-time LMPs by 10% on average for IEEE 14-bus

system and 3.3% for IEEE 118-bus system. In the 118-bus system, attacks on some target lines had effects on only the buses near the target lines, so the average perturbation was lower than the 14-bus case.

3.7 Proofs

3.7.1 Proof of Theorem 3.3.2

The *if* statement can be proved by constructing an undetectable attack following the arguments used to prove Theorem 3.3.1 and Theorem 3.3.5. Due to the space limit, we only provide the proof of the *only if* statement.

Let $\mathbf{a}$ be any attack with $\mathrm{Col}(H) \not\subseteq \mathrm{Col}(\bar{H}, \mathcal{U})$ where $\mathcal{U} \triangleq \{\mathbf{u}_1, \ldots, \mathbf{u}_K\}$ denotes the basis of $\mathcal{A}$ consisting of unit vectors in $\mathbb{R}^m$ and $U \in \mathbb{R}^{m \times K}$ is the matrix having the vectors in $\mathcal{U}$ as its columns. Without loss of generality, we assume that the columns of $\bar{H}$ and the unit vectors in $\mathcal{U}$ are linearly independent; if not, we can just work with a smaller set of $\mathcal{U}$ satisfying the independence condition.

Because $\mathrm{Col}(H) \not\subseteq \mathrm{Col}(\bar{H}, \mathcal{U})$, $\mathrm{Col}(H) \cap \mathrm{Col}(\bar{H}, \mathcal{U})$ is a subspace of $\mathrm{Col}(H)$ with a strictly smaller dimension. Hence, $\mathcal{S} \triangleq \{\mathbf{x} \in \mathbb{R}^n : H\mathbf{x} \in \mathrm{Col}(H) \cap \mathrm{Col}(\bar{H}, \mathcal{U})\}$ has the dimension less than n and thus a zero Lebesgue measure in $\mathbb{R}^n$. Let $\mathbf{x}$ be an arbitrary element of $\mathbb{R}^n \setminus \mathcal{S}$. Then, $\mathbf{y} \triangleq H\mathbf{x} \notin \mathrm{Col}(\bar{H}, \mathcal{U})$. When $\mathbf{x}$ is the true state, $\mathbf{z} = \mathbf{y} + \mathbf{e}$, and the $J(\hat{\mathbf{x}})$-test statistic for $\mathbf{a}$ is

$$J = \|W(\mathbf{y} + \mathbf{e} + \mathbf{a}(\mathbf{y} + \mathbf{e}))\|_{\Sigma^{-1}}$$

where $W = I - \bar{H}(\bar{H}^t \Sigma^{-1} \bar{H})^{-1} \bar{H}^t \Sigma^{-1}$. Since $\mathbf{a}(\mathbf{z}) \in \mathrm{Col}(\mathcal{U})$ for all $\mathbf{z}$, J is lower

bounded by

$$L \triangleq \min_{(a_k)_{k=1}^K} \|W(\mathbf{y} + \mathbf{e} + \sum_{k=1}^K a_k \mathbf{u}_k)\|_{\Sigma^{-1}}.$$

The minimization in L is achieved by the linear WLS solution, and one can show that $L = (\hat{W}(\mathbf{y} + \mathbf{e}))^t \Sigma^{-1} \hat{W}(\mathbf{y} + \mathbf{e})$ where $\hat{W} \triangleq W - (WU)[(WU)^t \Sigma^{-1}(WU)]^{-1}(WU)^t \Sigma^{-1} W$. W and $\hat{W}$ are idempotent and $\Sigma^{-1}W$ is symmetric. Using these properties, one may derive that

$$L = (\Sigma^{-\frac{1}{2}}(\mathbf{y} + \mathbf{e}))^t \Sigma^{\frac{1}{2}} \hat{W}^t \Sigma^{-\frac{1}{2}} (\Sigma^{-\frac{1}{2}}(\mathbf{y} + \mathbf{e})).$$

The above quadratic form has the following properties: (i) $\Sigma^{\frac{1}{2}} \hat{W}^t \Sigma^{-\frac{1}{2}}$ is idempotent and symmetric, (ii) $\Sigma^{-\frac{1}{2}}(\mathbf{y} + \mathbf{e}) \sim \mathcal{N}(\Sigma^{-\frac{1}{2}}\mathbf{y}, I_m)$, and (iii) $\text{rank}(\Sigma^{\frac{1}{2}} \hat{W}^t \Sigma^{-\frac{1}{2}}) = m - n - K$. With these three properties, Theorem B.33 and Theorem 1.3.3 in [75] imply that L has the noncentral chi-squared distribution with the $(m - n - K)$ degree of freedom and the noncentral parameter $\lambda \triangleq (\hat{W}\mathbf{y})^t \Sigma^{-1}(\hat{W}\mathbf{y})$.

It can be shown that $\mathbf{y} \notin \text{Col}(\bar{H}, \mathcal{U})$ implies $\hat{W}\mathbf{y} \neq \mathbf{0}$. Hence, if the diagonal entries of Σ (denoted by σ_{ii}^2, $1 \leq i \leq m$) uniformly decrease to 0, then $\lambda = \sum_{i=1}^m \sigma_{ii}^{-2}(\hat{W}\mathbf{y})_i^2$ grows to infinity. Suppose that the $J(\hat{\mathbf{x}})$-test uses a threshold τ. The detection probability of the attack is $\Pr(J > \tau)$, and it is lower bounded by $\Pr(L > \tau)$. And, $\Pr(L > \tau)$ approaches 1 as the noncentral parameter λ grows to infinity. Therefore, if the diagonal entries of Σ (*i.e.*, noise variances) uniformly decreases to 0, then λ grows to infinity and $\Pr(J > \tau)$ approaches 1. Hence, the only if statement and the additional statement are proved. $\blacksquare$

3.7.2 Proof of Theorem 3.3.3

Let $\bar{\mathcal{E}} \triangle \mathcal{E} = \{(a, b)\}$. We prove the statement for the case that the attack removes (a, b), and there are two line flow meters on (a, b) (one for each direction) and

injection meters at both a and b. For the line addition attack and other meter availabilities, the similar argument can be made.

Suppose there exists an undetectable attack with $\mathcal{A}$, and let $\mathcal{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_K\}$ denote the basis of $\mathcal{A}$ consisting of unit vectors in $\mathbb{R}^m$. Theorem 3.3.1 implies $\mathrm{Col}(H) \subset \mathrm{Col}(\bar{H}, \mathcal{A})$. It can be easily verified that $\mathbf{m}_{(a,b)} \in \mathrm{Col}(\bar{H}, \mathcal{A})$, and this implies $\mathbf{m}_{(a,b)} = \bar{H}\mathbf{x} + \sum_{k=1}^{K} \alpha_k \mathbf{u}_k$ for some $\mathbf{x} \in \mathbb{R}^n$ and $(\alpha_k)_{k=1}^{K} \in \mathbb{R}^K$. Then, $\bar{\mathbf{m}} \triangleq \mathbf{m}_{(a,b)} - \sum_{k=1}^{K} \alpha_k \mathbf{u}_k \in \mathrm{Col}(\bar{H})$.

Let $\bar{m}^{ij}$ ($\bar{m}^i$) denote the row entry of $\bar{\mathbf{m}}$ corresponding to the line flow from i to j (the injection at i) and $\mathbf{u}_{(i,j)}$ ($\mathbf{u}_{(i)}$) denote the m-dimensional unit vector with 1 at the row corresponding to the line flow from i to j (the injection at i). Physically, $\bar{\mathbf{m}} \in \mathrm{Col}(\bar{H})$ means that $\bar{\mathbf{m}}$ is a vector of meter data consistent with the topology $\bar{\mathcal{G}}$. It implies that (i) $\bar{m}^{ab}$ and $\bar{m}^{ba}$ are zeros, since (a, b) is disconnected in $\bar{\mathcal{G}}$, and (ii) the Kirchhoff's current laws (KCL) should hold at bus a and b in $\bar{\mathcal{G}}$, $i.e.$, the sum of all outgoing line flows from a should be equal to the injection amount at a. Using the special structure of $\mathbf{m}_{(a,b)}$ and $\bar{\mathbf{m}}$, the following can be proved. From (i), one can prove that $\mathbf{u}_{(a,b)}, \mathbf{u}_{(b,a)} \in \mathcal{U}$. From (ii), one can show that $\mathcal{U}$ should include $\mathbf{u}_{(a)}$ or some $\mathbf{u}_{(a,k)}$ (or $\mathbf{u}_{(k,a)}$) with a and k connected in $\mathcal{G}$. Similarly, $\mathcal{U}$ should include $\mathbf{u}_{(b)}$ or some $\mathbf{u}_{(b,l)}$ (or $\mathbf{u}_{(l,b)}$) with b and l connected in $\mathcal{G}$. Hence, $|\mathcal{U}|$ is no less than the total number of meters located on the target line (a, b) and the target buses a and b. ∎

3.7.3   Proof of Theorem 3.3.4

Suppose $\mathbf{a}$ is an undetectable attack with $\mathcal{A}$ for the target topology $\bar{\mathcal{G}}$ satisfying the theorem conditions. Let $\mathcal{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_K\}$ be the basis of $\mathcal{A}$ consisting of unit

vectors in $\mathbb{R}^m$, and $\mathcal{J} \subset \mathcal{V}$ denote the set of target buses with injection meters. For ease of presentation, we assume that each target line (i, j) has two line flow meters, one for each direction. For other meter availabilities, the similar argument can be made.

Theorem 3.3.1 implies that $\text{Col}(H) \subset \text{Col}(\bar{H}, \mathcal{U})$. It can be easily shown that if the target lines do not form a closed path in $\mathcal{G}$, then $\text{Col}(H) \subset \text{Col}(\bar{H}, \mathcal{U})$ implies that $\mathbf{m}_{(i,j)} \in \text{Col}(\bar{H}, \mathcal{U})$ for all target lines $(i, j) \in \mathcal{E} \setminus \bar{\mathcal{E}}$.

$\mathbf{m}_{(i,j)} \in \text{Col}(\bar{H}, \mathcal{U})$ means that it is possible to find a linear combination of vectors in $\mathcal{U}$, $\sum_{k=1}^{K} \alpha_k \mathbf{u}_k$, such that $\bar{\mathbf{m}}_{(i,j)} \triangleq \mathbf{m}_{(i,j)} + \sum_{k=1}^{K} \alpha_k \mathbf{u}_k \in \text{Col}(\bar{H})$. $\bar{\mathbf{m}}_{(i,j)} \in \text{Col}(\bar{H})$ implies that (i) the row entries of $\bar{\mathbf{m}}_{(i,j)}$ corresponding to the line flows of the disconnected lines in $\bar{\mathcal{G}}$ are zeros, and (ii) the entries of $\bar{\mathbf{m}}_{(i,j)}$ satisfy KCLs at all buses in $\bar{\mathcal{G}}$.

For each $(i, j) \in \mathcal{E} \setminus \bar{\mathcal{E}}$, since (i, j) is disconnected in $\bar{\mathcal{G}}$, $\bar{m}_{(i,j)}^{ij} = \bar{m}_{(i,j)}^{ji} = 0$. On the other hand, $m_{(i,j)}^{ij} = 1$ and $m_{(i,j)}^{ji} = -1$. Hence, $\mathcal{U}$ should include $\mathbf{u}_{(i,j)}$ and $\mathbf{u}_{(j,i)}$. Therefore, $\mathcal{U}$ should contain $\{\mathbf{u}_{(i,j)}, \mathbf{u}_{(j,i)} : (i, j) \in \mathcal{E} \setminus \bar{\mathcal{E}}\}$.

For each $i \in \mathcal{J}$, the assumptions imply that each line adjacent to i in $\bar{\mathcal{G}}$ has at least one line flow meter. We let $\mathbf{n}_i$ denote the set of the line flow meters on the lines incident to i in $\bar{\mathcal{G}}$, and $\mathbf{m}_{(i,j)}^{\mathbf{n}_i}$ denote the vector of the corresponding entries in $\mathbf{m}_{(i,j)}$. Because $\mathbf{m}_{(i,j)}$ has nonzero entries only for the injections at i and j and the line flows through (i, j), $\mathbf{m}_{(i,j)}^{\mathbf{n}_i}$ has all zero entries. On the other hand, $m_{(i,j)}^i = 1$. Hence, for $\bar{\mathbf{m}}_{(i,j)}$ to satisfy the KCL at bus i in $\bar{\mathcal{G}}$, at least one of $m_{(i,j)}^i$ or entries of $\mathbf{m}_{(i,j)}^{\mathbf{n}_i}$ has to be modified by $\sum_{k=1}^{K} \alpha_k \mathbf{u}_k$. Thus, $\mathcal{U}$ should contain $\mathbf{u}_{(i)}$ or $\mathbf{u}_{(a,b)}$ for some $(a, b) \in \mathbf{n}_i$.

In case that $\mathbf{u}_{(i)} \notin \mathcal{U}$, for $\bar{\mathbf{m}}_{(i,j)}$ to satisfy the KCL at bus i in $\bar{\mathcal{G}}$, at least one

entry of $\bar{\mathbf{m}}_{(i,j)}^{\mathbf{n}_i}$ should have a nonzero value: suppose $\bar{m}_{(i,j)}^{ik}$ takes a nonzero value. If $k \in \mathcal{J}$, we can make a similar argument based on the KCL at k: $\mathcal{U}$ should contain $\mathbf{u}_{(k)}$ or $\mathbf{u}_{(a,b)}$ for some $(a,b) \in \mathbf{n}(k) \setminus \{(i,k), (k,i)\}$. Following this line of argument, we can derive that for each $i \in \mathcal{J}$, $\mathcal{U}$ should contain unit vectors corresponding to at least one of the following sets: (i) injection meter at i, (ii) line flow meters on all the lines in some path $(i, v_2, \ldots, v_n)$ in $\bar{\mathcal{G}}^*$ and injection meter at v_n where $v_2, \ldots, v_n \in \mathcal{J}$, or (iii) line flow meters on all the lines in some path $(i, v_2, \ldots, v_n)$ in $\bar{\mathcal{G}}^*$ where $v_2, \ldots, v_{n-1} \in \mathcal{J}$ and v_n is either equal to one of $\{v_2, \ldots, v_{n-1}\}$ or not in $\mathcal{J}$. For each $i \in \mathcal{J}$, $\mathcal{U}$ should contain at least one set of unit vectors corresponding to any of the above three cases: we let $\mathcal{S}_i$ to denote an arbitrary one of such sets.

Note that $\{\mathbf{u}_{(i,j)}, \mathbf{u}_{(j,i)} : (i,j) \in \mathcal{E} \setminus \bar{\mathcal{E}}\}$ does not overlap with $\cup_{i \in \mathcal{J}} \mathcal{S}_i$. Hence, $|\mathcal{U}| \geq |\cup_{i \in \mathcal{J}} \mathcal{S}_i| + |\{\mathbf{u}_{(i,j)}, \mathbf{u}_{(j,i)} : (i,j) \in \mathcal{E} \setminus \bar{\mathcal{E}}\}|$. Proving $|\cup_{i \in \mathcal{J}} \mathcal{S}_i| \geq |\mathcal{J}|$ gives us the theorem statement, because $|\mathcal{J}| + |\{\mathbf{u}_{(i,j)}, \mathbf{u}_{(j,i)} : (i,j) \in \mathcal{E} \setminus \bar{\mathcal{E}}\}|$ is the exact number of meters the state-preserving attack modifies.

We will prove the following statement for all $n \leq |\mathcal{J}|$, by mathematical induction: for any subset $\bar{\mathcal{J}} \subset \mathcal{J}$ with $|\bar{\mathcal{J}}| = n$, $|\cup_{i \in \bar{\mathcal{J}}} \mathcal{S}_i| \geq n$. For $n = 1, 2, 3$, the statement can be easily verified. Suppose the statement is true for all $n \leq k$ $(k \geq 3)$, and $\bar{\mathcal{J}}$ is an arbitrary subset of $\mathcal{J}$ with $|\bar{\mathcal{J}}| = k + 1$. The tree condition guarantees that $\bar{\mathcal{J}}$ can be partitioned into two nonempty sets $\bar{\mathcal{J}}_1$ and $\bar{\mathcal{J}}_2$ such that for any $b_1 \in \bar{\mathcal{J}}_1$ and $b_2 \in \bar{\mathcal{J}}_2$, every path in $\bar{\mathcal{G}}^*$ between b_1 and b_2 contains a node not in $\mathcal{J}$. This implies that $\cup_{b \in \bar{\mathcal{J}}_1} \mathcal{S}_b$ and $\cup_{b \in \bar{\mathcal{J}}_2} \mathcal{S}_b$ are disjoint. By the induction hypothesis, we have $|\cup_{b \in \bar{\mathcal{J}}_1} \mathcal{S}_b| \geq |\bar{\mathcal{J}}_1|$ and $|\cup_{b \in \bar{\mathcal{J}}_2} \mathcal{S}_b| \geq |\bar{\mathcal{J}}_2|$. Thus, $|\cup_{b \in \bar{\mathcal{J}}} \mathcal{S}_b| = |\cup_{b \in \bar{\mathcal{J}}_1} \mathcal{S}_b| + |\cup_{b \in \bar{\mathcal{J}}_2} \mathcal{S}_b| \geq |\bar{\mathcal{J}}_1| + |\bar{\mathcal{J}}_2| = |\bar{\mathcal{J}}|$. Therefore, the induction implies $|\cup_{i \in \mathcal{J}} \mathcal{S}_i| \geq |\mathcal{J}|$, and the theorem statement follows. ∎

3.7.4 Proof of Theorem 3.5.1

Suppose meters are protected as described with $\mathcal{T}$ and $\mathcal{B}$. Let $\mathcal{A}$ be the resulting subspace of feasible attack vectors and $\mathcal{U} \triangleq \{\mathbf{u}_1, \dots, \mathbf{u}_K\}$ denote the basis of $\mathcal{A}$ consisting of unit vectors in $\mathbb{R}^m$. Assume that an undetectable attack can be launched for some target topology $\bar{\mathcal{G}}$ (different from $\mathcal{G}$). We will show that this assumption leads to a contradiction.

Note that $\mathcal{U}$ cannot contain the unit vectors corresponding to the protected measurements. In addition, Theorem 3.3.2 implies that $\mathrm{Col}(H) \subset \mathrm{Col}(\bar{H}, \mathcal{U})$. These two imply that the lines in $\mathcal{E}_{\mathcal{T}}$ cannot be removed by the attack, because each line has a protected line flow meter.

Let $\widehat{H}$ ($\widehat{\bar{H}}$) denote the submatrix of H ($\bar{H}$) obtained by selecting the rows corresponding to the protected meter measurements. One can easily verify that $\mathrm{Col}(H) \subset \mathrm{Col}(\bar{H}, \mathcal{U})$ if and only if $\mathrm{Col}(\widehat{H}) \subset \mathrm{Col}(\widehat{\bar{H}})$. Hence, we have $\mathrm{Col}(\widehat{H}) \subset \mathrm{Col}(\widehat{\bar{H}})$. This means that for all $\mathbf{x} \in \mathbb{R}^n$, there exists $\mathbf{y} \in \mathbb{R}^n$ such that $\widehat{\bar{H}}\mathbf{y} = \widehat{H}\mathbf{x}$. Let $H_{\mathcal{T}}$ denote the submatrix of $\widehat{H}$ obtained by selecting the rows corresponding to the protected line flow meters on the spanning tree $\mathcal{T}$. Since the lines in $\mathcal{E}_{\mathcal{T}}$ cannot be removed by the attack, the $H_{\mathcal{T}}$ part of H remains the same in $\bar{H}$; hence, $H_{\mathcal{T}}$ is also a submatrix of $\widehat{\bar{H}}$. Thus, $\widehat{\bar{H}}\mathbf{y} = \widehat{H}\mathbf{x}$ implies $H_{\mathcal{T}}\mathbf{y} = H_{\mathcal{T}}\mathbf{x}$. Since $\mathcal{T}$ is a spanning tree and it has one protected line flow meter per line, the protected line meters on $\mathcal{T}$ makes the grid observable [26]. Hence, $H_{\mathcal{T}}$ has full column rank. Consequently, $H_{\mathcal{T}}\mathbf{y} = H_{\mathcal{T}}\mathbf{x}$ implies $\mathbf{y} = \mathbf{x}$, and we have $\widehat{\bar{H}}\mathbf{x} = \widehat{H}\mathbf{x}$. This holds for all $\mathbf{x} \in \mathbb{R}^n$.

Let a be any element in $\mathcal{B}$. We will show that any line in $\mathcal{L}_a$ cannot be a target line. Note that the injection meter at a is protected, so $\widehat{H}$ and $\widehat{\bar{H}}$ have

the row corresponding to the injection at a. $\widehat{\bar{H}}\mathbf{x} = \widehat{H}\mathbf{x}$ for all $\mathbf{x} \in \mathbb{R}^n$ implies that the injection at bus a should be the same for $\mathcal{G}$ and $\bar{\mathcal{G}}$ as long as the state is the same for the two cases. When the state is $\mathbf{x}$, the injection at a in $\mathcal{G}$ is $\sum_{k:\{a,k\}\in\tilde{\mathcal{E}}} B_{ak}(x_a - x_k)$, and the injection at a in $\bar{\mathcal{G}}$ is $\sum_{l:\{a,l\}\in\tilde{\bar{\mathcal{E}}}} B_{al}(x_a - x_l)$. Thus we have,

$$\sum_{k:\{a,k\}\in\tilde{\mathcal{E}}} B_{ak}(x_a - x_k) = \sum_{l:\{a,l\}\in\tilde{\bar{\mathcal{E}}}} B_{al}(x_a - x_l), \ \forall \mathbf{x} \in \mathbb{R}^n,$$

which can be rewritten as follows: for all $\mathbf{x} \in \mathbb{R}^n$,

$$\sum_{k:\{a,k\}\in\tilde{\mathcal{E}}\setminus\tilde{\bar{\mathcal{E}}}} B_{ak}(x_a - x_k) - \sum_{l:\{a,l\}\in\tilde{\bar{\mathcal{E}}}\setminus\tilde{\mathcal{E}}} B_{al}(x_a - x_l) = 0.$$

If $\mathcal{L}_a \cap (\tilde{\bar{\mathcal{E}}}\triangle\tilde{\mathcal{E}})$ is not empty, the above statement is true only when $B_{ak} = 0$ for all $\{a,k\} \in \mathcal{L}_a \cap (\tilde{\bar{\mathcal{E}}}\triangle\tilde{\mathcal{E}})$. B_{ak} is the susceptance of the line $\{a,k\}$ when it is "connected", and this value is nonzero in practice for every line. Hence, $\mathcal{L}_a\cap(\tilde{\bar{\mathcal{E}}}\triangle\tilde{\mathcal{E}})$ should be empty; $i.e.$, a line in $\mathcal{L}_a$ cannot be a target line.

It was shown that the lines in $\mathcal{T}$ and $\cup_{a\in\mathcal{B}}\mathcal{L}_a$ cannot be a target line. Thus, the condition (3.15) implies that no line can be a target line, and this contradicts the assumption that there exists an undetectable topology attack. ∎

CHAPTER 4

DATA FRAMING ATTACK ON STATE ESTIMATION

4.1 Introduction

A promising feature of a future smart grid is the data-driven approach to automated monitoring, control, and decision. The paradigm shift to a data-driven framework enables deeper integration of data collection and sophisticated data processing. While extracting actionable information from real-time sensing data can make the grid more efficient and adaptive to real-time operating conditions, it exposes the grid to possible cyber data attacks aimed at disrupting grid operations and potentially causes blackouts.

In [18], Liu, Ning, and Reiter presented perhaps the first man-in-the-middle (MiM) attack on the power system state estimation where an adversary replaces "normal" sensor data with "malicious data." It was shown that, if the adversary could gain control of a sufficient number of meters, it could perturb the state estimate by an arbitrary amount without being detected by the bad data detector employed at the control center. Such undetectable attacks are referred to as *covert data attacks*.

The condition under which covert data attacks are possible is found to be equivalent to that of system observability. In particular, covert attacks are possible if and only if the system becomes unobservable when the meters under attack are removed [24] (or equivalently, the adversary controls a critical set of meters.) Therefore, the minimum number of meters that an adversary has to control in order to launch a covert data attack, referred to as a *security index*, is an important

measure of security against data attack. It represents a fundamental limit on the capability of an adversary to disrupt covertly the operation of the grid [20, 24].

In this chapter, we show that the barrier on the capability of an adversary can be circumvented by using a different form of attacks, one that exploits the vulnerabilities of the existing bad data detection and removal mechanisms. In particular, we show that the adversary only needs to gain control of about half of the meters required by the security index while achieving the same objective of perturbing the state estimate by an arbitrary amount without being detected by the control center.

The attacks considered in this chapter are referred to as *data framing attacks*, borrowing the notion of framing as providing false evidence to make someone innocent appear to be guilty of misconducts. In the context of state estimation, a data framing attack means that an adversary launches a data attack in such a way that the control center detects the presence of bad data and identifies normal meters as sources of bad data. To this end, the attacker does not try to make malicious data pass the bad data detection (as a covert attack tries to do). Instead, it purposely triggers the bad data detection and causes erroneous removal of good data. Unknown to the control center, the remaining data still contain adversary-injected malicious data, causing errors in the state estimate.

4.1.1 Summary of Results and Organization

We propose a data framing attack on power system state estimation. Specifically, we formulate the design of optimal data framing attack as a quadratically constrained quadratic program (QCQP). To analyze the efficacy of the data fram-

ing attack, we present a sufficient condition under which the framing attack can achieve an arbitrary perturbation of the state estimate by controlling only half of the critical set of meters. We demonstrate with the IEEE 14-bus and 118-bus networks that the sufficient condition holds in critical sets associated with cuts.

The optimal design of framing attack is based on a linearized system. In practice, a nonlinear state estimator is often used. We demonstrate that, under the nonlinear measurement model, the framing attacks designed based on linearized system model successfully perturb the state estimate, and the adversary can control the degree of perturbation as desired.

The rest of the chapter is organized as follows. Section 4.2 introduces the measurement and adversary models with preliminaries on state attacks. Section 4.3 presents the mathematical model of state estimation and bad data processing. In Section 4.4, we present the main idea of the data framing attack and the QCQP framework for the attack design. Section 4.5 provides a theoretical justification of the efficacy of the data framing attack. In Section 4.6, we test the data framing attack with the IEEE 14-bus and 118-bus networks.

4.2 Mathematical Models

This section introduces the topology and system state of a power network, the meter measurement model, and the adversary model. In addition, the covert state attack and its connection with network observability are explained. Throughout the chapter, boldface lower case letters ($e.g.$, $\mathbf{x}$) denote vectors, x_i denotes the ith entry of the vector $\mathbf{x}$, boldface upper case letters ($e.g.$, $\mathbf{H}$) denote matrices, $\mathbf{H}_{ij}$ denotes the (i, j) entry of $\mathbf{H}$, $\mathcal{R}(\mathbf{H})$ denotes the column space of $\mathbf{H}$, $\mathcal{N}(\mathbf{H})$ denotes

the null space of $\mathbf{H}$, and script letters ($e.g., \mathfrak{I}, \mathcal{A}$) denote sets. The multivariate normal distribution with the mean $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ is denoted by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

4.2.1   Network and Measurement Models

A power network is a network of buses connected by transmission lines, and thus the *topology* of the grid can be naturally defined as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is the set of buses, and $\mathcal{E}$ is the set of lines connecting buses ($\{i, j\} \in \mathcal{E}$ if and only if bus i and bus j are connected.) The *system state* of the power network is defined as the vector of bus voltage magnitudes and phase angles, from which all the other quantities (*e.g.*, power line flows, power injections, line currents) can be calculated.

For real-time estimation of the system state, the control center collects measurements from line flow and bus injection meters[1] deployed throughout the grid. The meter measurements are related to the system state $\mathbf{x}$ in a nonlinear fashion, and the relation is described by the AC model [51]:

$$\mathbf{z} = h(\mathbf{x}) + \mathbf{e}, \tag{4.1}$$

where $h(\cdot)$ is the nonlinear measurement function, and $\mathbf{e}$ is the Gaussian measurement noise with a diagonal covariance matrix.

If some of the meters malfunction or an adversary injects malicious data, the control center observes biased measurements,

$$\bar{\mathbf{z}} = h(\mathbf{x}) + \mathbf{e} + \mathbf{a}, \tag{4.2}$$

---

[1]Other types of meters can also be considered, but we restrict our attention to line flow and bus injection meters for simplicity.

where **a** represents a deterministic bias. In such a case, the data are said to be *bad*, and the biased meter entries are referred to as *bad data entries*. Note that even when a meter is protected from adversarial modification, it may still have a bias due to a physical malfunction or improper parameter setting; filtering out the measurements from such malfunctioning meters was the original objective of the legacy bad data processing and is adopted in practice today [40].

Even though the model in (4.1) is nonlinear, the state estimate is generally obtained by iterations of weighted linear least squares estimation with the locally linearized model [51]. Therefore, it is reasonable to analyze the performance of state estimation using the locally linearized model around the system operating point. To this end, in analyzing the attack effect on state estimation, we adopt the so-called DC model [51]. In the DC model, for the ease of analysis, the AC model (4.1) is linearized around the system state where all voltage phasors are equal to $1\angle 0$, and only real part of the measurements are retained:

$$\mathbf{z} = \mathbf{Hx} + \mathbf{e}, \tag{4.3}$$

where $\mathbf{z} \in \mathbb{R}^m$ is the measurement vector consisting of real part of line flow and bus injection measurements, the system state $\mathbf{x} \in \mathbb{R}^n$ is the vector of voltage phase angles at all buses except the reference bus ($\mathbf{x}$ is unknown, but deterministic), $\mathbf{H} \in \mathbb{R}^{m \times n}$ is the DC measurement matrix that relates the system state to bus injection and line flow amounts, and $\mathbf{e}$ is the Gaussian measurement noise with a diagonal covariance matrix $\mathbf{\Sigma}$. We represent the noise covariance matrix $\mathbf{\Sigma}$ as $\mathbf{\Sigma} = \sigma^2 \bar{\mathbf{\Sigma}}$, where $\bar{\mathbf{\Sigma}}$ is a diagonal matrix representing the variation of noise variances across different meters ($\sum_{i=1}^{m} \bar{\mathbf{\Sigma}}_{ii} = 1$), and σ^2 is a scaling factor.

Each row of $\mathbf{H}$ has a special structure depending on the type of the meter [51]. For ease of presentation, consider the noiseless measurement $\mathbf{z} = \mathbf{Hx}$. If an entry

z_k of $\mathbf{z}$ is the measurement of the line flow from bus i to bus j, z_k is $B_{ij}(x_i - x_j)$ where B_{ij} is the line susceptance and x_i is the voltage phase angle at bus i [51]. If z_k is the measurement of bus injection at i, it is the sum of all the outgoing line flows from i, and the corresponding row of $\mathbf{H}$ is the sum of the row vectors corresponding to all the outgoing line flows.

The analysis based on the DC model needs to be verified using the realistic AC model simulations; we demonstrate in Section 4.6 that the proposed attack strategy is also effective in the AC model simulations.

4.2.2   Adversary Model

We consider a man-in-the-middle attack on power system state estimation. As described in Fig. 4.1, an adversary is assumed to be capable of modifying the data from a subset of analog meters $\mathcal{I}_A$. We refer to the meters in $\mathcal{I}_A$ as *adversary meters*.

The control center observes the corrupted measurements $\bar{\mathbf{z}}$ instead of the original measurements $\mathbf{z}$ in (4.1). We assume that the adversary knows the line parameters (*i.e.*, the measurement function h and the measurement matrix $\mathbf{H}$).

The adversarial modification is mathematically modeled as follows:

$$\bar{\mathbf{z}} = \mathbf{z} + \mathbf{a}, \quad \mathbf{a} \in \mathcal{A}, \tag{4.4}$$

where $\mathbf{a}$ is an attack vector, and $\mathcal{A}$ is the set of feasible attack vectors defined as $\mathcal{A} \triangleq \{\mathbf{c} \in \mathbb{R}^m : c_i = 0, \ \forall i \notin \mathcal{I}_{\mathbf{a}}\}$. Note that $\mathcal{A}$ fully characterizes the ability of the adversary. In addition, the adversary is assumed to design $\mathbf{a}$ without observing any entry of $\mathbf{z}$, *i.e.*, the attack does not require any real-time observation.

4.2.3 Network Observability and Covert State Attack

For state estimation to be feasible, the control center needs to have enough meter measurements, from which the system state can be uniquely determined. Formally, a power network is said to be *locally observable* at a state $\mathbf{x}_0$ if the system state can be uniquely determined from the noiseless meter measurements $h(\mathbf{x})$ in a neighborhood of $\mathbf{x}_0$. This implies that the Jacobian of h at $\mathbf{x}_0$ has full rank. However, due to the intractability of checking local observability for all feasible operating points, the DC model (4.3) is generally adopted for observability analysis [26]: the network is said to be *observable* if the DC measurement matrix $\mathbf{H}$ has full rank. In practice, power networks should be designed to satisfy observability. Hence, we assume that the network of our interest is observable (*i.e.*, $\mathbf{H}$ has full rank.)

The concept of network observability is closely related to the feasibility of a covert state attack. The covert state attack was proposed in [18] under the DC model: if there exists $\mathbf{y} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ such that $\mathbf{Hy} \in \mathcal{A}$, then setting $\mathbf{a}$ equal to $\mathbf{Hy}$ results in

$$\bar{\mathbf{z}} = \mathbf{Hx} + \mathbf{e} + \mathbf{a} = \mathbf{H}(\mathbf{x} + \mathbf{y}) + \mathbf{e}, \tag{4.5}$$

and thus, $\bar{\mathbf{z}}$ cannot be distinguished from a normal noisy measurement vector with the state $\mathbf{x} + \mathbf{y}$. Furthermore, by properly scaling the attack vector (*e.g.*, $\alpha\mathbf{a}$), the adversary can perturb the state estimate by an *arbitrary* degree (*e.g.*, $\alpha\mathbf{y}$).

It is shown in [24] that a covert attack is feasible if and only if the adversary can control a *critical set* of meters, which is defined as a set of meters such that removing the set from the network renders the network unobservable while removing any proper subset of it does not [51]. Hence, the feasibility condition means that removing the adversary meters renders the measurement matrix rank deficient.
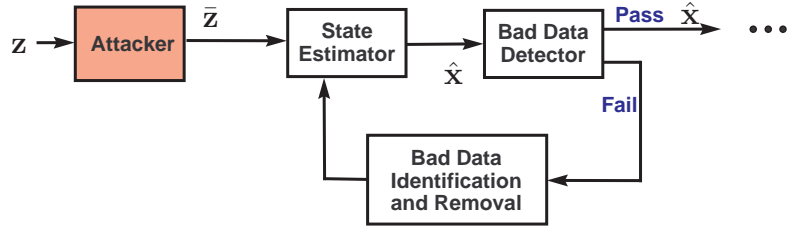
Figure 4.1: Adversary model with state estimation and bad data test

The intuition behind the condition is that, for any $\mathbf{y} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$, $\mathbf{Hy}$ is in $\mathcal{A}$ if and only if $\mathbf{Hy}$ has zero entries for all non-adversary meters; the latter implies that the measurement matrix after the removal of the rows corresponding to the adversary meters is rank deficient, because $\mathbf{y}$ is in its null space.

4.3 State Estimation and Bad Data Processing

This section introduces a popular approach of state estimation and bad data processing, which we assume to be employed by the control center. Once the control center receives the measurements $\mathbf{z}$, it aims to obtain the estimate $\hat{\mathbf{x}}$ of the system state $\mathbf{x}$. Because bad data entries in $\mathbf{z}$ may result in a bias in the state estimate, the control center iteratively conducts state estimation and bad data detection and identification to filter out possible bad data entries in $\mathbf{z}$.

Fig. 4.1 illustrates an iterative scheme for obtaining $\hat{\mathbf{x}}$, which consists of three function blocks: State Estimation, Bad Data Detection, and Bad Data Identification [40, 51]. The iteration begins with the initial measurement vector $\mathbf{z}^{(1)} \triangleq \mathbf{z}$ and the initial measurement function $h^{(1)} \triangleq h$ where the superscript denotes the index for the current iteration. In each iteration, (i) the state estimate is obtained (State Estimation), (ii) presence of bad data is tested (Bad Data Detection), and

101

Table 4.1: State estimation and bad data processing

Bad-Data-Processing($\mathbf{z}, h, \boldsymbol{\Sigma}$)
1:    $\mathbf{z}^{(1)} \leftarrow \mathbf{z}$; $h^{(1)} \leftarrow h$; $k \leftarrow 1$;
2:    while ($true$)
3:      $(\hat{\mathbf{x}}^{(k)}, \mathbf{r}^{(k)}) \leftarrow$ State-Estimation($\mathbf{z}^{(k)}, h^{(k)}$);
4:      $result \leftarrow$ Bad-Data-Detection($\mathbf{r}^{(k)}$);
5:      if $result == good$
6:        break;
7:      else
8:        $(\mathbf{z}^{(k+1)}, h^{(k+1)}) \leftarrow$ Bad-Data-ID($\mathbf{r}^{(k)}, \mathbf{z}^{(k)}, h^{(k)}$);
9:      end
10:    $k \leftarrow k + 1$;
11:   end
12:   return $\hat{\mathbf{x}}^{(k)}$;

(iii) if data are declared to be bad, one data entry is identified as bad and removed from the measurement vector (Bad Data Identification). Table 4.1 provides the pseudocode for the overall procedure. In the following subsections, the detailed operation of each function block will be presented.

4.3.1   State Estimation and Bad Data Detection

In the kth iteration, State Estimation uses $(\mathbf{z}^{(k)}, h^{(k)})$ as an input, and obtains the weighted least squares (WLS) estimate of the system state:

$$\hat{\mathbf{x}}^{(k)} \triangleq \arg\min_{\mathbf{x}} (\mathbf{z}^{(k)} - h^{(k)}(\mathbf{x}))^T (\boldsymbol{\Sigma}^{(k)})^{-1} (\mathbf{z}^{(k)} - h^{(k)}(\mathbf{x})), \qquad (4.6)$$

where $\boldsymbol{\Sigma}^{(k)}$ is the covariance matrix of the corresponding noise vector. Based on the state estimate, the residue vector is also evaluated:

$$\mathbf{r}^{(k)} \triangleq \mathbf{z}^{(k)} - h^{(k)}(\hat{\mathbf{x}}^{(k)}). \qquad (4.7)$$

We assume that the $J(\hat{\mathbf{x}})$-test [40, 51] is employed for bad data detection: Bad Data Detection makes a decision based on the sum of weighted squared residues:

$$
\begin{cases}
\text{bad data} & \text{if } (\mathbf{r}^{(k)})^T (\mathbf{\Sigma}^{(k)})^{-1} \mathbf{r}^{(k)} > \tau^{(k)}; \\
\text{good data} & \text{if } (\mathbf{r}^{(k)})^T (\mathbf{\Sigma}^{(k)})^{-1} \mathbf{r}^{(k)} \leq \tau^{(k)}.
\end{cases}
\tag{4.8}
$$

The $J(\hat{\mathbf{x}})$-test is widely used due to its low complexity and the fact that the test statistic has a χ^2 distribution if the data are good [40]. The latter fact is used to set the threshold $\tau^{(k)}$ for a given false alarm constraint.

4.3.2 Iterative Bad Data Identification and Removal

If Bad Data Detection (4.8) declares that the data are good, the algorithm returns the state estimate $\hat{\mathbf{x}}^{(k)}$ and terminates. However, if Bad Data Detection declares that the data are bad, Bad Data Identification is invoked to identify and remove one bad data entry from the measurement vector.

A widely used criterion for identifying a bad data entry is the normalized residue [40, 51], which is considered one of the most reliable criteria [41]. In the normalized residue analysis, each $r_i^{(k)}$ is divided by its standard deviation under the good data hypothesis (*i.e.*, the standard deviation of $r_i^{(k)}$ when there exists no bad data entry in $\mathbf{z}^{(k)}$.) If there exists no bad data entry in $\mathbf{z}^{(k)}$, and the state estimate $\hat{\mathbf{x}}^{(k)}$ is close to the actual state $\mathbf{x}$, the distribution of $\mathbf{r}^{(k)}$ can be approximated by $\mathcal{N}(\mathbf{0}, \mathbf{W}^{(k)} \mathbf{\Sigma}^{(k)})$ where

$$
\mathbf{W}^{(k)} \triangleq \mathbf{I} - \mathbf{H}^{(k)} ((\mathbf{H}^{(k)})^T (\mathbf{\Sigma}^{(k)})^{-1} (\mathbf{H}^{(k)}))^{-1} (\mathbf{H}^{(k)})^T (\mathbf{\Sigma}^{(k)})^{-1}
\tag{4.9}
$$

with $\mathbf{H}^{(k)}$ denoting the Jacobian of $h^{(k)}$ at $\hat{\mathbf{x}}^{(k)}$ and $\mathbf{I}$ denoting the identity matrix with the appropriate size (see Appendix of [40] for the detail.) Hence, the

normalized residue is calculated as

$$\tilde{\mathbf{r}}^{(k)} = \mathbf{\Omega}^{(k)} \mathbf{r}^{(k)}, \tag{4.10}$$

where $\mathbf{\Omega}^{(k)}$ is a diagonal matrix with

$$\mathbf{\Omega}_{ii}^{(k)} = \begin{cases} 0 & \text{if } \{i\} \text{ is a critical set}^2, \\ \dfrac{1}{\sqrt{(\mathbf{W}^{(k)}\mathbf{\Sigma}^{(k)})_{ii}}} & \text{otherwise.} \end{cases} \tag{4.11}$$

Once the normalized residue $\tilde{\mathbf{r}}^{(k)}$ is calculated, the meter with the largest $|\tilde{r}_i^{(k)}|$ is identified as a bad meter. Bad Data Identification removes the row of $\mathbf{z}^{(k)}$ and the row of $h^{(k)}$ that correspond to the bad meter and returns the updated measurement vector and measurement function for the next iteration, denoted by $\mathbf{z}^{(k+1)}$ and $h^{(k+1)}$.

Under the DC model (4.3), State Estimation, Bad Data Detection, and Bad Data Identification are the same with that in the AC model, except that the nonlinear measurement function $h^{(k)}(\mathbf{x})$ is replaced with the linear function $\mathbf{H}^{(k)}\mathbf{x}$ (so, the Jacobian is the same everywhere.) Note that the WLS state estimate (4.6) is replaced with a simple linear WLS solution:

$$\hat{\mathbf{x}}^{(k)} = ((\mathbf{H}^{(k)})^T (\mathbf{\Sigma}^{(k)})^{-1} (\mathbf{H}^{(k)}))^{-1} (\mathbf{H}^{(k)})^T (\mathbf{\Sigma}^{(k)})^{-1} \mathbf{z}^{(k)}, \tag{4.12}$$

and thus

$$\mathbf{r}^{(k)} = \mathbf{z}^{(k)} - \mathbf{H}^{(k)} \hat{\mathbf{x}}^{(k)} = \mathbf{W}^{(k)} \mathbf{z}^{(k)}. \tag{4.13}$$

---

[2]If $\{i\}$ is a critical set (*i.e.*, removing the meter i makes the grid unobservable), its residue is always equal to zero [51], and the corresponding diagonal entry of $\mathbf{W}^{(k)}\mathbf{\Sigma}^{(k)}$ is zero. For such a meter, the normalizing factor is 0 such that its normalized residue is equal to 0.

4.4 Data Framing Attack

This section presents a new attack strategy on state estimation, referred to as *data framing attack*, which exploits the bad data processing to remove data from some normally operating meters and make the adversary meters appear to be trustworthy. We present the main idea and the QCQP framework for the optimal design of the attack.

We focus our attention to the case where the adversary cannot control enough meters to launch a covert attack. The set of normal meters that the framing attack aims to remove (*i.e.*, frame as bad meters) is referred to as the *target set*, denoted by $\mathcal{I}_\mathrm{T}$. The target set $\mathcal{I}_\mathrm{T}$ is chosen such that after the target meters are removed from the grid, a covert attack becomes feasible. For instance, suppose that $\mathcal{I}$ is a critical set. The feasibility condition of the covert attack, explained in Section 4.2.3, implies that if $\mathcal{I} \setminus \mathcal{I}_\mathrm{A}$ is removed from the grid, then the adversary with $\mathcal{I}_\mathrm{A}$ can launch a covert state attack, because further removing all the meters in $\mathcal{I}_\mathrm{A}$ makes the grid unobservable. Therefore, $\mathcal{I} \setminus \mathcal{I}_\mathrm{A}$ can be set as the target set $\mathcal{I}_\mathrm{T}$.

The resulting state perturbation by the framing attack does depend on the choice of the target set. Finding the optimal target set for a given attack objective is certainly an important problem. However, it is out of scope of this chapter. We focus on the design of the attack vector for a fixed target set.

4.4.1 Effect of Attack on Normalized Residues

To analyze how the attack affects the bad data processing, we analyze, under the DC model (4.3), the adversarial effect on the normalized residue vector in the first iteration. In this subsection, we omit the superscript to simplify notation: all the quantities we consider are associated with the first iteration unless otherwise specified.

Suppose that $\mathbf{z}$ is the measurement vector without bad data under the DC model (4.3). The normalized residue in the first iteration is obtained as

$$\tilde{\mathbf{r}} = \mathbf{\Omega r} = \mathbf{\Omega W z}, \tag{4.14}$$

where $\mathbf{\Omega} = \mathbf{\Omega}^{(1)}$ is defined as in (4.11).

Due to the normalization, each entry $\tilde{r}_i$ is distributed as $\mathcal{N}(0, 1)$ unless $\{i\}$ is a critical set [51]; if $\{i\}$ is a critical set, the normalized residue for the meter i is always equal to zero for any $\mathbf{z}$.

If an attack vector $\mathbf{a}$ is added, the resulting normalized residue is

$$\tilde{\mathbf{r}} = \mathbf{\Omega W}(\mathbf{z} + \mathbf{a}) = \mathbf{\Omega W z} + \mathbf{\Omega W a}. \tag{4.15}$$

Thus, if $\{i\}$ is not a critical set, $\tilde{r}_i$ is distributed as $\mathcal{N}((\mathbf{\Omega W a})_i, 1)$; if $\{i\}$ is critical, $\tilde{r}_i = (\mathbf{\Omega W a})_i$ surely.

Recalling that the absolute normalized residues (*i.e.*, $|\tilde{r}_i|$) are the statistics used for identifying the bad data entries, one intuitive heuristic to get the target meters removed is to make the mean energy of the normalized residues at the target meters as large as possible. Making the target meters have large normalized residues in the first iteration is of course not a guarantee of their removal in the following

iterations. Nevertheless, this is a reasonable heuristic to avoid the difficult task of analyzing the dynamic adversarial effect in subsequent iterations. Note that

$$\mathbb{E}\left[\sum_{i \in \mathcal{I}_{\mathrm{T}}} (\tilde{r}_i)^2\right] = \sum_{i \in \mathcal{I}_{\mathrm{T}}} \mathbb{E}[(\tilde{r}_i)^2] = \sum_{i \in \mathcal{I}_{\mathrm{T}}} (\mathbf{\Omega W a})_i^2 + C, \tag{4.16}$$

where C is the number of the meters in $\mathcal{I}_{\mathrm{T}}$ that do not form a single-element critical set. Therefore, maximizing the mean energy of the normalized target residues is equivalent to maximizing $\sum_{i \in \mathcal{I}_{\mathrm{T}}} (\mathbf{\Omega W a})_i^2 = \|\mathbf{S}_{\mathrm{T}} \mathbf{\Omega W a}\|_2^2$ where $\mathbf{S}_{\mathrm{T}} \in \mathbb{R}^{|\mathcal{I}_{\mathrm{T}}| \times m}$ is the row-selection matrix which retains only the rows corresponding to the target meters.

4.4.2 Optimal Framing Attack via QCQP

The ultimate objective of the framing attack is to gain an ability to perturb the state estimate by an arbitrary degree. To this end, the framing attack aims to accomplish two tasks.

The first is to make the bad data processing remove the target meters such that the network with the remaining meters becomes vulnerable to a covert state attack by the adversary. As discussed in Section 4.4.1, we attempt to achieve this goal by maximizing the mean energy of the normalized target residues, which is equivalent to maximizing $\|\mathbf{S}_{\mathrm{T}} \mathbf{\Omega W a}\|_2^2$.

The second task is to ensure that the attack becomes covert after the target meters are removed, thereby making arbitrary state perturbation possible. Let $\mathbf{H}_0$ denote the $m \times n$ measurement matrix obtained from $\mathbf{H}$ by replacing the rows corresponding to the target meters with zero row vectors. Then, the attack becomes covert (*i.e.*, the attack vector lies in the column space of the measurement

107

matrix) after the target meters are removed, if and ony if $\mathbf{a}$ is in $\mathcal{R}(\mathbf{H}_0)$. Therefore, we restrict the attack vector $\mathbf{a}$ to be not only in the feasible set $\mathcal{A}$ but also in $\mathcal{R}(\mathbf{H}_0)$.

Based on the aforementioned intuition, we solve the following optimization to find the optimal *direction* to align the attack vector:

$$\begin{aligned} \max_{\mathbf{a}} \quad & \|\mathbf{S}_{\mathrm{T}}\boldsymbol{\Omega}\mathbf{W}\mathbf{a}\|_2^2 \\ \mathrm{subj.} \quad & \|\mathbf{a}\|_2^2 = 1, \quad \mathbf{a} \in \mathcal{R}(\mathbf{H}_0) \cap \mathcal{A}. \end{aligned} \tag{4.17}$$

The optimization (4.17) gives the optimal direction $\mathbf{a}^*$ of the attack vector to maximize the mean energy of the normalized target residues, among the feasible directions that render the attack covert after the target meters are removed.

To provide a more intuitive description of the feasible set in (4.17), we introduce the $(m - |\mathcal{I}_{\mathrm{A}}| - |\mathcal{I}_{\mathrm{T}}|) \times n$ matrix $\bar{\mathbf{H}}$ obtained from $\mathbf{H}$ by removing the rows corresponding to the adversary and target meters. It can be easily seen that $\mathbf{a} \in \mathcal{R}(\mathbf{H}_0) \cap \mathcal{A}$ if and only if $\mathbf{a} = \mathbf{H}_0 \mathbf{x}_0$ for some $\mathbf{x}_0 \in \mathcal{N}(\bar{\mathbf{H}})$. Therefore, the dimension of $\mathcal{R}(\mathbf{H}_0) \cap \mathcal{A}$ is equal to the dimension of $\mathcal{N}(\bar{\mathbf{H}})$. For instance, if $\mathcal{I}_{\mathrm{A}} \cup \mathcal{I}_{\mathrm{T}}$ is a critical set, $\bar{\mathbf{H}}$ has rank $n - 1$, and its null space has dimension one. Therefore, in this case, $\mathcal{R}(\mathbf{H}_0) \cap \mathcal{A}$ is a one-dimensional space, and there is no need to search for the optimal direction. On the other hand, if $\mathcal{I}_{\mathrm{A}} \cup \mathcal{I}_{\mathrm{T}}$ contains more than one critical sets, the dimension of $\mathcal{N}(\bar{\mathbf{H}})$ is greater than one, and the optimization (4.17) searches for the optimal direction among the infinite set of feasible directions.

Finally, we set an attack vector $\mathbf{a}$ as $\eta \mathbf{a}^*$ where $\eta \in \mathbb{R}$ is a parameter that adjusts the direction (*i.e.*, positive or negative depending on the sign of η) and the magnitude of the resulting state perturbation. It is important to point out that a sufficiently large $|\eta|$ is necessary for successful removal of the target meters. Because, the mean of the $J(\hat{\mathbf{x}})$-test statistic (of Bad Data Detection) increases linearly with respect to $|\eta|^2$ [40], and we want the test statistic to be larger than the

threshold in multiple iterations such that Bad Data Identification can be invoked enough times to remove all the target meters.

In practice, real-world power meters have very high signal-to-noise ratios (SNRs) [76], which means that even a small attack vector can be detected by the $J(\hat{\mathbf{x}})$-test. Therefore, the necessary size of $|\eta|$ to invoke Bad Data Identification in multiple iterations is expected to be reasonably small. The numerical examples in Section 4.6 demonstrate that the framing attack that perturbs the measurement vector by less than 1% in L_1-norm can succeed under a moderately high SNR setting.

The optimization (4.17) can be written as a QCQP:

$$
\begin{aligned}
\min_{\mathbf{q}} \quad & \mathbf{q}^T \mathbf{P} \mathbf{q} \\
\text{subj.} \quad & \mathbf{q}^T \mathbf{Q} \mathbf{q} - 1 = 0, \quad \mathbf{q} \in \mathbb{R}^p,
\end{aligned} \tag{4.18}
$$

where

$$
\mathbf{P} \triangleq -(\mathbf{S}_\mathrm{T} \mathbf{\Omega} \mathbf{W} \mathbf{B})^T (\mathbf{S}_\mathrm{T} \mathbf{\Omega} \mathbf{W} \mathbf{B}), \quad \mathbf{Q} \triangleq \mathbf{B}^T \mathbf{B}, \tag{4.19}
$$

and $\mathbf{B} \in \mathbb{R}^{m \times p}$ is the basis matrix of the p-dimensional vector space $\mathcal{R}(\mathbf{H}_0) \cap \mathcal{A}$. Note that the dimension p is nonzero because the target meters are set such that a covert attack becomes feasible after their removal. In addition, $\mathbf{P}$ is negative semidefinite, and $\mathbf{Q}$ is positive definite since $\mathbf{B}$ has full column rank. The positive definiteness of $\mathbf{Q}$ implies that a solution exists (*i.e.*, the objective function is bounded below.)

The KKT conditions for (4.18) are as follows:

$$
\mathbf{P} \mathbf{q} + \lambda (\mathbf{Q} \mathbf{q}) = 0, \quad \mathbf{q}^T \mathbf{Q} \mathbf{q} - 1 = 0, \tag{4.20}
$$

where λ is the Lagrange multiplier for the equality constraint. The optimal solution $\mathbf{q}^*$ of (4.18) is the one that results in the minimum objective function value among

all $(\lambda, \mathbf{q})$ pairs satisfying the KKT conditions (4.20).

The KKT conditions (4.20) imply that

$$
\begin{aligned}
\mathbf{Q}^{-1}\mathbf{P}\mathbf{q} &= \lambda\mathbf{q}; \\
\mathbf{q}^T\mathbf{P}\mathbf{q} &= \mathbf{q}^T(-\lambda\mathbf{Q}\mathbf{q}) = -\lambda\mathbf{q}^T\mathbf{Q}\mathbf{q} = -\lambda.
\end{aligned}
\tag{4.21}
$$

For any solution $(\lambda, \mathbf{q})$ of (4.20), the first equation means that λ should be an eigenvalue of $\mathbf{Q}^{-1}\mathbf{P}$, and $\mathbf{q}$ should be in the corresponding eigenspace. The second equation means that the objective function value at $\mathbf{q}$ is equal to $-\lambda$. Therefore, we can find an optimal solution $\mathbf{q}^*$ of (4.18) as follows: (i) find the maximum eigenvalue of $\mathbf{Q}^{-1}\mathbf{P}$, and (ii) find an eigenvector $\mathbf{q}^*$ in the corresponding eigenspace that satisfies $(\mathbf{q}^*)^T\mathbf{Q}\mathbf{q}^* - 1 = 0$. Once $\mathbf{q}^*$ is found, an optimal solution $\mathbf{a}^*$ of the original problem (4.17) is constructed as $\mathbf{a}^* = \mathbf{B}\mathbf{q}^*$.

4.5   Factor-of-Two Result

In this section, we demonstrate that the framing attack enables the adversary controlling only a half of a critical set of meters to perturb the state estimate by an arbitrary degree. Specifically, given a partition $\{\mathfrak{I}_1, \mathfrak{I}_2\}$ of a critical set of meters, we present a sufficient condition under which the adversary can control one of $\mathfrak{I}_1$ or $\mathfrak{I}_2$ to perturb the state estimate by an arbitrary degree. We provide numerical evidences from IEEE benchmark networks that for the critical sets associated with cuts, we can find a partition with $|\mathfrak{I}_1| \simeq |\mathfrak{I}_2|$ satisfying the sufficient condition.

4.5.1 Estimation of Adversarial State Estimate Perturbation

The exact analysis of how the framing attack would perturb the state estimate at the end of the iterative bad data processing is a difficult task. However, assuming that the meter SNRs are high, we can estimate the effect of the framing attack as follows. Since SNRs of most practical meters tend to be higher than 46 dB [76], the high meter SNR assumption is reasonable.

Suppose that the attacker adds the attack vector $\mathbf{a}$ to $\mathbf{z}$, and the bad data test is executed on $\bar{\mathbf{z}}$. The measurement vector in the kth iteration is

$$\bar{\mathbf{z}}^{(k)} = \mathbf{H}^{(k)}\mathbf{x} + \mathbf{a}^{(k)} + \mathbf{e}^{(k)}, \tag{4.22}$$

where $\mathbf{H}^{(k)}$, $\mathbf{a}^{(k)}$ and $\mathbf{e}^{(k)}$ are obtained from $\mathbf{H}$, $\mathbf{a}$ and $\mathbf{e}$ by removing the $(k-1)$ rows corresponding to the meters identified as bad until the $(k-1)$st iteration. The state estimate $\hat{\mathbf{x}}^{(k)}$ is

$$[(\mathbf{H}^{(k)})^T(\mathbf{\Sigma}^{(k)})^{-1}\mathbf{H}^{(k)}]^{-1}(\mathbf{H}^{(k)})^T(\mathbf{\Sigma}^{(k)})^{-1}\bar{\mathbf{z}}^{(k)}$$
$$= \mathbf{x} + [(\mathbf{H}^{(k)})^T(\mathbf{\Sigma}^{(k)})^{-1}\mathbf{H}^{(k)}]^{-1}(\mathbf{H}^{(k)})^T(\mathbf{\Sigma}^{(k)})^{-1}(\mathbf{a}^{(k)} + \mathbf{e}^{(k)}). \tag{4.23}$$

Hence, the state estimate perturbation after the kth iteration is

$$\hat{\mathbf{x}}^{(k)} - \mathbf{x} = [(\mathbf{H}^{(k)})^T(\mathbf{\Sigma}^{(k)})^{-1}\mathbf{H}^{(k)}]^{-1}(\mathbf{H}^{(k)})^T(\mathbf{\Sigma}^{(k)})^{-1}(\mathbf{a}^{(k)} + \mathbf{e}^{(k)}). \tag{4.24}$$

In addition, the residue vector is

$$\mathbf{r}^{(k)} = \mathbf{W}^{(k)}\bar{\mathbf{z}}^{(k)} = \mathbf{W}^{(k)}(\mathbf{H}^{(k)}\mathbf{x} + \mathbf{a}^{(k)} + \mathbf{e}^{(k)})$$
$$= \mathbf{W}^{(k)}(\mathbf{a}^{(k)} + \mathbf{e}^{(k)}). \tag{4.25}$$

From (4.24) and (4.25), we can see that both the state estimate perturbation and the residue vector do not depend on the actual state $\mathbf{x}$. Considering that bad

data detection and identification at each iteration exclusively rely on the residue vector, the observation from (4.24) and (4.25) implies that if we are interested in analyzing how much the attack *perturbs* the final state estimate, *i.e.*, $\hat{\mathbf{x}}^{(N)} - \mathbf{x}$, where N denotes the total number of iterations, we can simply work with $\mathbf{a} + \mathbf{e}$ by assuming that $\mathbf{x}$ is equal to $\mathbf{0}$.

Furthermore, if the meter SNRs are significantly large (*i.e.*, $\sigma^2 \ll 1$), we can estimate the resulting state perturbation by running the noiseless version of the bad data processing on the attack vector $\mathbf{a}$ and checking the resulting $\hat{\mathbf{x}}^{(N)}$. The *noiseless* version means the algorithm which the bad data processing converges to as σ^2 decays to 0. Specifically, $\boldsymbol{\Sigma}$ is replaced[3] by $\bar{\boldsymbol{\Sigma}}$, and in the kth iteration, the detector declares presence of bad data if and only if $(\mathbf{r}^{(k)})^T (\bar{\boldsymbol{\Sigma}}^{(k)})^{-1} \mathbf{r}^{(k)} > 0$ (*i.e.*, the data are declared to be good if and only if the state estimation results in a zero residue vector.)

4.5.2 Factor-of-Two Theorem for Critical Sets

Suppose that $\{\mathcal{I}_1, \mathcal{I}_2\}$ is a partition of a critical set, and let $\bar{\mathbf{H}}$ denote the measurement matrix after removing the meters in $\mathcal{I}_1 \cup \mathcal{I}_2$ from the grid. Since $\mathcal{I}_1 \cup \mathcal{I}_2$ is a critical set, $\bar{\mathbf{H}}$ has rank $n - 1$, and the dimension of its null space is one. Let $\Delta\mathbf{x}$ denote a unit basis vector of the null space of $\bar{\mathbf{H}}$. Recalling the discussion in Section 4.4.2, if $\mathcal{I}_1$ is the set of adversary meters, and $\mathcal{I}_2$ is the target set, then the framing attack aligns the attack vector along $\mathbf{H}_1 \Delta\mathbf{x}$, where $\mathbf{H}_1$ is the $m \times n$ matrix obtained from $\mathbf{H}$ by replacing the rows corresponding to the meters in $\mathcal{I}_2$ with zero

---

[3]Note that State Estimation and Bad Data Identification are not affected by the value of σ^2. Because, σ^2 gets cancelled out in the state estimate expression (4.12), and Bad Data Identification depends on the relative magnitudes of each residue with respect to other residues, which are not affected by the value of σ^2. Only Bad Data Detection is affected by the decaying σ^2.

row vectors ($\mathbf{H}_2$ is defined in the same way by replacing the rows corresponding to $\mathcal{I}_1$.)

The following theorem provides a sufficient condition that guarantees that the framing attack can use one of $\mathcal{I}_1$ and $\mathcal{I}_2$ to perturb the state estimate by an arbitrary degree under the high SNR setting. The condition is based on the result of running the deterministic test described in Section 4.5.1.

Theorem 4.5.1 *Suppose that if we run the noiseless version of the state estimation and the bad data processing on $\mathbf{H}_1\Delta\mathbf{x}$, then there exists a unique state $\mathbf{y} \in \mathbb{R}^n$ such that the final state estimate is always equal to $\mathbf{y}$ (i.e., $\hat{\mathbf{x}}^{(N)} = \mathbf{y}$) regardless of whatever decisions are made under tie[4] situations in Bad Data Identification. Under this condition, the following hold for any true state $\mathbf{x} \in \mathbb{R}^n$:*

(1) Suppose $\mathbf{y} \neq \mathbf{0}$. If the framing attack using $\mathcal{I}_1$ as adversary meters and $\mathcal{I}_2$ as target meters (i.e., $\mathbf{a} = \eta\mathbf{H}_1\Delta\mathbf{x}$ where $\eta \in \mathbb{R}$ is a scaling factor) is launched, then

$$\lim_{\sigma^2 \to 0} \Pr(\bar{\mathbf{z}}^{(N)} = \mathbf{H}^{(N)}(\mathbf{x} + \eta\mathbf{y}) + \mathbf{e}^{(N)}) = 1, \tag{4.26}$$

where N is the random variable representing the total number of iterations in the bad data processing.

(2) Suppose $\mathbf{y} \neq \Delta\mathbf{x}$. If the framing attack using $\mathcal{I}_2$ as adversary meters and $\mathcal{I}_1$ as target meters (i.e., $\mathbf{a} = \eta\mathbf{H}_2\Delta\mathbf{x}$) is launched, then

$$\lim_{\sigma^2 \to 0} \Pr(\bar{\mathbf{z}}^{(N)} = \mathbf{H}^{(N)}(\mathbf{x} + \eta(\Delta\mathbf{x} - \mathbf{y})) + \mathbf{e}^{(N)}) = 1. \tag{4.27}$$

---

[4]It is possible that a *tie* may occur in Bad Data Identification at some iteration: *i.e.*, the largest absolute normalized residue is assumed by more than one meter. In a tie situation, we assume that Bad Data Identification chooses an arbitrary meter with the largest absolute normalized residue.

Proof: See Section 4.7 ■

The event $\{\bar{\mathbf{z}}^{(N)} = \mathbf{H}^{(N)}(\mathbf{x} + \eta\mathbf{y}) + \mathbf{e}^{(N)}\}$ means that the final measurement vector at the end of the bad data processing is a noisy measurement vector with the state perturbed by $\eta\mathbf{y}$. Theorem 4.5.1 implies that if the condition is met, then at least one of $\mathfrak{I}_1$ and $\mathfrak{I}_2$ can be used by the framing attack to perturb the state estimate by an arbitrary degree, because $\mathbf{y}$ cannot be simultaneously $\mathbf{0}$ and $\Delta\mathbf{x}$. Especially, if the condition holds for the partition with $|\mathfrak{I}_1| = |\mathfrak{I}_2|$, then the adversary controlling only a half of the critical set can perturb the state estimate by an arbitrary degree.

One important question is whether a partition $\{\mathfrak{I}_1, \mathfrak{I}_2\}$ with $|\mathfrak{I}_1| \simeq |\mathfrak{I}_2|$ that satisfies the condition of Theorem 4.5.1 can be found in general. To answer this question, we investigated critical sets *associated with cuts*[5] in the IEEE 14-bus and 118-bus networks, where every bus has an injection meter and every line has line meters for both directions. The spanning tree observability criterion in [26] implies that the set $\mathfrak{I}$ of the meters associated with a cut (*i.e.*, the set of the line meters on the cut-set lines and the injection meters on the both ends of the cut-set lines) forms a critical set if removing the cutset decomposes the topology into two connected graphs. For instance, the cut in Fig. 4.2 disconnects the bus 3 from the rest of the network, and $\{\{2, 3\}, \{3, 4\}\}$ is the associated cut-set. The set of circled red meters is the critical set associated with the cut.

We executed 20,000 runs of the random contraction algorithm by Karger and Stein [77]—a randomized algorithm for finding a cut—and found 118 cuts in the 14-bus network and 290 cuts in the 118-bus network. For each cut, we built a

---

[5]A cut of an undirected graph $(\mathcal{V}, \mathcal{E})$ is defined as a partition $\{\mathcal{V}_1, \mathcal{V}_2\}$ of $\mathcal{V}$ consisting of two nonempty subsets, and the associated cut-set is the subset of lines connecting two vertices in different partitions.
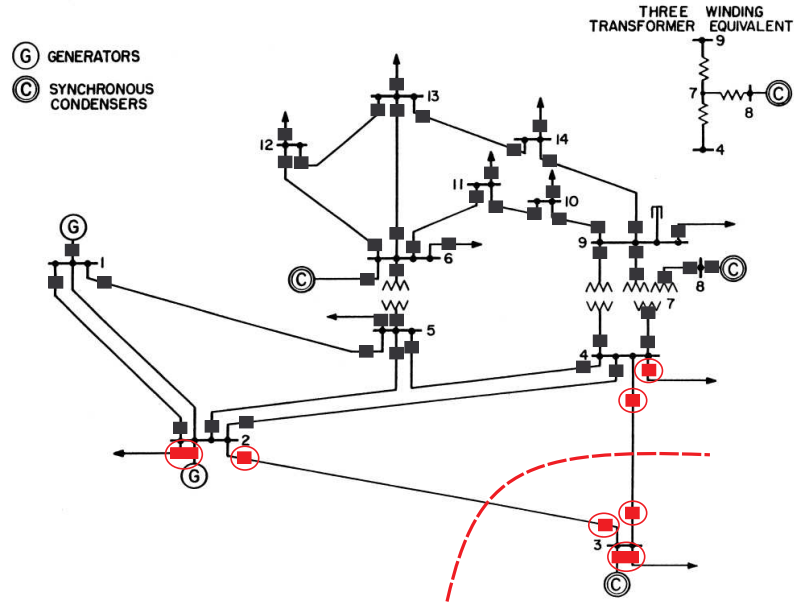
Figure 4.2: IEEE 14-bus network: the rectangles on lines and buses represent line flow meters and bus injection meters respectively. The line meter on the line $\{i,j\}$, that is closer to i, measures the power flow from i to j. The red dashed line describes a cut, and the circled meters are the meters associated with the cut.

partition $\{\mathcal{I}_1, \mathcal{I}_2\}$ of the critical set $\mathcal{I}$ associated with the cut such that $|\mathcal{I}_1| \simeq \frac{|\mathcal{I}|}{2}$: $\mathcal{I}_1$ consists of only the line meters (both directions) associated with a subset of lines in the cut-set such that $\left| |\mathcal{I}_1| - \frac{|\mathcal{I}|}{2} \right| \leq 1$, and $\mathcal{I}_2$ is set to be $\mathcal{I} \setminus \mathcal{I}_1$. In both networks, for every cut we considered, the partition constructed in the aforementioned manner satisfied the condition of Theorem 4.5.1; this suggests that the sufficient condition is not stringent, at least for critical sets associated with cuts[6].

_____

[6]The average size of the critical sets we considered is 15.7 for the 14-bus case and 12.7 for the 118-bus case.

4.6 Numerical Results

We tested the performance of the framing attack with the IEEE 14-bus and 118-bus networks under both the DC and AC models. The AC simulation results demonstrate the efficacy of the framing attack under the real-world power system setting. Because the ultimate goal of the attack is to perturb the state estimate, we measure the mean L_2-norm of the resulting state estimate error:

$$\mathbb{E}[\|\hat{\mathbf{x}} - \mathbf{x}\|_2],$$

where $\hat{\mathbf{x}}$ is the state estimate, and $\mathbf{x}$ is the true state.

4.6.1 Simulation Setting

In the IEEE 14-bus and 118-bus networks, we chose representative attack scenarios (*i.e.*, adversary meters and target meters) and tested the performance of the framing attack. For each case, we ran Monte Carlo simulations to evaluate the mean state estimate perturbation. In each Monte Carlo run, the true state $\mathbf{x}$ was generated by a multivariate Gaussian distribution with small variances. Its mean was set as the operating state given by the IEEE 14-bus and 118-bus data [78]. Based on the generated state $\mathbf{x}$, the noisy measurements were generated by the measurement model (*i.e.*, $h(\mathbf{x}) + \mathbf{e}$). The attack vector was constructed based on the DC measurement matrix $\mathbf{H}$ as described in Section 4.4. Once constructed, the atack vector was added to the noisy measurements, and state estimation and bad data processing[7] were executed on the corrupted measurements. After the bad data processing finished, we measured $\|\hat{\mathbf{x}}^{(N)} - \mathbf{x}\|_2$.

---

[7]The false alarm rate of the bad data detector is set to be 0.04 throughout all the simulations.

The main difference between the DC and AC simulations is that we used different measurement models for data generation. Note that the design of the framing attack was studied for the DC model which has only the real part of the measurements. For the AC simulations, we designed the attack vector based on the DC model, and the attack modified only the corresponding real part of the measurements. Considering the linear decoupled model (see Chapter 2.7 in [51]), such addition of the attack vector is expected to modify primarily the bus voltage phase angles and have little effect on the bus voltage magnitudes. Hence, in interpreting the AC results, we focus on the perturbation in the phase-angle part of the state estimate.

For comparison, we also executed the conservative scheme in [24], which aims to perturb the state estimate by the maximum degree while not raising any alarm in the bad data processing. This scheme has been considered as the best the adversary incapable of a covert state attack can do. In the conservative scheme, the attack vector was designed as a solution to

$$
\begin{aligned}
\max_{\mathbf{a} \in \mathcal{A}} \quad & \|(\mathbf{H}^T \mathbf{\Sigma}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{\Sigma}^{-1} \mathbf{a}\|_2^2 \\
\text{subj.} \quad & \mathbf{r}^T \mathbf{\Sigma}^{-1} \mathbf{r} \leq \tau,
\end{aligned}
\tag{4.28}
$$

where the constraint guarantees that the alarm is not raised at all, and the objective function is the resulting perturbation of the state estimate due to the attack vector.

4.6.2   Simulation Results with 14-Bus Network

We first tested the case where the adversary can control only a half of a critical set. Specifically, we considered the adversary who can control $(2, 3)$, $(3, 4)$, and $(4, 3)$: (i, j) denotes the line meter for the power flow from i to j, and (i) denotes the

injection meter at bus i. The target meters were set to be $(3, 2)$, (2), (3), and (4) such that the set of adversary meters and target meters is the critical set associated with the cut in Fig. 4.2. We tested the framing attack with three different attack magnitudes: $\|\mathbf{a}\|_1$ is 1%, 2%, or 3% of $\|\mathbf{z}\|_1$.

Fig. 4.3 shows the resulting state estimate perturbations versus the meter SNR in the DC simulations. The meter SNR ranges from 26 dB to 46 dB (equivalently, the noise-to-signal amplitude ratio ranges from 5% to 0.5%.) Note that the SNR range we tested is no greater than the SNR of most practical meters deployed in real-world power networks [76]. The normal state estimate error and the state estimate error under the conservative scheme are very close, and both decay to zero as the SNR increases. However, the state estimate error under the framing attack converges to a constant, which is proportional to the attack magnitude, as the SNR increases. The result implies that the framing attack can adjust the state estimate perturbation by choosing a proper attack magnitude. The effect of the framing attack becomes distinct from the normal state estimate error when the SNR is high. To demonstrate the relative effect of the framing attack with respect to the normal error, Fig. 4.4 shows the resulting state estimate perturbation normalized with respect to the state estimate error under the non-attack scenario. Under the same attack setting, Fig. 4.5 shows the state estimate perturbation versus the meter SNR in the AC simulations. It can be observed that, especially in the high SNR region, the perturbation amount is proportional to the attack magnitude. The plots imply that the effect of the framing attack persists in the AC model, thereby suggesting that the framing attack can be detrimental to the real-world power system state estimation.

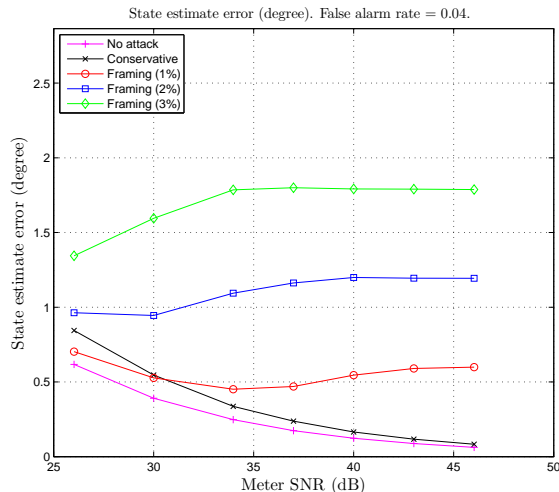Second, we demonstrate that the framing attack may pursue perturbation in

Figure 4.3: DC simulations with the 14-bus network: 1,000 Monte Carlo runs. The adversary meters are $(2,3)$, $(3,4)$, and $(4,3)$, and the target meters are $(3,2)$, (2), (3), and (4).
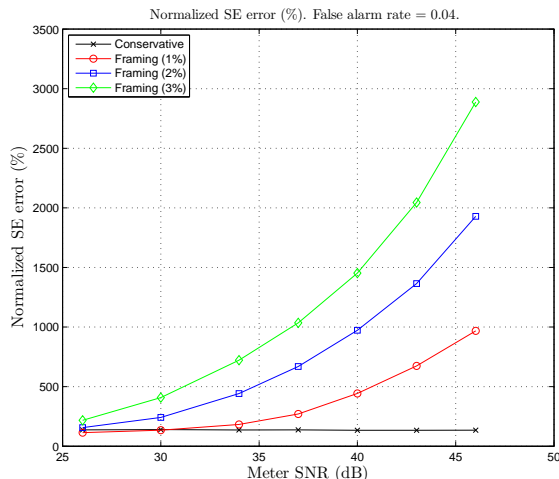

Figure 4.4: DC simulations with the 14-bus network: 1,000 Monte Carlo runs. The adversary meters are $(2,3)$, $(3,4)$, and $(4,3)$, and the target meters are $(3,2)$, (2), (3), and (4).

various directions by choosing a different target set. We considered the case that the adversary controls $(2,3)$, $(3,4)$, $(4,3)$, $(6,12)$, $(12,6)$, and $(12,13)$. Note that the adversary still cannot control any critical set, and thus a covert attack is infeasible. The framing attack with any of the following three different target sets successfully perturbed the state estimate: (i) (2), (3), (4), $(3,2)$, (6), (12), (13),
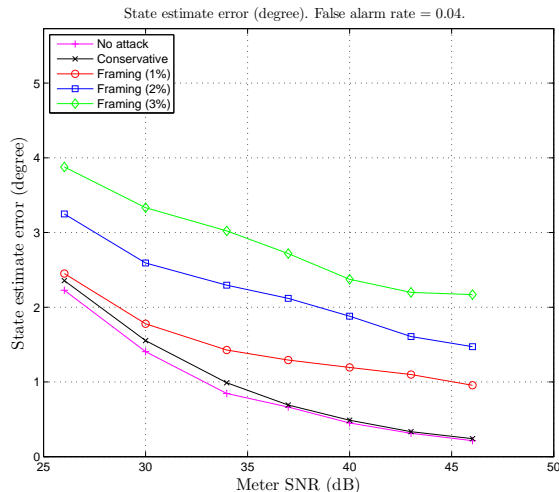
Figure 4.5: AC simulations with the 14-bus network: 1,000 Monte Carlo runs. The adversary meters are $(2, 3)$, $(3, 4)$, and $(4, 3)$, and the target meters are 2, 3, 4, $(3, 2)$.

and $(13, 12)$; (ii) (2), (3), (4), and $(3, 2)$; (iii) (6), (12), (13), and $(13, 12)$. For instance, Fig. 4.6 shows the state estimate perturbation versus the meter SNR in the AC simulations for the first target set. While the three target sets all resulted in successful state estimate perturbation, each resulted in a different direction of perturbation. For each target set, Table 4.2 shows the three buses, whose phase angle estimates were most significantly perturbed, and the mean perturbation of their phase angle estimates; positive perturbation means overestimation, and negative perturbation means underestimation. The table demonstrates that the adversary controlling a large number of meters may adjust the direction of perturbation by choosing a proper target set. Note that with the second target set, whose associated critical set (*i.e.*, the critical set contained in $\mathcal{I}_A \cup \mathcal{I}_T$) isolates bus 3, the framing attack perturbed the bus-3 phase angle estimate significantly while having little effect on other bus phase angle estimates. This is expected because once the target meters are successfully removed from the network, the adversary can control *all* the real meter measurements that depend on the bus-3 phase angle. The similar effect can be observed for the framing attack with the third target
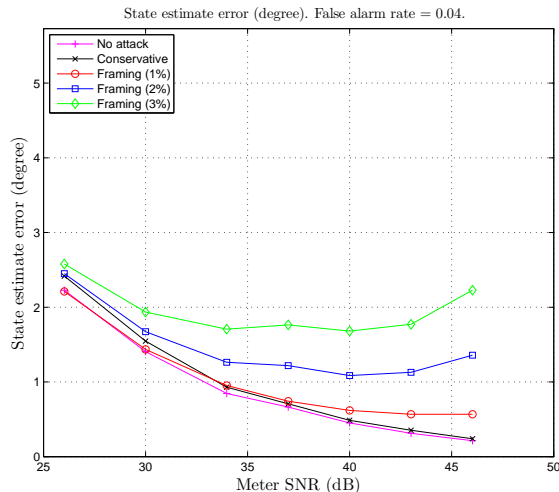
120

State estimate error (degree). False alarm rate = 0.04.

Figure 4.6: AC simulations with the 14-bus network: 1,000 Monte Carlo runs. The adversary meters are $(2,3)$, $(3,4)$, $(4,3)$, $(6,12)$, $(12,6)$, and $(12,13)$, and the target meters are 2, 3, 4, $(3,2)$, 6, 12, 13, and $(13,12)$.

Table 4.2: The three buses whose phase angles are most significantly perturbed by each attack: AC simulations, 1,000 Monte Carlo runs, SNR = 46dB.

(2), (3), (4), $(3,2)$, (6), (12), (13), $(13,12)$	(2), (3) (4), $(3,2)$	(6), (12), (13), $(13,12)$
1) bus 12: $2.075°$	1) bus 3: $-2.183°$	1) bus 12: $2.878°$
2) bus 3: $0.272°$	2) bus 14: $0.182°$	2) bus 14: $0.005°$
3) bus 14: $-0.180°$	3) bus 9: $0.168°$	3) bus 9: $0.004°$

set, whose associated cut isolates bus 12. On the other hand, for the first target set, once the target meters are removed, the adversary controls all the real meter measurements that depend on the bus-3 phase angle or the bus-12 phase angle. In this case, the framing attack, constructed by the QCQP framework in (4.17), perturbed both bus-3 and bus-12 phase angle estimates.
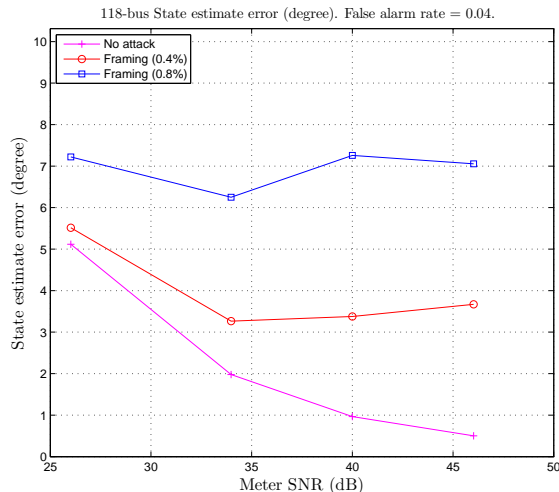
Figure 4.7: AC simulations with the 118-bus network: 250 Monte Carlo runs. The adversary meters are $(20, 21)$, $(21, 20)$, and $(21, 22)$, and the target meters are (20), (21), (22), and $(22, 21)$.

4.6.3 Simulation Results with 118-Bus Network

Through the simulations with the 118-bus network, we aim to demonstrate the effect of the framing attack on a larger network. We considered the scenario where the adversary controls $(20, 21)$, $(21, 20)$, and $(21, 22)$, and the target meters are (20), (21), (22), and $(22, 21)$; *i.e.*, the set of the adversary meters and the target meters is the critical set associated with the cut isolating the bus 21 from the rest of the network. Fig. 4.7 shows the state estimate errors under the non-attack scenario and the framing attacks with different attack magnitudes in the AC simulations. The plots imply that the framing attack successfully perturbs the state estimate, and thus its adversarial effect persists in a larger network.

4.7 Proof of Theorem 4.5.1

Let $\mathcal{S}$ denote the set of sequences of meter removals that can possibly happen when the noiseless version of bad data processing is executed on $\mathbf{H}_1\Delta\mathbf{x}$: *i.e.*, $(a_1, \ldots, a_M) \in \mathcal{S}$ if and only if some decisions under tie situations may result in the removal of the meters $\{a_1, \ldots, a_M\}$ in the order of $a_1, \ldots, a_M$. The cardinality of $\mathcal{S}$ can be greater than 1 since different decisions under tie situations may result in different sequences of meter removals.

For any sequence $(a_1, \ldots, a_M) \in \mathcal{S}$, the existence of such $\mathbf{y}$—as described in the condition—implies that if all the meters in the sequence are removed, the remaining part of $\mathbf{H}_1\Delta\mathbf{x}$, denoted by $\mathbf{H}_1^{(M)}\Delta\mathbf{x}$, is equal to $\mathbf{H}^{(M)}\mathbf{y}$, where $\mathbf{H}_1^{(M)}$ and $\mathbf{H}^{(M)}$ are obtained from $\mathbf{H}_1$ and $\mathbf{H}$ respectively, by removing the rows corresponding to all meters in the sequence.

Now, consider running the bad data test on $\mathbf{H}\mathbf{x} + \mathbf{H}_1\Delta\mathbf{x} + \mathbf{e}$. The equation (4.25) implies that the residue vector in each iteration only depends on $\mathbf{H}_1\Delta\mathbf{x} + \mathbf{e}$. In addition, as σ^2 decreases to zero, the results of bad data detection and identification heavily depend on $\mathbf{H}_1\Delta\mathbf{x}$, and thus the sequence of removed meters becomes highly likely to be in $\mathcal{S}$. Formally,

$$\lim_{\sigma^2 \to 0} \Pr((a_1, \ldots, a_N) \in \mathcal{S}) = 1, \tag{4.29}$$

where $(a_1, \ldots, a_N)$ is a random sequence of meters removed by the bad data test. Let $\mathbf{H}^{(N)}$ and $\mathbf{e}^{(N)}$ denote the random matrix and vector obtained from $\mathbf{H}$ and $\mathbf{e}$ respectively by removing the rows corresponding to $\{a_1, \ldots, a_N\}$.

The event $\{(a_1, \ldots, a_N) \in \mathcal{S}\}$ implies that $\mathbf{H}_1^{(N)}\Delta\mathbf{x} = \mathbf{H}^{(N)}\mathbf{y}$, and thus

$$\bar{\mathbf{z}}^{(N)} = (\mathbf{H}^{(N)}\mathbf{x} + \mathbf{e}^{(N)}) + \mathbf{H}_1^{(N)}\Delta\mathbf{x} = (\mathbf{H}^{(N)}\mathbf{x} + \mathbf{e}^{(N)}) + \mathbf{H}^{(N)}\mathbf{y}. \tag{4.30}$$

123

Therefore,

$$\lim_{\sigma^2 \to 0} \Pr(\bar{\mathbf{z}}^{(N)} = (\mathbf{H}^{(N)}\mathbf{x} + \mathbf{e}^{(N)}) + \mathbf{H}^{(N)}\mathbf{y}) = 1. \qquad (4.31)$$

Note that replacing the attack vector $\mathbf{H}_1\Delta\mathbf{x}$ with $\mathbf{H}_1\eta\Delta\mathbf{x}$ simply changes the above to

$$\lim_{\sigma^2 \to 0} \Pr(\bar{\mathbf{z}}^{(N)} = (\mathbf{H}^{(N)}\mathbf{x} + \mathbf{e}^{(N)}) + \mathbf{H}^{(N)}\eta\mathbf{y}) = 1. \qquad (4.32)$$

Now, consider running the bad data test over $\mathbf{H}\mathbf{x} + \mathbf{H}_2\Delta\mathbf{x} + \mathbf{e}$; this is the case when the framing attack is launched with the partition $\mathcal{I}_2$. First, note that

$$\mathbf{H}\Delta\mathbf{x} = \mathbf{H}_1\Delta\mathbf{x} + \mathbf{H}_2\Delta\mathbf{x}. \qquad (4.33)$$

Therefore, running the bad data test on $\mathbf{H}\mathbf{x} + \mathbf{H}_2\Delta\mathbf{x} + \mathbf{e}$ is equivalent to running it on $\mathbf{H}(\mathbf{x} + \Delta\mathbf{x}) - \mathbf{H}_1\Delta\mathbf{x} + \mathbf{e}$.

Suppose we run the noiseless version of the bad data processing on $-\mathbf{H}_1\Delta\mathbf{x}$. The set of sequences of meter removals that can possibly happen is equivalent to $\mathcal{S}$, because the sign change only flips the signs of residue entries; it does not affect their absolute values, which are the statistics used for detection and identification of bad data entries. Furthermore, it can be easily seen that the final state estimate is always equal to $-\mathbf{y}$ regardless of whatever decisions are made under the tie situations.

Now, consider again running the bad data test on $\mathbf{H}(\mathbf{x} + \Delta\mathbf{x}) - \mathbf{H}_1\Delta\mathbf{x} + \mathbf{e}$, which is equivalent to $\mathbf{H}\mathbf{x} + \mathbf{H}_2\Delta\mathbf{x} + \mathbf{e}$. In exactly the same manner as we derived (4.31), we can derive the following:

$$\lim_{\sigma^2 \to 0} \Pr(\bar{\mathbf{z}}^{(N)} = \mathbf{H}^{(N)}(\mathbf{x} + \Delta\mathbf{x}) + \mathbf{e}^{(N)} + \mathbf{H}^{(N)}(-\mathbf{y})) = 1, \qquad (4.34)$$

or equivalently,

$$\lim_{\sigma^2 \to 0} \Pr(\bar{\mathbf{z}}^{(N)} = \mathbf{H}^{(N)}\mathbf{x} + \mathbf{e}^{(N)} + \mathbf{H}^{(N)}(\Delta\mathbf{x} - \mathbf{y})) = 1. \qquad (4.35)$$

When the attack vector $\mathbf{H}_2\Delta\mathbf{x}$ is scaled by η (*i.e.*, $\mathbf{a} = \mathbf{H}_2\eta\Delta\mathbf{x}$), repeating the same steps as above, we can easily derive the following:

$$\lim_{\sigma^2 \to 0} \Pr(\bar{\mathbf{z}}^{(N)} = \mathbf{H}^{(N)}\mathbf{x} + \mathbf{e}^{(N)} + \mathbf{H}^{(N)}\eta(\Delta\mathbf{x} - \mathbf{y})) = 1. \tag{4.36}$$

Therefore, the proof is complete. ■

CHAPTER 5

CONCLUSIONS

In this dissertation, we studied attacks and countermeasures in communications and power networks. Specifically, stepping-stone attacks in communications networks and data attacks in power networks were considered.

Although the attack framework varied significantly depending on the network type and the adversary's goal, there were some common observations. In particular, it was commonly observed in all attack problems that an adversary having enough control on network operations can achieve its goal while hiding its presence. Therefore, a challenging task for a network administrator is to find an affordable protection strategy that can prevent such undetectable attacks. The main contribution of dissertation was to provide conditions on the adversary's ability under which attacks can be detected, build network protection strategies based on the detectability condition, and study what attacks can achieve if a network is not secured properly.

In the following sections, we provide concluding remarks for each topic and comments for future works.

5.1 Detection of Information Flows

In Chapter 2, we have studied timing-based detection of information flows in a network. We formulate flow detection as a binary composite hypothesis testing problem and present a detector that requires neither a parametric model nor a training data set. The detector requires a constant memory, and it has linear computational complexity with respect to the sample size. The simulations with

real-world TCP and VoIP traces demonstrate that the proposed detector is superior to the benchmark passive detectors and more suitable for the nonparametric and unsupervised setting.

The test results with the real-world traces suggest that the proposed detector may perform well in a more practical setting. Nevertheless, as all other passive detectors, our detector has a fundamental limit that it cannot detect the flow if the presence of the flow generates no correlation between the timing measurements at all.

5.2   Topology Attack of a Power Grid

In Chapter 3, we have considered undetectable malicious data attack aimed at creating a false topology at the control center. We obtain a necessary and sufficient condition for an attack launched by a strong attacker to be undetectable. We also present a class of undetectable line removal attacks that can be launched by weak attackers with only local information. Finally, we present a countermeasure against strong attackers by protecting a subset of meters.

Some of the results presented in Chapter 3 are obtained under strong conditions. Here, we mention several of such limitations as pointers for further study. First, the DC model assumed in Section 3.3 makes the results valid only near the operating point. It has been demonstrated in [39] that the DC model tends to exaggerate the effect of state attacks, and the nonlinear state estimator has the ability to significantly reduce the attacks' impact on the state estimate. Obtaining conditions for undetectable topology attacks under the AC model is of considerable interest.

Second, we have focused mostly on state-preserving topology attacks. Even though such attacks are optimal under certain scenarios, to understand the full implication of topology attacks, it is necessary to consider attacks that affect both topology and states.

Finally, we consider only one particular form of countermeasure, namely implementing authentication at a subset of meters. Other mechanisms should be studied, including one with more sophisticated bad data detection and those taking into accounts of system dynamics.

5.3   Data Framing Attack on State Estimation

In Chapter 4, we have presented the data framing attack on power system state estimation. Controlling only a half of a critical set, the data framing attack can perturb the state estimate by an arbitrary degree. A theoretical justification was provided, and numerical experiments demonstrated the efficacy of the framing attack.

Our results indicate that most known countermeasures, that are aimed at merely preventing covert state attacks, are not sufficient for protection against the attacks aimed at state perturbation. In designing countermeasures, the possibility of the framing attack needs to be taken into account.

One important direction for future work is to find an easily verifiable necessary condition for the framing attack to succeed with given adversary meters. Such a condition is essential for designing a countermeasure.

BIBLIOGRAPHY

[1] D. Donoho, A. Flesia, U. Shankar, V. Paxson, J. Coit, and S. Staniford, "Multiscale stepping-stone detection: Detecting pairs of jittered interactive streams by exploiting maximum tolerable delay," in *5th International Symposium on Recent Advances in Intrusion Detection, LNCS vol. 2516*, Zurich, Switzerland, Oct 2002.

[2] X. Wang and D. Reeves, "Robust correlation of encrypted attack traffic through stepping stones by manipulation of inter-packet delays," in *Proc. of the 2003 ACM Conference on Computer and Communications Security*, 2003, pp. 20–29.

[3] P. Peng, P. Ning, D. Reeves, and X. Wang, "Active Timing-Based Correlation of Perturbed Traffic Flows with Chaff Packets," in *Proc. 25th IEEE International Conference on Distributed Computing Systems Workshops*, Columbus, OH, June 2005, pp. 107–113.

[4] X. Wang, S. Chen, and S. Jajodia, "Tracking anonymous peer-to-peer voip calls on the internet," in *Proc. of the 2005 ACM Conference on Computer and Communications Security*, Alexandra, VA, nov 2005.

[5] Y. H. Park and D. S. Reeves, "Adaptive Watermarking Against Deliberate Random Delay for Attack Attribution Through Stepping Stones," in *Proc. of the Ninth International Conference on Information and Communications Security (ICICS 2007)*, dec 2007.

[6] Y. J. Pyun, Y. H. Park, X. Wang, D. Reeves, and P. Ning, "Tracing Traffic through Intermediate Hosts that Repacketize Flows," in *INFOCOM 2007. 26th IEEE International Conference on Computer Communications. IEEE*, may 2007.

[7] D. Ramsbrock, X. Wang, and X. Jiang, "A First Step towards Live Botmaster Traceback," in *Proceedings of the 11th international symposium on Recent Advances in Intrusion Detection*, Cambridge, MA, 2008.

[8] A. Houmansadr, N. Kiyavash, and N. Borisov, "RAINBOW: A Robust And Invisible Non-Blind Watermark for Network Flows," in *Proc. of the 16th Annual Network and Distributed System Security Symposium*, San Diego, CA, Feb 2009.

[9] A. Houmansadr and N. Borisov, "SWIRL: A Scalable Watermark to Detect Correlated Network Flows," in *Proc. of the 18th Annual Network and Distributed System Security Symposium*, San Diego, CA, Feb 2011.

[10] X. Wang and D. S. Reeves, "Robust Correlation of Encrypted Attack Traffic through Stepping Stones by Flow Watermarking," *IEEE Transactions on Dependable and Secure Computing*, vol. 8, no. 3, pp. 434 –449, may-june 2011.

[11] L. Zhang, A. Persaud, A. Johnson, and Y. Guan, "Stepping Stone Attack Attribution in Non-Cooperative IP Networks," Iowa State University, Tech. Rep. TR-2005-02-1, Feb. 2005.

[12] ——, "Detection of Stepping Stone Attack under Delay and Chaff Perturbations," in *Proc. of The 25th IEEE International Performance Computing and Communications Conference*, Phoenix, AZ, Apr. 2006.

[13] A. Blum, D. Song, and S. Venkataraman, "Detection of Interactive Stepping Stones: Algorithms and Confidence Bounds," in *7th International Symposium on Recent Advance in Intrusion Detection (RAID)*, Sophia Antipolis, French Riviera, France, September 2004.

[14] T. He and L. Tong, "Detection of Information Flows," *IEEE Transactions on Information Theory*, vol. 54, pp. 4925–4945, Nov. 2008.

[15] B. Coskun and N. Memon, "Efficient Detection of Delay-Constrained Relay Nodes," in *Proceedings of the 2007 Annual Computer Security Applications Conference*, dec 2007, pp. 353–362.

[16] ——, "Online Sketching of Network Flows for Real-Time Stepping-Stone Detection," in *Proceedings of the 2009 Annual Computer Security Applications Conference*, Washington, DC, 2009, pp. 473–483.

[17] V. Paxson and S. Floyd, "Wide-Area Traffic: The Failure of Poisson Modeling," *IEEE/ACM Transactions on Networking*, vol. 3, no. 3, pp. 226–244, June 1995.

[18] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," in *Proceedings of the 16th ACM conference on Computer and communications security*, 2009, pp. 21–32.

[19] R. B. Bobba, K. M. Rogers, Q. Wang, H. Khurana, K. Nahrstedt, and T. J. Overbye, "Detecting false data injection attacks on dc state estimation,"

in *First Workshop on Secure Control Systems,CPSWEEK 2010*, Stockholm, Sweeden, Apr 2010.

[20] H. Sandberg, A. Teixeira, and K. H. Johansson, "On security indices for state estimators in power networks," in *First Workshop on Secure Control Systems,CPSWEEK 2010*, Stockholm, Sweeden, Apr 2010.

[21] G. Dán and H. Sandberg, "Stealth attacks and protection schemes for state estimators in power systems," in *Proc. IEEE 2010 SmartGridComm*, Gaithersburg, MD, USA., Oct 2010.

[22] O. Vukovic, K. C. Sou, G. Dan, and H. Sandberg, "Network-layer protection schemes against stealth attacks on state estimators in power systems," in *Smart Grid Communications (SmartGridComm), 2011 IEEE International Conference on*, oct. 2011, pp. 184 –189.

[23] O. Kosut, L. Jia, R. J. Thomas, and L. Tong, "Malicious data attacks on smart grid state estimation: attack strategies and countermeasures," in *Proc. IEEE 2010 SmartGridComm*, Gaithersburg, MD, USA, Oct 2010.

[24] ——, "Malicious data attacks on the smart grid," *IEEE Transactions on Smart Grid*, vol. 2, no. 4, pp. 645 –658, dec. 2011.

[25] A. Monticelli and F. F. Wu, "Network observability: Theory," *IEEE Trans. Power Apparatus and Systems*, vol. PAS-104, no. 5, pp. 1042–1048, May 1985.

[26] G. R. Krumpholz, K. A. Clements, and P. W. Davis, "Power system observability: a practical algorithm using network topology," *IEEE Trans. Power Apparatus and Systems*, vol. 99, no. 4, pp. 1534–1542, July 1980.

[27] K. Clements and P. Davis, "Detection and identification of topology errors in electric power systems," *IEEE Transactions on Power Systems*, vol. 3, no. 4, pp. 1748 –1753, nov 1988.

[28] F. F. Wu and W. E. Liu, "Detection of topology errors by state estimation," *IEEE Trans. Power Systems*, vol. 4, no. 1, pp. 176–183, Feb 1989.

[29] I. Costa and J. Leao, "Identification of topology errors in power system state estimation," *IEEE Transactions on Power Systems*, vol. 8, no. 4, pp. 1531 –1538, nov 1993.

[30] A. Monticelli, "Modeling circuit breakers in weighted least squares state estimation," *IEEE Transactions on Power Systems*, vol. 8, no. 3, pp. 1143 –1149, aug 1993.

[31] A. Abur, H. Kim, and M. Celik, "Identifying the unknown circuit breaker statuses in power networks," *IEEE Transactions on Power Systems*, vol. 10, no. 4, pp. 2029 –2037, nov. 1995.

[32] L. Mili, G. Steeno, F. Dobraca, and D. French, "A robust estimation method for topology error identification," *IEEE Transactions on Power Systems*, vol. 14, no. 4, pp. 1469 –1476, nov 1999.

[33] E. Lourenco, A. Costa, and K. Clements, "Bayesian-based hypothesis testing for topology error identification in generalized state estimation," *IEEE Transactions on Power Systems*, vol. 19, no. 2, pp. 1206 – 1215, may 2004.

[34] A. Jaen, P. Romero, and A. Exposito, "Substation data validation by a local three-phase generalized state estimator," *IEEE Transactions on Power Systems*, vol. 20, no. 1, pp. 264 – 271, feb. 2005.

[35] F. Vosgerau, A. Simoes Costa, K. Clements, and E. Lourenco, "Power system state and topology coestimation," in *Bulk Power System Dynamics and Control (iREP) - VIII (iREP), 2010 iREP Symposium*, aug. 2010, pp. 1 –6.

[36] A. L. Ott, "Experience with pjm market operation, system design, and implementation," *IEEE Trans. Power Systems*, vol. 18, no. 2, pp. 528–534, May 2003.

[37] L. Xie, Y. Mo, and B. Sinopoli, "False data injection attacks in electricity markets," in *Proc. IEEE 2010 SmartGridComm*, Gaithersburg, MD, USA., Oct 2010.

[38] L. Jia, R. J. Thomas, and L. Tong, "Malicious data attack on real-time electricity market," in *Proc. 2011 IEEE Intl. Conf. Acoust. Speech & Sig. Proc. (ICASSP)*, Prague, Czech Republic, May 2011.

[39] ——, "On the Nonlinearity Effects on Malicious Data Attack on Power System," in *Power and Energy Society General Meeting, 2012 IEEE*, july 2012.

[40] E. Handschin, F. C. Schweppe, J. Kohlas, and A. Fiechter, "Bad data analysis for power system state estimation," *IEEE Trans. Power Apparatus and Systems*, vol. PAS-94, no. 2, pp. 329–337, Mar/Apr 1975.

[41] T. Van Cutsem, M. Ribbens-Pavella, and L. Mili, "Bad data identification methods in power system state estimation-a comparative study," *IEEE Transactions on Power Apparatus and Systems*, vol. 104, no. 11, pp. 3037–3049, 1985.

[42] A. Monticelli and F. F. Wu, "Network observability: Identification of observable islands and measurement placement," *IEEE Trans. Power Apparatus and Systems*, vol. PAS-104, no. 5, pp. 1035–1041, May 1985.

[43] T. Kim and H. Poor, "Strategic protection against data injection attacks on power grids," *IEEE Transactions on Smart Grid*, vol. 2, no. 2, pp. 326 –333, june 2011.

[44] S. Bi and Y. Zhang, "Defending mechanisms against false-data injection attacks in the power system state estimation," in *2011 IEEE GLOBECOM Workshops*, Houston, TX, USA., Dec 2011.

[45] A. Giani, E. Bitar, M. Garcia, M. McQueen, P. Khargonekar, and K. Poolla, "Smart grid data integrity attacks: characterizations and countermeasures," in *2011 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, Oct 2011, pp. 232–237.

[46] J. Kim and L. Tong, "On topology attack of a smart grid: undetectable attacks and countermeasures," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 7, July 2013.

[47] K. Clements and P. Davis, "Multiple bad data detectability and identifiability: A geometric approach," *IEEE Transactions on Power Delivery*, vol. 1, no. 3, pp. 355–360, 1986.

[48] G. Korres and G. Contaxis, "Identification and updating of minimally dependent sets of measurements in state estimation," *IEEE Transactions on Power Systems*, vol. 6, no. 3, pp. 999–1005, 1991.

[49] A. Monticelli and A. Garcia, "Reliable bad data processing for real-time state estimation," *IEEE Transactions on Power Apparatus and Systems*, vol. 102, no. 5, pp. 1126–1139, 1983.

[50] L. Mili, T. Van Cutsem, and M. Ribbens-Pavella, "Hypothesis testing identification: A new method for bad data analysis in power system state estimation," *IEEE Transactions on Power Apparatus and Systems*, vol. 103, no. 11, pp. 3239–3252, 1984.

[51] A. Abur and A. G. Expósito, *Power System State Estimation: Theory and Implementation*. CRC, 2000.

[52] A. Monticelli, *State Estimation in Electric Power Systems: A Generalized Approach (Power Electronics and Power Systems)*. Springer US, 1999.

[53] W. W. Kotiuga and M. Vidyasagar, "Bad data rejection properties of weughted least absolute value techniques applied to static state estimation," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-101, no. 4, pp. 844–853, 1982.

[54] A. Abur and M. Celik, "A fast algorithm for the weighted least absolute value state estimation [for power systems]," *IEEE Transactions on Power Systems*, vol. 6, no. 1, pp. 1–8, 1991.

[55] M. Celik and A. Abur, "A robust wlav state estimator using transformations," *IEEE Transactions on Power Systems*, vol. 7, no. 1, pp. 106–113, 1992.

[56] H. Singh and F. Alvarado, "Weighted least absolute value state estimation using interior point methods," *IEEE Transactions on Power Systems*, vol. 9, no. 3, pp. 1478–1484, 1994.

[57] L. Mili, M. Cheniae, and P. Rousseeuw, "Robust state estimation of electric power systems," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 41, no. 5, pp. 349–358, 1994.

[58] M. Cheniae, L. Mili, and P. Rousseeuw, "Identification of multiple interacting bad data via power system decomposition," *IEEE Transactions on Power Systems*, vol. 11, no. 3, pp. 1555–1563, 1996.

[59] K. Morrow, E. Heine, K. Rogers, R. Bobba, and T. Overbye, "Topology perturbation for detecting malicious data injection," in *2012 45th Hawaii International Conference on System Science (HICSS)*, Jan 2012, pp. 2104–2113.

[60] A. Tajer, S. Kar, H. Poor, and S. Cui, "Distributed joint cyber attack detection and state recovery in smart grids," in *2011 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, Oct 2011, pp. 202–207.

[61] Y. Huang, H. Li, K. Campbell, and Z. Han, "Defending false data injection attack on smart grid network using adaptive cusum test," in *2011 45th Annual*

Conference on Information Sciences and Systems (CISS), march 2011, pp. 1–6.

[62] S. Cui, Z. Han, S. Kar, T. Kim, H. Poor, and A. Tajer, "Coordinated data-injection attack and detection in the smart grid: A detailed look at enriching detection solutions," *IEEE Signal Processing Magazine*, vol. 29, no. 5, pp. 106–115, 2012.

[63] Y. Huang, M. Esmalifalak, H. Nguyen, R. Zheng, Z. Han, H. Li, and L. Song, "Bad data injection in smart grid: attack and defense mechanisms," *IEEE Communications Magazine*, vol. 51, no. 1, pp. 27–33, 2013.

[64] P. Kruus, D. Sterne, R. Gopaul, M. Heyman, B. Rivera, P. Budulas, B. Luu, T. Johnson, N. Ivanic, and G. Lawler, "In-Band Wormholes and Countermeasures in OLSR Networks," in *2nd International Conference on Security and Privacy in Communication Networks (SecureComm 2006)*, Baltimore, MD, Aug. 2006.

[65] T. Chothia and K. Chatzikokolakis, "A Survey of Anonymous Peer-to-Peer File-Sharing," in *Embedded and Ubiquitous Computing Workshops, LNCS vol. 3823*, 2005, pp. 744–755.

[66] J. Ren and J. Wu, "Survey on anonymous communications in computer networks," *Computer Communications*, vol. 33, no. 4, pp. 420–431, March 2010.

[67] ITU-T Recommendation G.114, "One way transmission time," 2003.

[68] J. Shao, *Mathematical Statistics.* Springer, 2003.

[69] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Transactions on Networking*, vol. 2, no. 1, pp. 1–15, Feb 1994.

[70] M. E. Crovella and A. Bestavros, "Self-similarity in World Wide Web traffic: evidence and possible causes," *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, pp. 835–846, Dec 1997.

[71] Z. Sahinoglu and S. Tekinay, "On multimedia networks: self-similar traffic and network performance," *IEEE Communications Magazine*, vol. 37, no. 1, pp. 48 – 52, jan 1999.

[72] "Vulnerability Analysis of Energy Delivery Control Systems," Idaho National Laboratory, September 2011, INL/EXT-10-18381.

[73] C. Paar and J. Pelzl, *Understanding Cryptography: A Textbook for Students and Practitioners.* Springer, 2010.

[74] O. Alsac, N. Vempati, B. Stott, and A. Monticelli, "Generalized state estimation," *IEEE Transactions on Power Systems*, vol. 13, no. 3, pp. 1069 –1075, aug 1998.

[75] R. Christensen, *Plane answers to complex questions: the theory of linear models.* Springer, 2011.

[76] "Accuracy of Digital Electricity Meters," Electric Power Research Istitute white paper, May 2010.

[77] D. R. Karger and C. Stein, "A new approach to the minimum cut problem," *Journal of the ACM*, vol. 43, no. 4, pp. 601–640, Jul. 1996.

[78] "Power Systems Test Case Archive." [Online]. Available: http://www.ee.washington.edu/research/pstca/