Cyberintrusion Detection in Critical Infrastructure

by

Nancy Bissinger

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
December 12, 2015

Keywords: Cyberintrusion, Power Systems, Natural Gas Transmission, Principal Component Analysis, Multivariate Analysis

Approved by

Jorge Valenzuela, Chair, Professor of Industrial and Systems Engineering
Saeed Maghsoodloo, Professor Emeritus of Industrial and Systems Engineering
Chan Park, Daniel F. and Josephine Breeden Professor of Industrial and Systems Engineering
Jianhui Wang, Affiliated Professor of Industrial and Systems Engineering

Abstract

Sophisticated cyberterrorists have sufficient knowledge to devise an attack through the Internet which could compromise critical resource delivery. As the threat of such cybercrime escalates, defending critical infrastructure is a primary focus of the United States government, industry executives, and the research community. Current research and development primarily focuses on preventing the cyberterrorist from accomplishing his mission of disruption. This research focuses not on prevention, but on *detection*. Its main objective is the development of an algorithm that can be used to detect data anomalies which may be the result of security breaches.

Grounded in multivariate statistical process control, the algorithm uses principal component analysis to separate data variability into common-cause and assignable-cause subspaces. Analysis using the common-cause subspace determines whether the data has been compromised. Successful results will add a dimension of protection for critical infrastructure systems that has not previously been addressed in the literature. Implementation of the algorithm in a process control system could significantly improve the security of operational and planning practices today and in the future. In process control operations, wireless transmission of measurements could be interrupted or data storage in databases on the TCP/IP network could be corrupted or compromised through malware or other human interference. Data errors resulting from any of these occurrences could disrupt physical processes in critical infrastructure. This innovative algorithm provides a solution to this problem.

Acknowledgments

I dedicate this dissertation to all of those mentors, friends, and family members who have helped me to be successful.

Dr. Valenzuela and my committee—Dr. Park, Dr. Maghsoodloo, Dr. Wang— inspired me and taught me things that books could not. The ISE faculty helped me remember how much I love learning and provided the academic environment necessary for my success. My friends, both in Auburn and in New Orleans, were there when I needed someone with whom to share exciting news or a shoulder to cry on. My husband, Allan, coped with my time away from home and helped me to understand the physics of electric power. My children—Brook, Brett, and Brandon made me feel special for even attempting this feat and Brett shared his knowledge of statistical techniques and his educational experience. My extended family gave me the strength that only they can give. Everyone encouraged and supported me and without all of you, this dream would not now be a reality. Thank you all.

In addition, I would like to thank Dr. Jianhui Wang and Argonne National Laboratory for the research topic and financial support.

Table of Contents

List of Figures

List of Tables

List of Symbols

β	Energy price (K-euro per kW)
$\bar{\mathbf{x}}$	Mean of variables
σ	Standard deviation of variables
$\mathbf{C}$	Cost of electric power generation (\$ per MW)
$\mathbf{D}$	Inner diameter of a pipeline (mm)
$\mathbf{d}$	Real power demand (MW)
$\mathbf{f^P}$	Real power flow (MW)
$\mathbf{F^{MAX}}$	Transmission line capacity (MVA)
$\mathbf{G^{MAX}}$	Real power generation maximum (MW)
$\mathbf{G^{MIN}}$	Real power generation (MW)
$\mathbf{H}$	Adjacency matrix
$\mathbf{L^G}$	Length of a pipeline (km)
$\mathbf{L}$	Eigenvalues of covariance matrix
$\mathbf{Q^{MAX}}$	Reactive power generation maximum (VAR)
$\mathbf{Q^{MIN}}$	Reactive power generation minimum (VAR)
$\mathbf{q}$	Reactive power (VAR)
$\mathbf{R}$	Line resistance (Ω)

$\mathbf{S}$	Covariance matrix of variables
$\mathbf{U_c}$	Common cause subspace
$\mathbf{U}$	Eigenvectors of covariance matrix
$\mathbf{V^{MAX}}$	Voltage magnitude maximum (volts)
$\mathbf{V^{MIN}}$	Voltage magnitude minimum (volts)
$\mathbf{v}$	Voltage magnitude (volts)
$\mathbf{w}$	Broken stick segments
$\mathbf{X}$	Variables
$\boldsymbol{\chi}$	Line reactance (Ω per mile)
$\boldsymbol{\mu}$	Eigenvalues of the Laplacian matrix
$\boldsymbol{\psi}$	Strength of a node in a graph
$\boldsymbol{\rho}$	Effective resistance
$\boldsymbol{\Theta^{MAX}}$	Voltage angle maximum (degrees)
$\boldsymbol{\Theta^{MIN}}$	Voltage angle minimum (degrees)
$\boldsymbol{C^{G}}$	Gas constant that incorporates properties of the environment and the pipeline
ΔR_G^l	Graph resistance metric
δ	Density of gas relative to air
ϵ	Absolute rugosity (roughness) of a pipeline (mm)
γ_1	Compressor-specific constant
γ_2	Compressor-specific constant

γ_3	Compressor-specific ratio of outlet pressure to inlet pressure
ϕ	Gleason Staelin statistic
A^a	Set of natural gas active arcs
A^p	Set of natural gas passive arcs
E	Set of graph edges
J	Number of variables
k	Isentropic exponent
N	Number of observations
N^a	Number of active pipelines
N^b	Number of electrical buses
N^d	Number of natural gas demand nodes
N^e	Number of edges
N^f	Number of pipelines
N^g	Number of generators
N^l	Number of transmission lines
N^n	Number of natural gas network nodes
N^p	Number of passive pipelines
N^s	Number of natural gas supply nodes
N^t	Number of natural gas transshipment nodes
P	Size of common cause subspace

Q	Sum of squares of residual
Q_α	Detection statistic
R^G	Gas constant
R_G	Effective graph resistance
S_g	Specific gravity of gas relative to air
T	Gas temperature (K)
Z	Gas compressibility factor
$\mathbf{f^A}$	Apparent power flow(MVA)
$\mathbf{f^G}$	Gas mass flow (m^3 per hour)
$\mathbf{g}$	Real power generation injection at each bus (MW)
$\mathbf{p}$	Gas nodal pressures (bar)
$\mathbf{s}$	Gas supplies(+) and demands(-)
$\mathbf{W^{MAX}}$	Maximum energy used by a compressor station (kW)
$\mathbf{W}$	Energy used by a compressor station (kW)
$\boldsymbol{\eta}$	Thermic efficacy factor of the compressor stations
$\boldsymbol{\theta}$	Voltage angles at each bus (Degrees)
L^G	Laplacian matrix of a graph

Chapter 1

Introduction

The economic stability and general well being of the United States depend on the secure functioning of its critical infrastructure, including such diverse sectors as energy, chemical, communications, water, wastewater, and transportation. The sectors included in the critical infrastructure are all supported by information technology (IT) and industrial control systems. As the communication between IT and control systems has increased in complexity, so has the cyber risk to operations. Industry as well as government agencies have consequently focused their efforts on reducing this risk to critical infrastructure in the United States. In February, 2013, President Obama issued an executive order directing the National Institute of Standards and Technology (NIST) to work with stakeholders and develop a voluntary framework for reducing cyber risks to critical infrastructure [2]. The developed Framework is made up of standards, guidelines, and practices that promote the protection of critical infrastructure and is meant to be a model document that will be adopted by each critical infrastructure sector [3].

The Framework core consists of five functions—Identify, Protect, **Detect**, Respond, Recover. This research documents a method of detection which, grounded in multivariate process control, affords a level of protection for critical infrastructure which is nonexistent today. It focuses on a generic, data-driven algorithm that accepts trending data and uses multivariate statistical analysis to compare the current data observation to recent trending. This analysis identifies anomalous data by transforming the multidimensional observation into a single value, that is compared to a statistical threshold to determine if the observation conforms to the recent trending. The statistical threshold represents the limit of the acceptable difference between the structure of the observation and the trending data. The algorithm maintains surveillance of data in a system to ensure that data used in operation and control is secure.

1.1 Research Goal and Objectives

The goal of this research is to detect cyberattacks to critical infrastructure by creating a practical detection tool that can assist a system control center in detecting cyberintrusion. Using multivariate analysis, the tool monitors the output of a computer module to determine if the input data has been compromised. The objectives are

- Accurately discover anomalies.
- Result in a reasonable false alarm rate. A false alarm is denoted by an incorrect assertion that an anomaly has occurred when no anomaly is actually present.

1.2 Research Contributions

The main contribution of this dissertation is the development of a generic data-driven algorithm that when incorporated into programs and procedures in system control centers will enhance the security of critical systems. The algorithm applies principal component analysis (PCA) to analyze the results from a processor program and detect data that may have been compromised by cyberterrorism. PCA has successfully identified anomalous data in communication networks [4] and its use in the chemical processing [5] and manufacturing industries [6], for process control is well-documented. The algorithm applies established techniques in an innovative way in critical infrastructures.

This research defines a new class of cyberattacks to energy systems—malicious modification of network data stored in an accessible database. Data anomalies could be the result of unauthorized access to and modification of data by an intruder or by malicious code that is not quarantined by preventive software. Since modern control centers use state-of-the-art security to prevent cyberintrusion based on recommendations, regulations, and requirements from agencies like The North American Electric Reliability Corporation (NERC), NIST, and the Department of Energy (DOE) [7], operators expect the data stored in protected databases to be secure. A cybercriminal could manage to get around security measures and modify data in a database without the system operator's knowledge [8].

The techniques for anomaly detection are applicable to data in such diverse areas as electric power operation, natural gas and oil transportation, and water processing plants. In general, the research applies to critical infrastructures where input parameters are stored in a network-accessible database.

1.3 Organization of Research

Chapter 2 provides background on the problem statement with an extensive literature review of prior work on both bad (caused by equipment or human error) and malicious (intentionally modified by an intruder) data detection. A conceptional architecture and framework of the algorithm with a simple illustrative example are included in chapter 3. Principal Component Analysis, the principle tool used in the algorithm is defined and described in chapter 4. Chapter 5 explains the approach that is followed to detect and identify anomalous data. Because the algorithm can be used in multiple industries, two different industries were chosen to test the efficacy of the algorithm: the electric power transmission system and the transportation of natural gas. The two industries are similar in that the state variable data trends over time, but different in the data itself and the application processor used to calculate the state variable data. In chapter 6 the algorithm is applied to two case studies from the electric power industry and in chapter 7 the algorithm is applied to a case study from the natural gas industry. A summary of conclusions in support of the research ends the manuscript in chapter 8.

Chapter 2

Literature Review

The Internet has become a major means of communication and its vulnerabilities have been discovered by cyberterrorists. According to Industrial Control Systems Cyber Emergency Response Team (ICS-CERT) there are many different ways for intruders to gain access to data or communication systems. Fig. 2.1 documents one type of attack which could be thwarted by the algorithm developed through this research. "Nearly every production control system logs to a database on the control system LAN that is then mirrored into the business LAN. Often administrators go to great lengths to configure firewall rules, but spend no time securing the database environment. A skilled attacker can gain access to the database on the business LAN and use specially crafted SQL statements to take over the database server on the control system LAN. Nearly all modern databases allow this type of attack if not configured properly to block it," states the ICS-CERT website [9].

NERC develops and enforces standards to assure the reliability of the bulk power system in North America [10]. In addition to the reliability standards, NERC also develops and enforces Critical Infrastructure Protection Standards (CIP) which focus on cybersecurity. A detailed summary of the current version of standards is available at [11].

NIST defines cybersecurity as "the process of protecting information by preventing, detecting, and responding to attacks" [3]. NIST released Cybersecurity Framework Version 1.0 in February, 2014 [12], a document to help critical systems better protect their information and physical assets from cyberattack.

Supervisory control and data acquisition (SCADA) networks as illustrated in Fig. 2.2 are made up of computers and applications and are part of the nation's critical infrastructure. SCADA networks were originally designed to maximize functionality and as such, the security

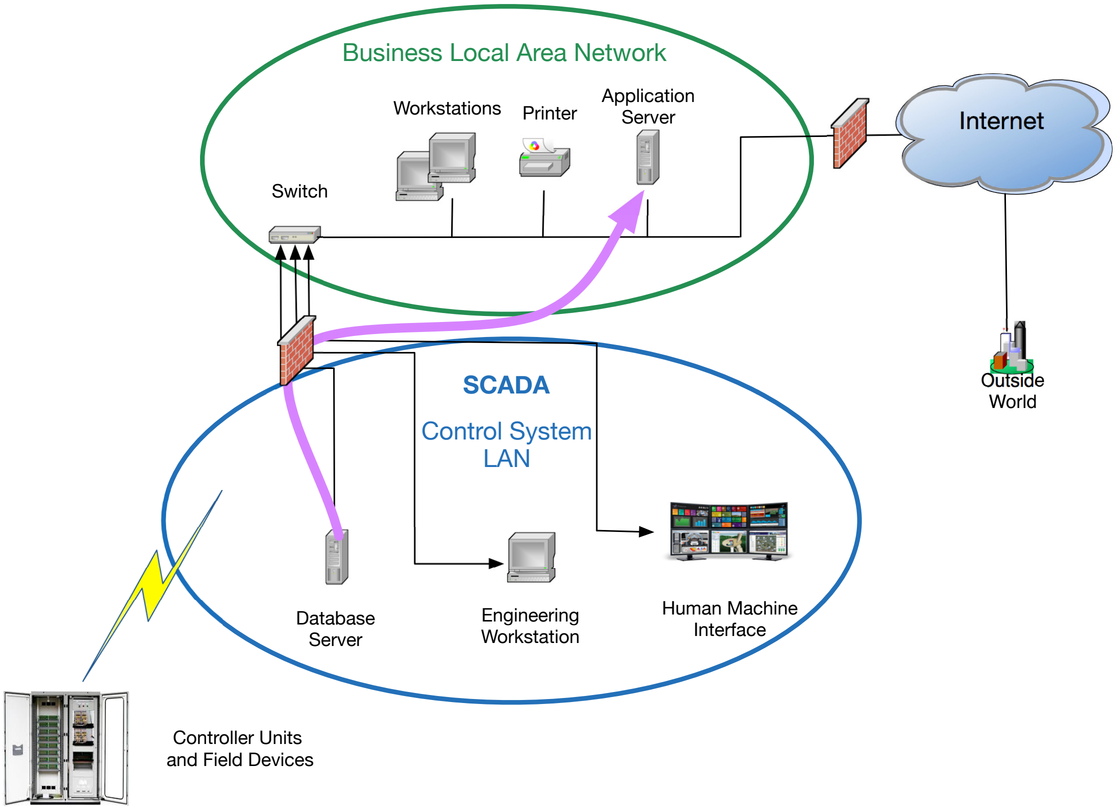


Figure 2.1: Database Links

of these systems is often weak, causing them to be vulnerable to disruption of service, process redirection or manipulation of operational data. The President's Critical Infrastructure Protection Board and the Department of Energy developed a list of 21 steps to help improve the security of SCADA networks [13]. Among the list of steps is 8–Implement internal and external intrusion detection systems and establish 24-hour incident monitoring. The algorithm developed in this research is such an intrusion detection system.

In addition to these standards, NIST also makes available guidelines on firewalls and firewall policy [14]. NIST SP 800-41, *Guidelines on Firewalls and Firewall Policy*, provides general guidance for the selection of firewalls and the firewall policies. As critical systems transition to state-of-the-art technology, upgrading hardware, software, and infrastructure increases the threat of cyberattack to these necessary resources. The Government Accountability Office (GAO) reported in 2009 that cyberthreats to critical infrastructure like the

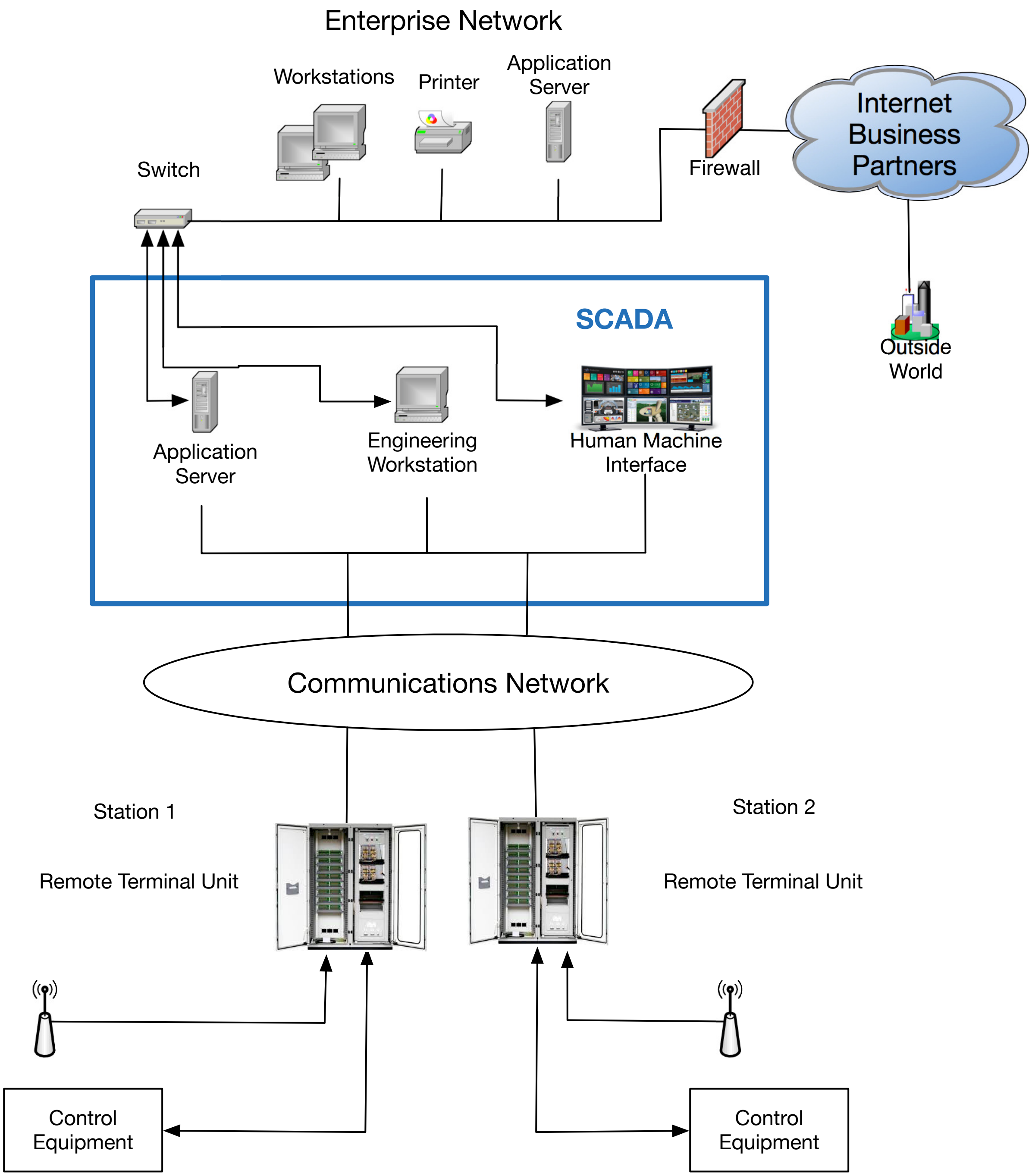


Figure 2.2: SCADA

power grid are increasing and evolving. The sources of these threats—hackers, foreign nations, disgruntled employees and terrorists—coupled with the sophistication of technology and widespread documentation of intrusion techniques on the Internet have led the government to become increasingly concerned about the potential for cyberattack. Fig. 2.3, from the United States Computer Emergency Readiness Team (US CERT), illustrates that threats categorized as unauthorized access and malicious code made up 43% of cyberthreats to federal information systems and cyberbased critical infrastructures in 2010, up from 32% in the three year timespan beginning in 2006 [15]. Unauthorized access is considered logical (or physical) access to data without permission. And, malicious code is the successful installation of a virus, worm or other code-based entity that infects a system and is not quarantined by anti-virus software.

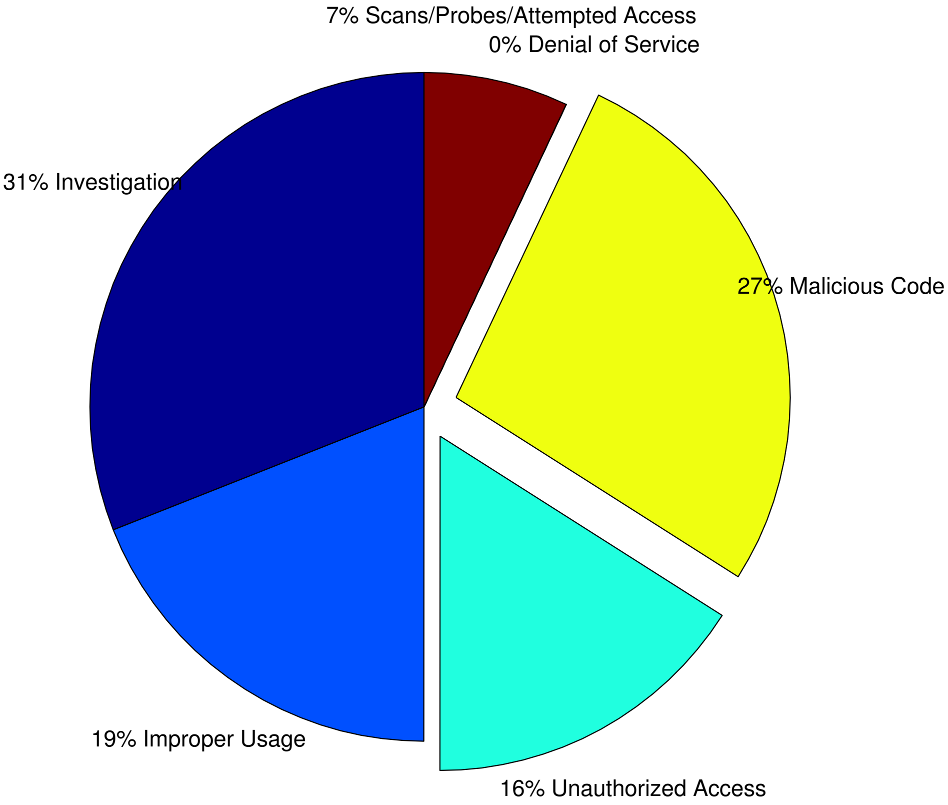


Figure 2.3: Percentage of Incidents Reported to US-CERT in 2010 by Category [1]

Whereas the threat is real, cyberattacks have not disrupted critical resource delivery in the United States [16]. However, advanced intruders can circumvent computer and network security. The Stuxnet worm is one recent example of malicious code that gained access to and damaged critical control systems in Iran's nuclear program [17].

2.1 Electric Power Transmission System

The trustworthiness of the data passed among SCADA and Energy Management System (EMS) and Business Management System (BMS) [18] program modules in the Enterprise Network is important for the proper operation of the power grid. In addition, integrated topology processing, such as in the PowerWorld software [19] provides for information exchanges across operations and planning modules. Output data from an operations module is often used as input data to a planning module and vice versa. At any point in the process of data collection and decision making data errors could lead to unnecessary and costly outages if not recognized in a timely manner. The data could be compromised not only through equipment or human errors but also by the intervention of an intruder who tampers with the data stored in a database or interferes with the module that processes real time network topology.

The optimal power flow (OPF) program (explained further in Chapter 6.2) solves a set of nonlinear equations using stored data to compute a steady state operation point with the objective function varying according to the target of the optimization. OPF is widely used in power system control and since it runs often, sometimes every 30 seconds [20], an undetected cyberattack on data supplied to the OPF program could cause power to be dispatched erroneously. Network configuration, transmission line capacity, and other transmission line parameters are some of the data input to the OPF program that could be maliciously modified by an adversary through unauthorized access or the introduction of malicious code.

2.1.1 Cybersecurity in State Estimation

Recent research in power system security has focused extensively on cyberintrusion related to measuring devices like, phasor measurement units (PMUs). These attacks are referred to as malicious data injection attacks. The research documented here is different from the research on malicious data injection attacks. However, because of the importance of state estimation to the power grid, a review of techniques used in bad data detection in state estimation is included here.

The state estimation program uses the measurements from metering devices such as PMUs to estimate state variables like voltage angles and magnitudes at each bus in a power system. Statistical techniques successfully identify and remove obvious bad data from state estimation procedures. And, since state estimation cleans the data, this process also prevents the bad data from being stored in databases for future use.

2.1.2 Bad Data Detection in State Estimation

Dealing with erroneous data has been a concern of state estimation computer programs since their application to power systems by Schweppe, *et al.* in the 1960's[21]. State estimation programs give system operators a relatively up to date picture of the power system through the use of state variable estimates and computer modeling. Since state estimation uses telemetered real time data to approximate the actual values of system variables, the validity of the process is subject to bad data caused by such things as equipment installation problems, localized equipment failures, and communication errors. These computerized static state estimation programs use statistical techniques to reduce the effects of bad data on the estimates.

Iterative and residual analysis. Weighted least squares (WLS) is a commonly used technique that iteratively solves Gauss' normal equations and has proven successful in accurately estimating state variables [21]. A bad data suppression (BDS) estimator to improve on the WLS technique by detecting and identifying bad data that previously skewed WLS results is proposed in [22]. The technique uses least normalized residuals (LNR) to detect and eliminate bad data. This use of residual analysis and non-quadratic estimation criteria led to the concept of interacting vs. non-interacting bad data and the ability to probabilistically predict false alarms [23]. This alternative also suggests that for detecting bad data local metering redundancy is more important than metering redundancy across the system. Hypothesis testing identification (HTI) has been used to eliminate problems faced by previous techniques related to conditions such as multiple and interacting bad data. Through an iterative process, HTI selects all suspect data based on normalized residuals, estimates the errors, makes a decision (valid or invalid) based on a hypothesis related to the data error, refines the hypothesis and repeats. HTI is better able to identify data errors caused by multiple and interacting bad measurements [24]. In [25] the authors borrow from decision theory and use a branch and bound algorithm to successfully identify and eliminate bad data in cases where the LNR algorithm fails. The method is successful even when multiple bad data are interacting and conforming and also eliminates the assumption that data errors will be small.

Orthogonal transformation. The Bad Data Detection, Identification and Elimination (BDDIE) method described in [26] incorporates orthogonal transformation from Golub's Method [27] and Givens' Rotation Method [28] with previously documented LNR. Combining orthogonal transformations with LNR, Vempati and Shoults [29] sequentially process measurements in order to detect multiple bad data occurrences. These orthogonal transformation techniques have not been widely accepted by the power system community, probably because of the computation effort they require when applied to large systems.

The state estimation techniques referenced above all use a single frequency, balanced, and symmetric power system model under steady state conditions. The methods have been enhanced, but in most commercial applications, they remain much the same as when first implemented in the latter half of the twentieth century. These techniques successfully identify and remove obvious bad data from state estimation procedures. Since state estimation cleans the data and only sends a portion of collected data to SCADA systems, this process also prevents the bad data from being stored in databases for future use.

2.1.3 Malicious Data Detection in State Estimation

Since the late 1990s when terrorism began moving to the forefront of American consciousness, more consideration has been placed on detecting data errors intentionally injected into the power system. These errors will likely not be detectable by commonly used methods like LNR and others referenced above. False data maliciously incorporated into databases in a power system are designed to be undetectable. An intruder could craftily formulate an attack to inject bad data in a way that would optimize damage to the power system while minimizing detection.

Liu *et al.*, in [30], described a new class of cyberattacks called false data injection attacks. Results of their research indicate that it is possible by compromising meter measurements to construct an attack vector that changes the results of state estimation and is undetectable by commonly used methods of bad data detection like LNR. Sensor measurement protection through the use of network observability rules as a solution to detecting false data injection attacks was the focus of research in [31]. To measure the vulnerability of a network, the research in [32] defined a security index as the minimum number of meters necessary to perform an unobservable attack. An algorithm for such an index that helps to locate power flows whose measurements are potentially easy to manipulate can be found in [33] where the authors urge the incremental deployment of protected measurements to increase grid security. In an expansion on the research by Liu *et al.*, Kosut, *et al.* proposed a detector based on the generalized likelihood ratio test (GLRT) to detect attacks where the adversary does not have access to a sufficient number of meters to launch an unobservable attack. They posit that the key to defending against such malicious data attacks is the introduction of redundant and trustworthy measurements that ensure network observability [34]. Understanding attacks in non-linear state estimation is the focus of [35] where a methodology is developed to simulate an attacker's use of online data during an attack to change nonlinear state estimation solutions. To defend against and detect false data-injection attacks to state estimation, Yang, *et al.*[36] identifies and protects critical sensors to make the state estimation more resilient to attack and develops spatial and temporal-based schemes to identify data-injection attacks. The financial effects of the aforementioned false data injection attacks on electric power market operations is studied in [37].

The most recent research on smart grid data integrity attacks with the goal of biasing state estimation results, focuses on strategic methods of identifying, foiling, and counteracting attacks on IEDs. In [38] unobservable, coordinated attacks are described and strategic placement of secure phasor measurement units (PMU) are shown to be an effective defense. The research by Kim, *et al.* develops an algorithm to optimize the choice of PMUs to secure and a separate algorithm to optimize their placement [39]. A risk mitigation model is proposed by [40] to respond to cyberattacks in PMU networks. The mixed integer linear programming problem prevents cyber-attack propagation while maintaining observability of the network.

2.2 Cyberattack Detection

Because of the continuing escalation in cyberattacks in every sector, research continues to expand in cyberattack detection in the energy sector. The current research on timely detection is limited with the majority of research still focused on prevention.

The results of state estimation and fault diagnosis matrices are used as input to identify nodal attacks in power system networks in [41]. Artificial neural networks are employed in [42] to monitor power flows and detect anomalies. A recent thesis on cyberattack detection in electric power distribution systems is attributed to [43].

2.3 Communication Network

Implementation of the results from the recent research is necessary to protect the power grid; however, it is not sufficient. Network data stored in databases is also vulnerable to cyberattack. These cyberattacks are different from previously researched data integrity attacks in the sense that these physical transmission line data do not depend on the measurements from devices like PMUs. If a cyberattack that changed network parameters stored in a database were to significantly alter the system state estimate, then the alarm raised to the operator during bad data detection would likely discover the parameter change. However, situations could arise where the state estimation bad data detection schemes would not detect malicious modification of network data in an accessible database.

For example inadequate measurement redundancy could cause the parameter to be undetectable [44] or the network database could be modified by cyberattackers after the most recent state estimation program has run and before an instance of the OPF module has run. The algorithm described here, implemented in the network topology processing portion of the OPF module, provides an additional security measure to protect the power grid.

In this day and time, malicious data detection in communication networks is a common problem seeking a solution and PCA is a multivariate statistical technique used successfully in this industry [4, 45].

2.4 Industrial Control System

Industrial control systems are integral to the operation of critical infrastructure systems like energy transmission networks. Cyberintrusion and its impact on industrial control systems is a growing problem and research in this area has highlighted some fundamental risks to these necessary systems. ICS-CERT recommends that a framework often referred to as "defense-in-depth" be applied for improving cybersecurity defenses. Defense-in-depth is the strategy of implementing multiple layers of defense to deal with multiple security situations such as those found in industrial control systems [46]. In Fig. 2.4 the identified risk is System Data Model Access and each layer in the figure addresses a particular part of the overall system necessary to mollify the risk. One of the defense strategies in the Network Layer is an Intrusion Detection System (circled in red in Fig. 2.4). To improve cybersecurity defenses all layers identified in the risk assessment strategy need to actively participate.

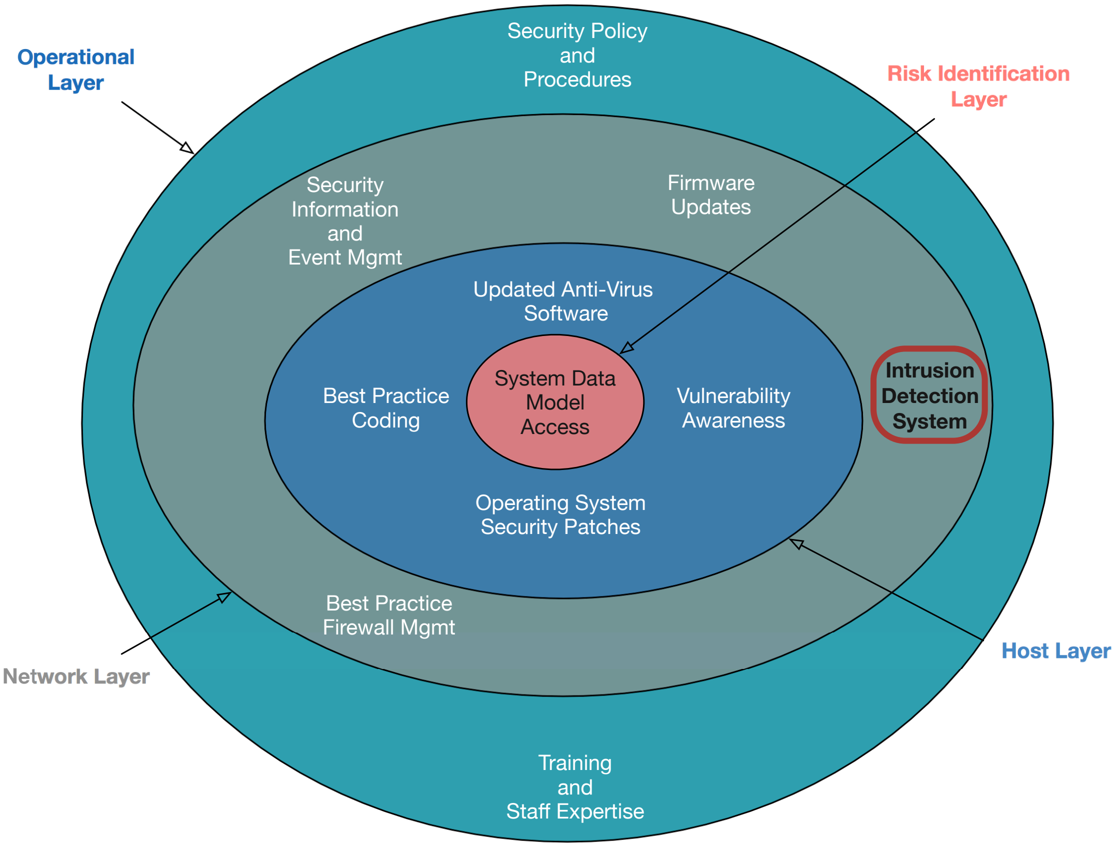

Operational Layer
Security Policy and Procedures
Risk Identification Layer
Security Information and Event Mgmt
Firmware Updates
Updated Anti-Virus Software
Best Practice Coding
System Data Model Access
Vulnerability Awareness
Intrusion Detection System
Operating System Security Patches
Best Practice Firewall Mgmt
Network Layer
Host Layer
Training and Staff Expertise

Figure 2.4: Defense-in-Depth

Cybersecurity work related to intrusion detection in industrial control systems has focussed mainly on identifying signatures specific to control systems that can be used to monitor for attacks. See [47] for an example of a commercially available set of SCADA intrusion detection system signatures. This type of intrusion detection system falls apart before a new signature is identified since it cannot be detected until it is identified.

Cardenas, *et al.* [48] proposed an automatic detection and response module based on estimates of the state of the system to detect computer attacks that change the behavior of the targeted control system.

A replay attack is a series of commands to a control system that is copied and replayed to the system. Since the commands are identical, but replayed at a different point in time, it is difficult to detect the incorrect data. Mo, *et al.* [49] used watermarked input and an optimal Neyman-Pearson (using a likelihood ratio test) detector to determine if a system is under attack.

2.5 Natural Gas Transmission System

In May 2012 the Christian Science Monitor [50] reported,"A major cyberattack is currently under way aimed squarely at computer networks belonging to US natural gas pipeline companies, according to alerts issued by the US Department of Homeland Security (DHS)." Statements such as this have been reported over the years, but neither pipeline operations nor their industrial control systems have yet been affected. In its quarterly report, ICS-CERT [51] reported in April 2014 that Internet accessible control systems are at risk and documented instances of a public utility being compromised.

Working with the Transportation Security Administration (TSA) and Department of Homeland Security (DHS), the pipeline industry established the Industrial Control Systems Joint Work Group (ICS_JWG) in 2012. "The ICS_JWG enhances the collaborative efforts of the industrial control systems stakeholder community in securing critical infrastructure by accelerating the design, development, and deployment of secure industrial control systems. Cybersecurity is a particular focus for this group," according to the Interstate Natural Gas Association of America (INGAA) [52]. The INGAA believes that public-private partnerships between the gas industry and law enforcement are "the most effective means for securing the nation's critical infrastructure and addressing any cyberthreats." At this point in time, educating their employees about the importance of implementing best practices for improving security at critical pipeline facilities across the country is the primary cyber focus of the pipeline industry.

Since much of the interstate flow of natural gas is managed through SCADA systems, any effort to improve security for automated control systems improves security for natural gas transmission.

2.6 Standard Protection Practice

Standard protection practices in the U.S. critical infrastructure focus on *preventing* the cybercriminal from accessing and altering network data using such devices as known signature intrusion detection systems, virus protection software and encryption technologies. In electric power transmission, the state estimation programs use telemetered real world measurements to forecast changes in demand which require changes in electric generation. Though the state estimation could identify cyberactivity, in situations where network observability is incomplete, cyberactivity could go unnoticed without an algorithm designed to find it. The anomaly detection tool described below is such an algorithm.

Chapter 3

Conceptual Architecture

Conceptually, the anomaly detection tool works as protection for the results from application software. The tool analyzes the output from the application software and alerts the controller of anomalous data before damage is incurred. Fig. 3.1 illustrates the fact that processing occurs behind the firewall and with intrusion prevention in place.

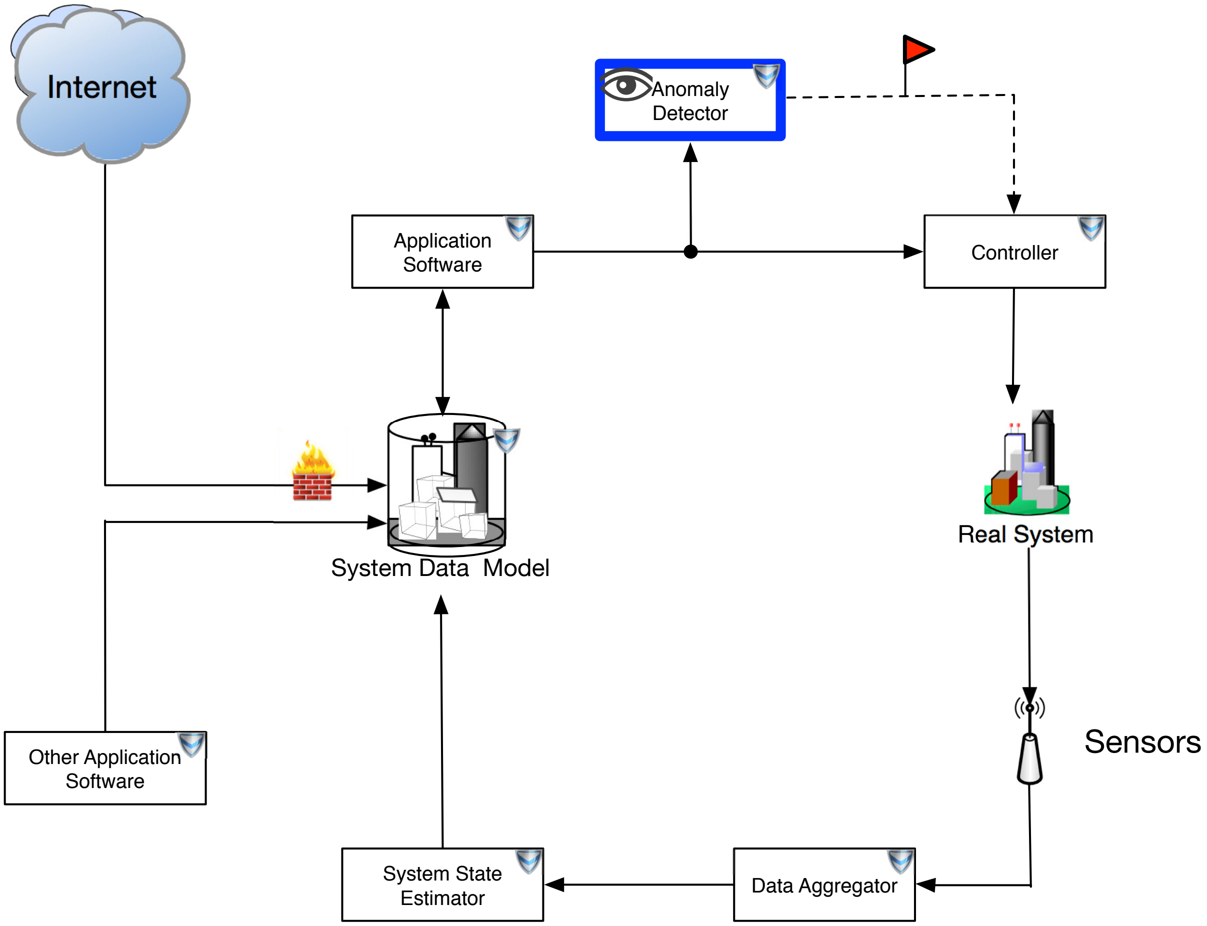


Figure 3.1: Conceptual Architecture

Even with an active firewall and intrusion prevention in place, a cyberevent (unauthorized access to the protected system) could alter the system data model, a digital representation of the real system that is used by application software to facilitate decisions related to controlling the real system. The altered system data model would no longer match the real system and without an anomaly detection tool in place, both operator decisions and automatic control decisions would be made using incorrect information.

An example from the energy sector is used below to illustrate the conceptual architecture. An explanation of the real transmission system modeled in the example is given in section 3.1. The system data model as it is stored in the database is described in section 3.2. The DCOPF application software and its input and output are related in section 3.3. And the attack and its effect are elaborated on in the final section 3.4.

3.1 Real System

An electric power transmission system (denoted by green lines in Fig. 6.1) is designed to transport high-voltage electrical energy from power generating plants to electrical substations which are typically located near customer demands. This chapter references three major parts of the electric power transmission system—buses, lines, and generators. The bus is the common connection point for the lines, demand sources, and generation sources. The lines are the structures through which the electrical energy is transported between buses. And, the generators are the machines in a power plant that generate the electrical energy.

The 6-Bus system, used as an example, is a small transmission network of 6 buses, 11 lines, and 3 generators.

3.2 System Data Model

The system data is stored in a database and protected by an active firewall as well as other security measures like passwords and updated virus protection software. The data model for the example is given in Tables 3.1 and 3.2. Additional detail about the 6-Bus system data model can be found in [53]. If a cyberterrorist were to subvert preventive measures, gain access to, and modify the power system data model, stored in a protected network database, then the model would no longer accurately represent the real power system. Decisions and actions based on the altered model could result in damage to the energy system if left undetected.

Table 3.1: Bus Descriptions

Bus	**Demand (MWH)**	**G$^{\mathbf{MAX}}$(MW)**	**G$^{\mathbf{MIN}}$(MW)**	**Cost ($/MWH)**
1	0	200	0	11.669
2	0	150	0	10.333
3	0	180	0	10.833
4	80	NA	NA	NA
5	80	NA	NA	NA
6	80	NA	NA	NA

Table 3.2: Line Descriptions

Line	**From Bus**	**To Bus**	**Reactance (p.u.)**	**Capacity (MW)**
1	1	2	0.2	40
2	1	4	0.2	60
1	3	5	0.3	40
4	2	3	0.25	40
5	2	4	0.10	60
6	2	5	0.30	30
7	2	6	0.20	90
8	3	5	0.26	70
9	3	6	0.10	80
10	4	5	0.40	20
11	5	6	0.30	40

In this example, it is assumed that the attacker changed line 1 to transmit between buses 2 and 3 (from between buses 1 and 2) and changed the destination of line 5 from bus 4 to bus

Table 3.3: Atacked Line Descriptions

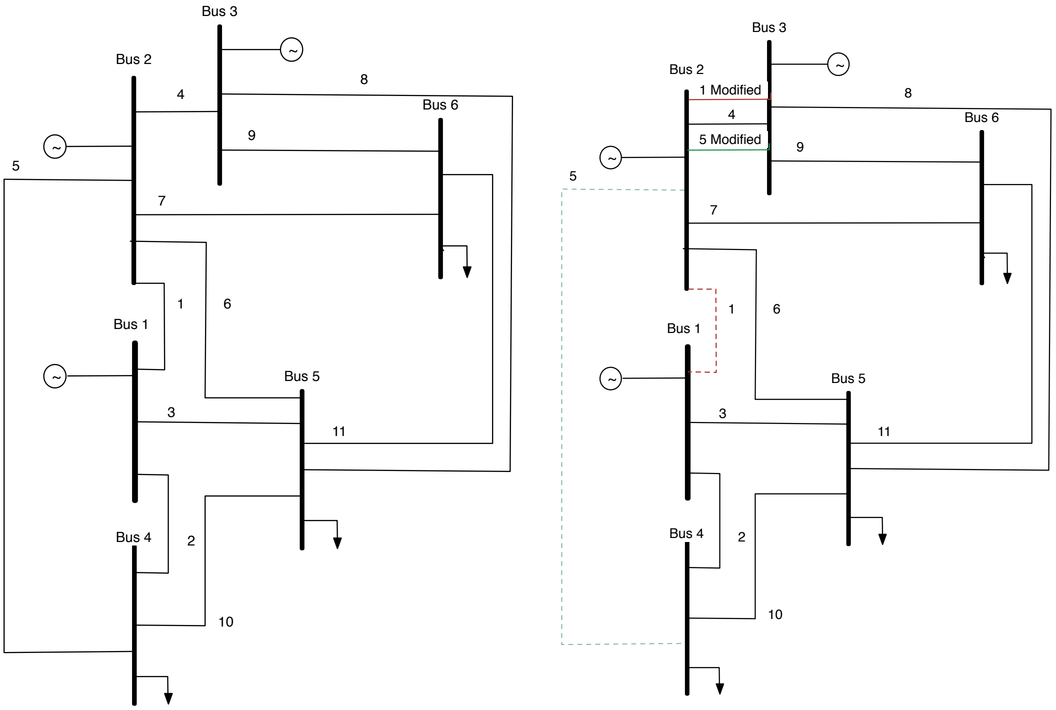
Line	From Bus	To Bus	Reactance (p.u.)	Capacity (MW)
1	**2**	**3**	0.2	40
2	1	4	0.2	60
3	1	5	0.3	40
4	2	3	0.25	40
5	2	**3**	0.10	60
6	2	5	0.30	**60**
7	2	6	0.20	90
8	3	5	0.26	70
9	3	6	0.10	80
10	4	5	0.40	**40**
11	5	6	0.30	40

3. In addition to the from and to bus changes, the cyberattacker also raised the line capacity where he expected the line overuse to occur (lines 6 and 10) thus avoiding technical program problems in the application software. Line overloads are typically flagged by application software because of line capacity constraints. The demand and generation amounts and all other system parameters remain the same. Table 3.3 gives the line descriptions in the database after the attack.

Fig. 3.2 shows both the model that accurately represents the real system and the model as it was modified by the cyberevent.

3.3 Application Software

The application software runs an optimization module called DCOPF. When used for power dispatch, DCOPF minimizes the cost of generation and is constrained by demand and generation balance at each bus, upper and lower limits on the output of generating units, and line capacity. The decision variables are voltage angles, $\boldsymbol{\theta}$, and injected generation, $\mathbf{g}$, at each bus. The power flow through a line is related to the voltage angles at the from and to



(a) Accurate Model (b) Attacked Model

Figure 3.2: 6-Bus System Model

buses, θ_i, θ_j, and the reactance, χ_{ij} of the transmission line between the two buses according to equation (3.2)[1]

$$f_{ij}^{P} = \frac{\theta_i - \theta_j}{\chi_{ij}} \tag{3.2}$$

Under the described attack scenario, the system operator would run the application software, that in this case is the DCOPF program, to determine the power dispatch, unaware of the attack. The mathematical model in equations (3.3), (3.4), (3.5), (3.6), is used by the DCOPF program to determine the economic dispatch of the 6-Bus system. The program is run periodically, often every 30 seconds, and is consequently a crucial tool for the power system operator.

[1] For power flow calculations using the data from the cases in Matlab, the voltage angles must be converted from degrees to radians and multiplied by the base MVA.

$$f_{ij}^{P} = \frac{1}{\chi_{ij}} \times \frac{\pi}{180} \times (\theta_i - \theta_j) \times 100 \tag{3.1}$$

$$\underset{g,\theta}{\text{minimize}} \sum_{i=1}^{6} C_i g_i \tag{3.3}$$

Subject to

$$\sum_{i=1}^{6} H_{ij} \frac{(\theta_i - \theta_j)}{\chi_{ij}} + g_i - d_i = 0 \qquad \text{for } j = 1 \ldots 6 \tag{3.4}$$

$$\left| \frac{(\theta_i - \theta_j)}{\chi_{ij}} \right| \leq F_{ij}^{\text{MAX}} \qquad \text{for } i = 1 \ldots 6, \quad \text{for } j = 1 \ldots 6 \tag{3.5}$$

$$G_i^{\text{MIN}} \leq g_i \leq G_i^{\text{MAX}} \qquad \text{for } i = 1 \ldots 6 \tag{3.6}$$

In equation (3.3) C_i is a parameter that represents the cost of generation at the ith generator and g_i is a decision variable—the amount of generation in KW generated from the ith generator.

Equation (3.4) is the set of power balance constraints. Each power balance equation constrains the amount of electrical energy in and out of bus j so that it is balanced. H_{ij} is a parameter that represents the adjacency matrix, indicating the relationships among the buses and the lines; θ_i is a decision variable—the voltage angle at bus i; χ_{ij} is the reactance parameter for the line from bus i to bus j; g_i is the decision variable described above; and d_i is a variable representing the amount of energy demanded by customers at bus i. The power balance equation ensures that the sum of electrical energy in and electrical energy out of bus j is equal to zero. Note that there is one equation (3.4) for each bus.

Equation (3.5) is the set of variable constraints. F_{ij}^{MAX} is the parameter that defines the maximum power flow that can safely and reliably be accommodated by the line from bus i to bus j. Equation (3.2) defines the power flow through the line from bus i to bus j. There is one of equation (3.5) for each transmission line.

Equation (3.6) is the set of generator capacity constraints. The generation at bus i is constrained by the parameters G_i^{MIN} and G_i^{MAX} based on the size of the generator unit and other environmental and economic considerations. There is one equation (3.6) for each bus at which there is generation.

Note that in equations (3.3), (3.4), (3.5), and (3.6) all upper case letters represent parameters stored in the database.

3.4 Effect of the Attack

If the operator is unaware of the attack, the application software, DCOPF, will be run with the faulty parameters from the database in Table 3.3 and find values for $\boldsymbol{\theta}$ and $\mathbf{g}$. Table 3.4 gives the result of the DCOPF after the attack. The operator will implement the faulty results. System results in the *real* system and the flows on the *real* transmission lines that were calculated using equation (3.2), are given in Table 3.5.

Table 3.4: DCOPF Result after Attack.

Bus	$\boldsymbol{\theta}$ (Degrees)	$\mathbf{g}$ (MW)
1	0	78.4
2	2.53	37.4
3	3.14	124.4
4	-7.04	0
5	-2.85	0
6	-0.57	0

In Table 3.5 it is clear that the cyberevent has actually resulted in line 3 being overloaded and lines 1 and 2 seriously close to being overloaded.

This simple example highlights the importance of anomaly detection software. If the attack were committed by a sophisticated team of programmers and engineers, serious damage could be inflicted on the power transmission network. The anomaly detection software

Table 3.5: Real System Results after the Attack

	From	**To**	**MVA**	**Power**	
Line	**Bus**	**Bus**	**Rating**	**Flow**	**% Used**
1	1	2	40	39.80	99.5
2	1	4	60	55.31	92.2
3	1	5	40	45.29	113.2
4	2	3	40	6.09	15.2
5	2	4	60	33.20	55.3
6	2	5	30	17.29	57.6
7	2	6	90	31.49	35.0
8	3	5	70	18.75	26.8
9	3	6	80	47.28	59.1
10	4	5	20	5.29	26.5
11	5	6	40	3.18	8.0

described below, grounded in multivariate statistical analysis, is an important addition to standard preventive protection practices.

Chapter 4

Principal Component Analysis

Many of the networks in the U.S. critical infrastructure are made up of thousands of lines and connection points. Measurements along the lines and at the connection points are important to the economic and reliable delivery of product. Analyzing individual measurements will reveal some information, but often there is a need to understand the underlying structure of the network. Multivariate analysis tools are needed to discover the complex and possibly hidden relationships among the data. Principal Component Analysis is such a tool and is used for this analysis.

PCA is a powerful multivariate analysis technique used extensively in the social sciences to reduce the dimensionality of data and more recently in such fields as facial recognition and image compression. It was first documented by Pearson in 1901 [54] and developed independently by Hotelling in 1933 [55]. Whereas its use as a dimension reduction tool is well documented, PCA is also one of the best-known statistical methods for identifying anomalies in communication network traffic [56]. PCA transforms a (typically) large set of correlated variables into a set of uncorrelated variables on a new coordinate system. These uncorrelated variables, called principal components, are ordered so that the first few principal components contain most of the variability in all of the original set of variables while maintaining as much information as possible from the original data.

The method of principal components is derived from a key concept in matrix algebra: A $p \times p$ symmetric, nonsingular matrix, such as the covariance matrix,

$$\mathbf{S} = \frac{\mathbf{X}^T\mathbf{X}}{(N-1)} \tag{4.1}$$

may be reduced to a diagonal matrix **L** by premultiplying and post multiplying it by a particular orthonormal matrix **U** such that

$$\mathbf{U^T S U} = \mathbf{L}.$$

4.1 Definition of Principal Components

Principal components (PCs) are the orthogonal axes formed when PCA is applied to an $N \times J$ matrix of data, $\mathbf{X}$. Each of the principal components points in the direction of maximum variance remaining in the data, given the variance already accounted for by the previous principal components. The method of principal components produces the orthogonal regression line that minimizes the deviations perpendicular to the line itself. This orthogonal regression line is the first principal component.

PCA is an iterative process that in step 1 looks for a linear function of the elements of the matrix $\mathbf{X}$,

$$\mathbf{u}_1^T\mathbf{x} = u_{11}\mathbf{x}_1 + u_{12}\mathbf{x}_2 + \ldots + u_{1J}\mathbf{x}_J = \sum_{i=1}^{J} u_{1i}\mathbf{x}_i \tag{4.2}$$

where $\mathbf{u}_1^T\mathbf{x}$ accounts for the maximum variance among the J variables and is called the first principal component. In the next iteration, $\mathbf{u}_2^T\mathbf{x} = \sum_{i=1}^{J} u_{2i}\mathbf{x}_i$ is orthogonal to $\mathbf{u}_1^T\mathbf{x}$, accounts for the next highest variance and is called the second principal component, etc. until the Jth linear function, $\mathbf{u}_J^T\mathbf{x} = \sum_{i=1}^{J} u_{Ji}\mathbf{x}_i$, which is orthogonal to $\mathbf{u}_1^T\mathbf{x}$, $\mathbf{u}_2^T\mathbf{x}$, ..., $\mathbf{u}_{J-1}^T\mathbf{x}$ and accounts for the least amount of variance. The J principal components are uncorrelated and mutually orthogonal.

U is the $J \times J$ orthogonal basis set of principal component coefficients. Each column of **U** is an eigenvector of the covariance matrix, **S**, corresponding to its jth largest eigenvalue L_j

and contains the coefficients for one principal component. Since $\mathbf{u}_j^T\mathbf{u}_j = 1$ for each column of $\mathbf{U}$, L_j is the variance of the jth column of $\mathbf{S}$.

The coefficients, u_{ij}, are referred to as loadings since they are the weights by which the original data elements are multiplied when calculating the principal components. Each $u_{ij}\mathbf{x}_i$ in equation (4.2) is a score and a principal component is the sum of all of the scores for one column of loadings [57].

4.2 Method of PCA

Whereas one of the valuable traits of PCA is that it transforms correlated variables into uncorrelated variables, another important result of PCA is its ability to adequately represent multivariate data in a much smaller dimension. In multivariate analysis of critical infrastructure systems like energy transmission systems, this result is crucial since the number of transmission lines or pipelines in a single analysis can be in the thousands.

Prior to beginning the method of principal components, the Gleason and Staelin statistic for covariance matrices,

$$\phi = \sqrt{\frac{\|\mathbf{S}\|^2 - \sum_{i=1}^{p}(\sigma_i^2)^2}{\sum_{i=1}^{p}\sum_{j\neq i}^{p}(\sigma_i\sigma_j)^2}} \tag{4.3}$$

provides an indication of whether or not PCA is worth the effort by measuring the average level of correlation among the variables. A Gleason and Staelin statistic near the perfect correlation of all of the variables, $\phi = 1$, is an indication to go ahead with PCA [58, 59].

PCA can be run on either the correlation matrix or the covariance matrix of the mean-centered data. The correlation matrix is typically used when the variables in question are measured in different units whereas the covariance matrix is used when all of the variables are measured in the same units and are similar in size. For instance, if the researcher were analyzing individual information which included physical size measurements of extremities (in meters), Intelligence Quotients (IQs) and the measured force exerted by the extremities (in PSI), the correlation matrix would likely yield a better result with PCA. Since each set of data analyzed in this paper will be measured in the same units, the covariance matrix will be used for analysis.

4.3 Analysis Using Subspaces

Reducing the dimensionality of a large dataset through PCA concentrates the variability associated with the original data into a relatively small number of principal components. The decision as to how many principal components should be retained in the analysis to most accurately represent the original data has been researched extensively [60, 61] and varies depending on the criterion (stopping rule) used. The stopping rule determines the number of PCs used for analysis. The subspace defined by the stopping rule should be a fitting model for the original data with the majority of the information (variation) included.

Most of the stopping rules are either subjective or either over estimate or under estimate the data's dimension [60] and different analyses result depending on the number of PCA components retained. The list of criteria used in the literature is long, but a representative sample of criteria is described below:

Eigenvalue-one criterion. Also known as the Kaiser or Kaiser-Guttman criterion, this criterion retains components with eigenvalues greater than the average eigenvalue. Any component contributing more than the average accounts for more than the amount contributed by one variable. It is therefore worthy of being retained [62]. Components contributing less than the average are contributing less than the value contributed by one variable and are therefore considered trivial.

This criterion has been shown to be effective when a small (less than 30) number of variables are being analyzed and the variable commonallities[1] are high (greater than .70) or when the number of observations is high (greater than 250) and the mean commonality is greater than or equal to .60 [63].

[1] Variable commonality refers to the variance shared by two variables in a multiple regression.

Whereas this criterion is simple to administer, interpretation is difficult when the difference between the eigenvalues of two components is slight but one is greater than the average and one is less than the average. This criterion can be used with greater confidence when the difference between the last component to be retained and the first to be dismissed is significant.

Kaiser-Guttman with modification by Jolllife. Joliffe [57] suggested incorporating the effect of sample variance by retaining those eigenvalues whose values exceed the average eigenvalue multiplied by 0.7. This procedure is recommended for covariance matrices.

Cattell's Scree test [64, 65]. To use the scree test, a plot of the eigenvalues associated with each component is generated. The analysis looks for a break between components with relatively large eigenvalues and relatively small eigenvalues, often referred to as the elbow. Components with large eigenvalues are retained. If several breaks occur, the decision as to which components to retain is made after the first break [59]. In Cattell's original article [64], he recommended including the components prior to and including the break; however in subsequent research [65], his recommendation included the first one after the break. Both of these techniques are used in practice.

The scree test appears to provide accurate results as long as the number of observations is greater than 200 and most of the variable commonalities are large [63]. Ambiguity in finding a significant break is the most obvious deterrent to using this criterion as a standalone measure of which components to retain.

Log-eigenvalue diagram. For each eigenvalue, graph log (L_j) against j and look for the point at which the eigenvalues decay linearly. This technique was developed to help with the situation where it is difficult to identify the elbow using the eigenvalues alone.

Variance accounted for. Another criterion involves retaining a component if it accounts for a specified percentage of total variance in the dataset, often 10%. Or, alternatively, retaining all components accounting for a total cumulative percent of variance, usually greater than 70% or 80%. Both of these criteria are used extensively by researchers today, but the subjectivity of the criteria has been criticized.

Jackson suggests a better method of accounting for variance is to specify a total sum of variability to keep in the common subspace instead of a proportion [59].

Broken stick. The broken stick model [57] randomly divides a "stick" of unit length into n segments with the expected value of the kth longest segment defined by equation (4.4). The distribution defined by the broken stick model was identified by MacArthur [66] in his study of the distribution of bird species and first applied to PCA for dimension reduction by Frontier [67].

$$w_k = \frac{1}{N^n} \sum_{i=k}^{N^n} \left(\frac{1}{i} \right) \tag{4.4}$$

The kth PC is included in the model if the proportion explained by that PC is larger than the corresponding w_k. This model retains those PCs for which the amount of variance accounted for by the PC is more than would be expected by chance alone.

In their study of cDNA microarrays, Cangelosi and Gorily [61] recommend that the analyst should look for a "consensus dimension" by using the modified broken stick model, Velicer's MAP, Jolliffe's modification of Kaiser-Guttman, the LEV diagram, parallel analysis, the scree test, and consider the actual information dimension as the upper bound.

In the extensive literature on this topic, the decision as to which stopping rule to use varies according to the reason for using PCA, whether the analysis is based on the correlation or covariance matrix, and the industry using PCA.

For detecting anomalous data in a large complex system, Cattell's Scree test suggests a reasonable number of PCs to keep. Experience with the specific data in question provides the user with knowledge to raise or lower the number.

Chapter 5

Anomaly Detection

"In data mining, anomaly detection is the identification of items, events or observations which do not conform to an expected pattern or other items in a dataset" according to Chandola, *et al.* [68]. Anomaly detection has been studied in the statistics community as far back as 1887 when Edgeworth published an article [69] about "discordant observations." The subsequent study of anomaly detection has produced a large number of techniques in many different research areas including medicine [70], fraud detection [71] and manufacturing [72]. A detailed survey of anomaly detection methods by Chandola, *et al.* is available in [68] for the interested reader.

The survey [68] refers to three different types of anomalies:

- Point Anomalies: A data instance is anomalous with respect to the data.
- Contextual Anomalies: A data instance is anomalous within an understood set of conditions [73].
- Collective Anomalies: A collection of related data instances is anomalous. The individual data instances within a collective anomaly are not anomalous by themselves.

The anomalies that are targeted by the developed algorithm are contextual anomalies and require an understanding of the underlying structure of the multivariate data to be discovered.

5.1 PCA as an Anomaly Detection Tool

The use of PCA in cyberintrusion detection is a relatively recent use of the tool, but was preceded by similar uses in multivariate quality control [74, 75]. Because of the complexity of multivariate data matrices, PCA came into its own for applications like quality control, face recognition and image processing as computers improved in their ability to process a large amount of complicated data in a relatively short amount of time.

5.2 Subspace Method of Detection

The algorithm uses PCA to compute an orthogonal linear transformation of data and splits the resulting space into two subspaces, separating the data's variability into common cause and assignable cause variability. Common cause variability is described as naturally occurring and inherent to the process whereas assignable cause variability is unnatural and due to a shock or disruption to the process. Assignable cause variability cannot be accounted for by naturally occurring events and its source needs to be identified and corrected [76]. An anomaly in an observation is detected if the summed squares of the residual exceeds a statistical threshold. The subspace method used for anomaly detection requires the use of a parameter for the size of the common cause subspace, P, and the threshold or detection statistic, Q_α. Both have been discussed extensively in the literature [57, 59]. The size of the common cause subspace is determined by the number of principal components that are necessary and sufficient to model the data while retaining as much information (variance) as possible. For each new observation the preceding data is analyzed by PCA and split into the two subspaces described above. If data in the new observation can be accurately modeled by the two subspace model, then the residual will be minimal.

In recent research regarding anomaly detection in communication networks [4], where volume anomalies associated with a sudden positive or negative change in packet flows are analyzed, P is determined by the "Proportion of variance accounted for" criteria as described in section 4.3. Subsequent research in this area follows Lakhina *et al.* [45, 77].

5.3 Detection Model

The original variables may be stated as a function of the principal components as in (5.1)

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{U}\mathbf{z}, \tag{5.1}$$

where $\mathbf{z} = \mathbf{U}^T[\mathbf{x} - \bar{\mathbf{x}}]$. This equality only holds true when all of the principal components are used. In this analysis, where fewer than all of the PCs are used, the estimate for $\mathbf{x}$, $\hat{\mathbf{x}} = \bar{\mathbf{x}} + \mathbf{U}_c\mathbf{z}$, is used. $\mathbf{U}_c = \mathbf{U}(:, 1 : P)$ is the common cause subspace where most of the variance in all of the data remains. Each new observation is modeled according to equation 5.2.

$$\mathbf{x}^{\mathbf{NEW}} = \bar{\mathbf{x}} + \mathbf{U_c}\mathbf{z} + (\mathbf{x}^{\mathbf{NEW}} - \hat{\mathbf{x}}) \tag{5.2}$$

where $\bar{\mathbf{x}}$ represents the contribution of the multivariate mean, $\mathbf{U}_c\mathbf{z}$ represents the contribution of the PCs and $\mathbf{x}^{\mathbf{NEW}} - \hat{\mathbf{x}}$ is the amount unexplained by the subspace model, the residual.

5.4 Q_α statistic

The Q_α statistic, as defined by Jackson and Mudholkar [78] is used as a threshold for anomaly detection in [4] among others. Jackson and Mudholkar show in the appendix of [78] that

$$c = \frac{\kappa_1[(Q/\kappa_1)^{h_0} - 1 - \kappa_2 h_0(h_0 - 1)/\kappa_1^2]}{\sqrt{2\kappa_2 h_0^2}}$$

is approximately normally distributed with zero mean and unit variance so that Q_α is defined as:

$$Q_\alpha = \kappa_1 \left[\frac{c_\alpha \sqrt{2\kappa_2 h_0^2}}{\kappa_1} + 1 + \frac{\kappa_2 h_0(h_0 - 1)}{\kappa_1^2} \right]^{\frac{1}{h_0}} \tag{5.3}$$

where

$$\kappa_i = \sum_{j=P+1}^{J} L_j^i \quad \text{for } i = 1, 2, 3$$

L_j is the variance accounted for by the jth PC and

$$h_0 = 1 - (2\kappa_1\kappa_3)/(3\kappa_2^2).$$

Q_α is the control limit for the residual where c is the normal deviate corresponding to the upper $(1-\alpha)$ percentile when $h_0 \geq 0$ and lower (α) percentile when $h_0 < 0$.

5.5 Testing the Residual

The residual term from the current observation, equation (5.2), is tested by summing the squares of the residuals as in equation (5.4).

$$Q = (\mathbf{x}^{\mathbf{NEW}} - \hat{\mathbf{x}})'(\mathbf{x}^{\mathbf{NEW}} - \hat{\mathbf{x}}) \tag{5.4}$$

Q represents the sum of squares of the distance of $\mathbf{x}^{\mathbf{NEW}} - \hat{\mathbf{x}}$ from the common cause subspace that the P-dimensional model defines. If $Q > Q_\alpha$ from equations (5.4) and (5.3) then an anomaly exists in the current observation and the operator is alerted.

5.6 Computer System Specifications.

The generic algorithm is written using Matlab R2014b [79] for general programming and MATPOWER Version 5.1 [80] for procedures related to electric power transmission. Experiments were run on a 1.8 Ghz Intel Core 17 MacBook Air with 4GB ram.

5.7 Example of the Anomaly Detection Tool

In this step-by-step illustration of the use of PCA to detect anomalous data, the application processor has produced state variable data to be used in decision making. The anomaly detection tool analyzes the data and sends an alarm to operators if an anomaly is detected. The anomaly detection process is simulated as closely as possible to that which would be implemented in the real world where the algorithm would run periodically.

In this example, a matrix $\mathbf{X}$ of 45 observations on 6 variables was generated from the online Pennsylvania, Jersey, Maryland (PJM) database [81] using Monte Carlo simulation and is shown in Fig. 5.1. The elements of $\mathbf{X}$ represent data collected under normal operating conditions. Notice that the context of the data over time could have an impact on the discovery of anomalous data and must be considered in the analysis.

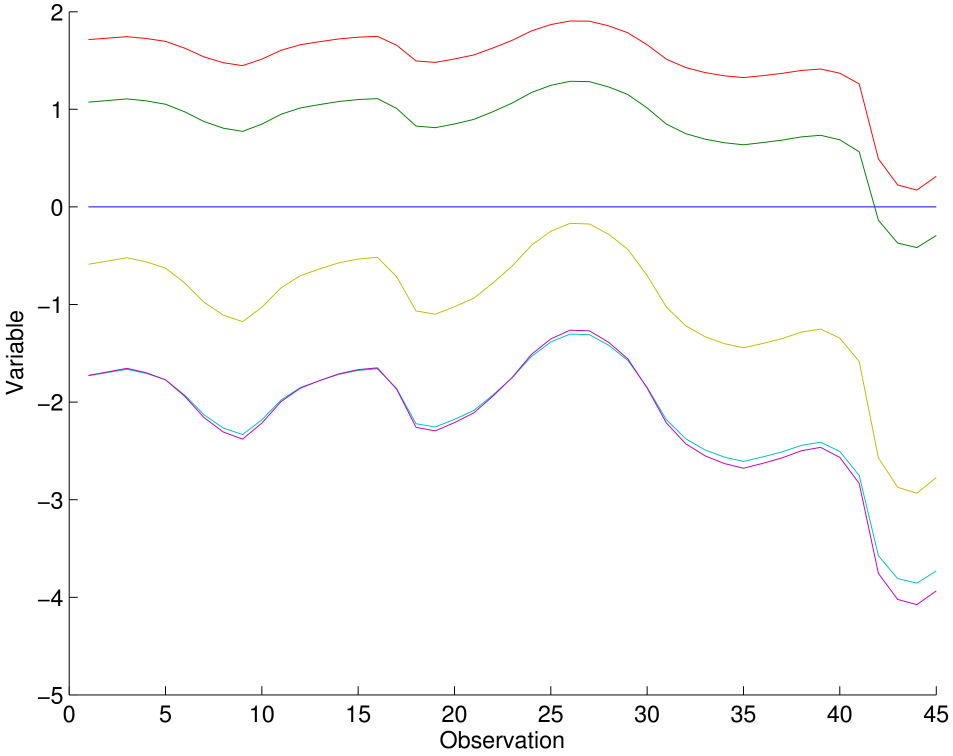


Figure 5.1: Observations

Step1: Observe the data. The column means, $\bar{\mathbf{x}}$ and the sample covariance matrix, $\mathbf{S}$ as in equation 5.5 are given in Table 5.1 and examined for a basic overview of the data.

$$\mathbf{S} = \frac{\sum_{i=1}^{6}(\mathbf{x_i} - \bar{\mathbf{x}})^T(\mathbf{x_i} - \bar{\mathbf{x}})}{(45-1)} \tag{5.5}$$

Table 5.1: Covariance Matrix, **S** with $\bar{x}$

	1	**2**	**3**	**4**	**5**	**6**
1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.0000	0.1640	0.1657	0.2506	0.2758	0.2694
3	0.0000	0.1657	0.1680	0.2506	0.2759	0.2704
4	0.0000	0.2506	0.2506	0.3934	0.4322	0.4186
5	0.0000	0.2758	0.2759	0.4322	0.4747	0.4601
6	0.0000	0.2694	0.2704	0.4186	0.4601	0.4470
$\bar{x}$	0.0000	0.8272	1.4788	-2.1574	-2.1927	-1.0251

Before continuing with PCA, the Gleason-Staelin statistic is calculated from equation (4.3). A Gleason and Staelin statistic near the perfect correlation of all of the variables, $\phi = 1$, is an indication to go ahead with PCA [58]. For this system $\phi = 1.0$, indicating that the variables are 100% positively correlated and PCA is warranted as a multivariate analysis technique.

Step2: Run PCA. PCA is run on the covariance matrix of mean-centered data, **S**. The resulting **u**-vectors (characteristic vectors or eigenvectors of S) and variances (characteristic roots or eigenvalues of S) are given in Table 5.2. **u**-vectors are orthonormal (orthogonal with unit length) and scaled to unity so that

$$\mathbf{u_i'} \times \mathbf{u_i} = 1 \quad \forall i \quad \text{and} \quad \mathbf{u_i'} \times \mathbf{u_j} = 0 \quad \forall i, j.$$

Table 5.2: U vectors and variances

Variable	$\mathbf{u}_1$	$\mathbf{u}_2$	$\mathbf{u}_3$	$\mathbf{u}_4$	$\mathbf{u}_5$	$\mathbf{u}_6$
1	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
2	0.3147	0.4514	0.8084	-0.0777	-0.1942	0.0000
3	0.3158	0.7196	-0.4707	-0.2402	0.3212	0.0000
4	0.4892	-0.3907	0.2129	0.0561	0.7481	0.0000
5	0.5377	-0.3539	-0.1754	-0.6000	-0.4415	0.0000
6	0.5223	0.0231	-0.2212	0.7571	-0.3234	0.0000
Eigenvalue	1.6383	0.0088	1.53e-18	2.85e-22	3.37e-31	0
% Explained	99	100	100	100	100	100

Note that the sum of the variances of the original variables from the diagonal of the covariance matrix, **S**, in Table 5.1 is 1.6471 which is equal to the sum of the variances of the PCs in Table 5.3.

In $\mathbf{u_1}$ all of the coefficients are positive and except for the first one, close to the same size. The first characteristic vector, $\mathbf{u_1}$, represents the overall variability in the dataset which amounts to a total of 99% of the total variability. In the second characteristic vector, $\mathbf{u_2}$, the negative values of the fourth and fifth coefficients contrast with the other positive values. This represents a difference between these variables and the other four, but accounts for less than 1% of the total variability. The other three characteristic vectors represent only a minuscule amount of variability.

The ith PC in Table 5.3 is formed according to formula (5.1) and is the transformed ith variable. The transformed observations are referred to as z-scores, a term first used in psychology and education particularly in conjunction with factor analysis, but now ubiquitous in PCA [59].

Table 5.3: PCs (variables) and z-scores (observations)

	PC$_1$	**PC$_2$**	**PC$_3$**	**PC$_4$**	**PC$_5$**	**PC$_6$**
z-score$_1$	0.8409	-0.0406	-4.5042e-10	1.3722e-11	9.4718e-16	0
z-score$_2$	0.9033	-0.0475	-4.0307e-10	1.6307e-11	3.5449e-16	0
z-score$_3$	0.9666	-0.0544	-3.1625e-10	1.8927e-11	9.9147e-17	0
z-score$_4$	0.8877	-0.0458	-4.1787e-10	1.5656e-11	7.1670e-16	0
z-score$_5$	0.7621	-0.0320	4.7545e-10	1.0614e-11	4.2195e-16	0
z-score$_6$	0.4686	0.0002	4.1420e-10	1.7012e-12	4.7189e-16	0
z-score$_7$	0.0917	0.0416	-2.2090e-10	-3.5331e-12	6.0578e-16	0
z-score$_8$	-0.1651	0.0698	-6.9199e-11	-4.1666e-12	7.6954e-17	0
z-score$_9$	-0.2915	0.0836	7.6070e-12	-3.8990e-12	3.3568e-16	0
z-score$_{10}$	-0.0082	0.0525	-1.6284e-10	-4.0011e-12	3.6388e-16	0
z-score$_{11}$	0.3742	0.0106	-3.7177e-10	-2.0747e-13	7.9185e-16	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
z-score$_{45}$	-3.3392	-0.1540	7.9781e-11	1.5105e-11	7.8230e-16	0
Variance	1.6383	0.0088	1.53e-18	2.85e-22	3.37e-31	0

Step3: Determine the subspace cutoff criteria, P. The subspace cutoff, P, is used to reduce the dimensionality of the data, $\mathbf{X}$, and ultimately to discover anomalous observations through residual analysis.

To demonstrate the differences among the various cutoff criteria from the literature, the cutoff value is determined using each criterion described in section 4.3 above.

- Average eigenvalue criterion (Kaiser-Guttman): In Table 5.2 the average eigenvalue is 1.6472/6=0.27. Since only the first eigenvalue, 1.6383 is greater than 0.27, $P = 1$.

- Average eigenvalue criterion with Jolliffe modification: The average eigenvalue is 0.27 as found in Table!5.2. Jolliffe suggests for covariance matrices keeping PCs where the eigenvalue is greater than 70% of the average or $0.27 \times 0.70 = 0.2515$. Only the first eigenvalue meets the criterion therefore $P = 1$.

- Cattell's Scree Test: In Fig. 5.2 it appears that the elbow occurs at the second component so $P = 2$ including the elbow or $P = 3$ including the first component past the elbow.

- Cumulative variance accounted for: To detect anomalies, the cumulative variance accounted for by the common subspace needs to be very close to 100%. Choosing the cutoff to be .99999, $P = 2$.

- Broken Stick Method: In Fig. 5.4 the green bars represent the proportions of 6 independent variables defined by equation (4.4) with values (2.45, 1.45, 0.95, 0.6167, 0.3667, 0.1667) and proportions (40.8333, 24.1667, 15.8333, 10.2778, 6.1111, 2.7778). The blue line represents the the proportion of variance explained by each PC (99.4629, 0.5371, 9.3182e-17, 1.7298e-20, 2.04646e-19, 0). Since only the first PC proportion is greater than that expected by chance, $P = 1$.

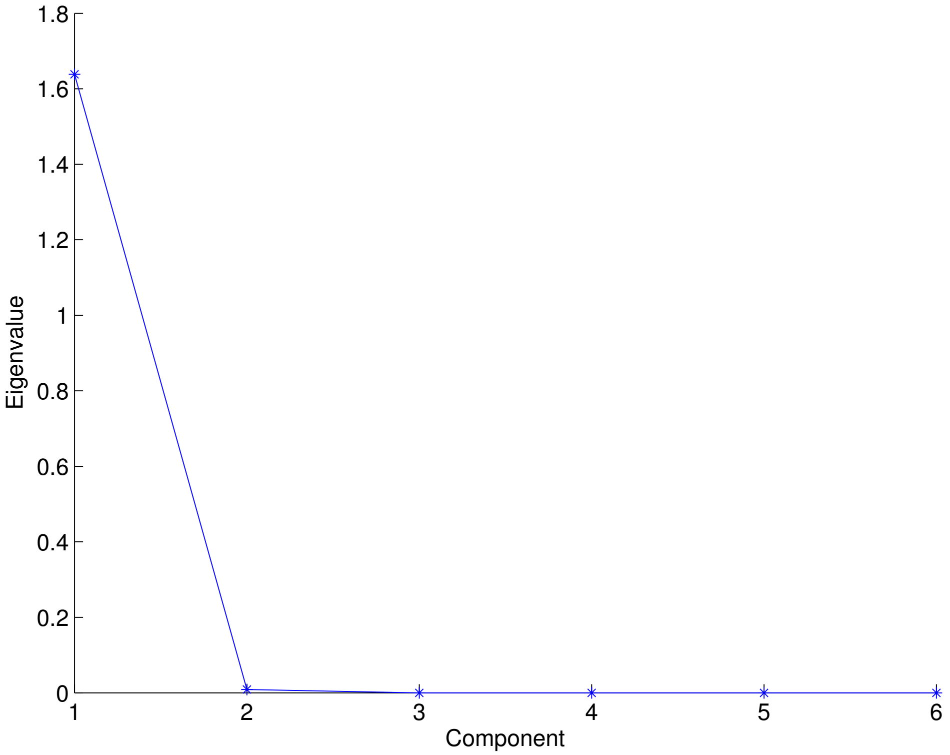


Figure 5.2: Cattell's Scree Test

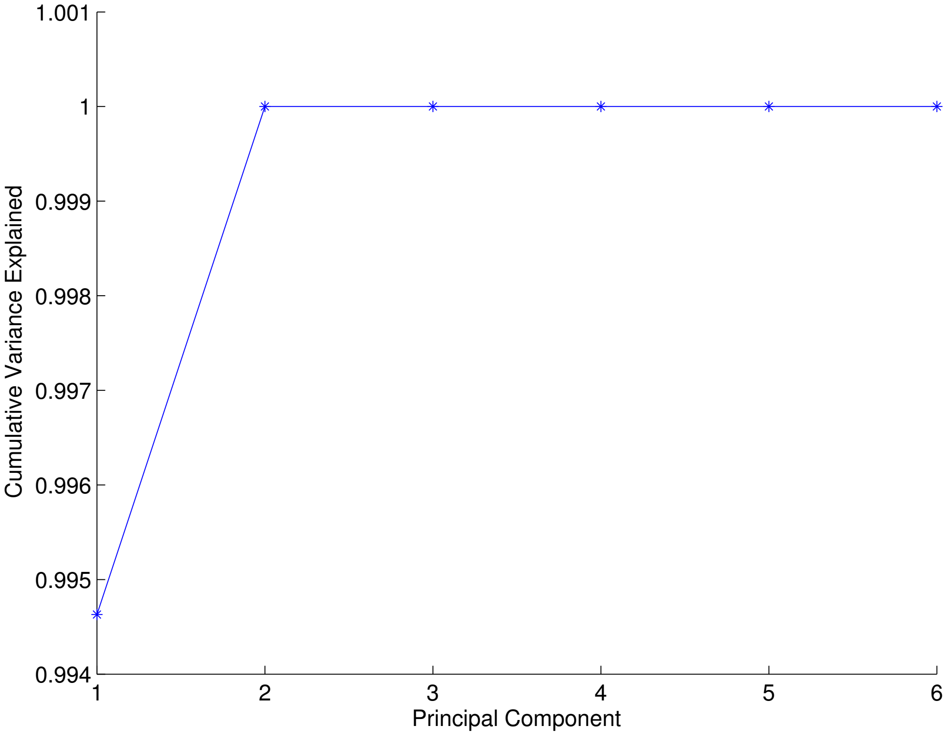


Figure 5.3: Cumulative Variance Accounted For

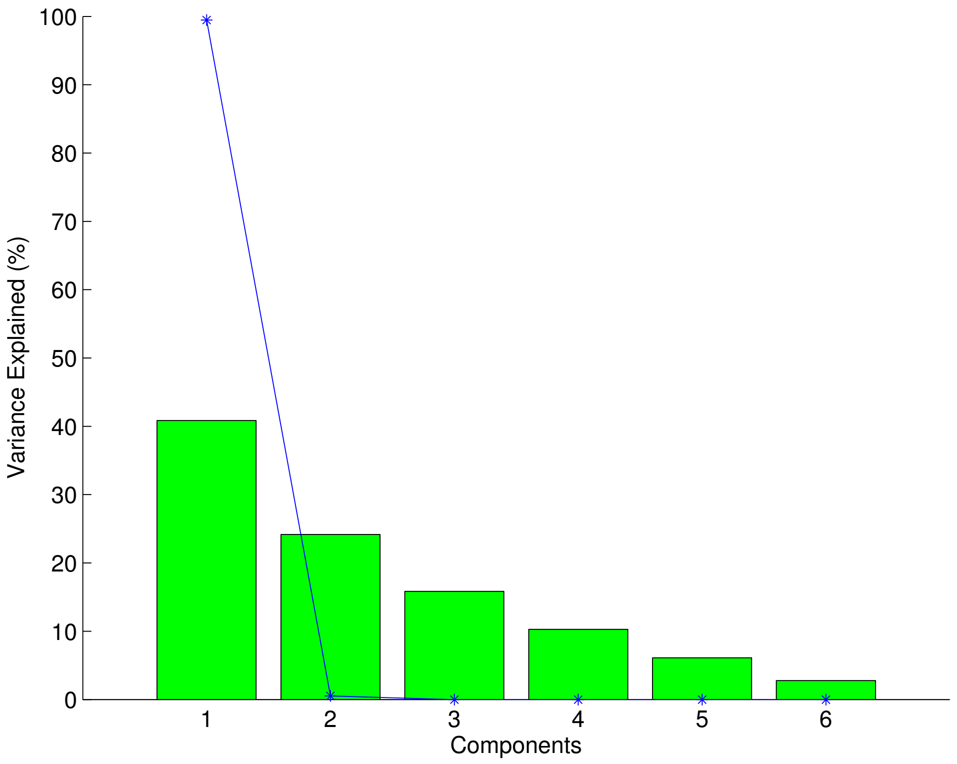


Figure 5.4: Broken Stick Method

- Log Eigenvalue Diagram: Fig. 5.5 does not improve the decision-making process from the Scree Test in Fig. 5.2, but it appears that the components after 2 are in a line, so $P = 2$.

For the remainder of the study, the *Cumulative Variance Accounted For* criterion will be used to determine P. It should also be noted here that each time the anomaly detection program is run is considered a new instance since conditions change over time. The common cause subspace size, P, is determined for each specific instance and will likely differ over time.

If all of the PCs are used then the original variables in $\mathbf{X}$ can be recreated from Equation (5.1).

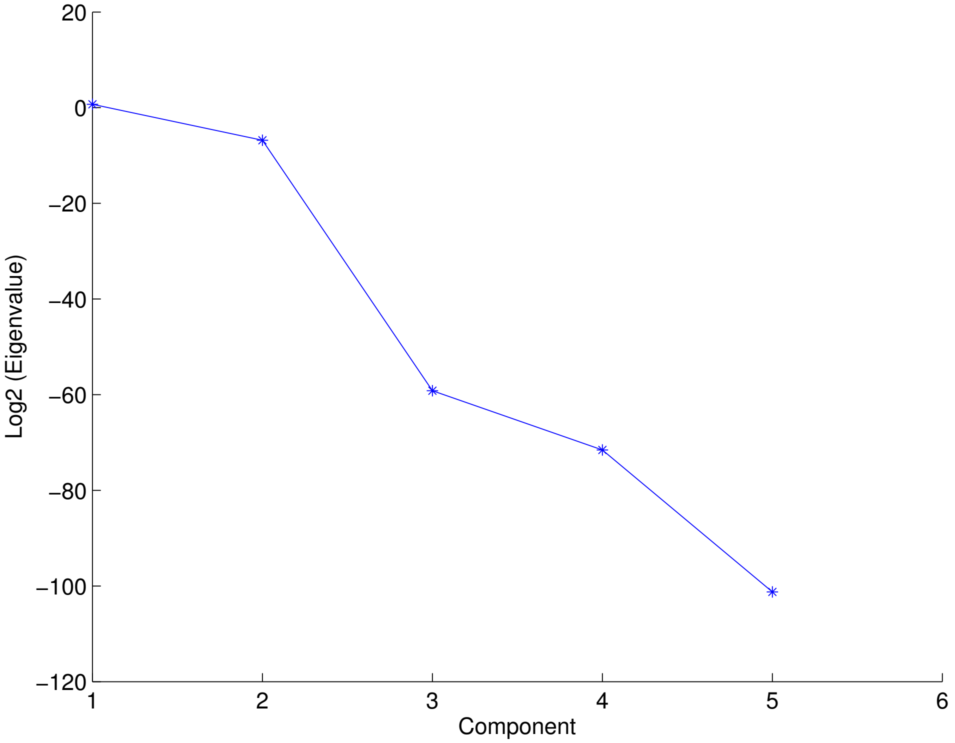


Figure 5.5: Log Eigenvalue Test

Step4: Insert an anomaly. To simulate the anomaly detection process, an observation that contains anomalous data is inserted as observation 46 in the original data. In this simulation, it is assumed that the cyberintruder has access to the database that stores parameters and has modified one parameter, resulting in the following anomalous observation.

$$\mathbf{x^{NEW}} = \begin{bmatrix} 0 & 0.0817 & 0.8585 & -3.3560 & -3.4707 & -2.2216 \end{bmatrix} \tag{5.6}$$

Step5: Calculate Q_α using Equation (5.3), $Q_\alpha = 3.3280 \times 10^{-17}$.

Step6: Perform residual analysis. The mean of the observations on which PCA was performed summed with the contribution of the 2 PCs is an approximation of the dataset $\mathbf{X}$. Subtracting that sum, $\hat{\mathbf{x}}$, from the new observation, $\mathbf{x}^{NEW}$ results in the difference between the model and the new observation, the residual, as in Equation (5.2). The 2 PC model is expected to well represent the variation in **X**, so the sum of the residuals should be negligible. The values discussed above are given in Table 5.4.

Table 5.4: Residuals

Variable	$\bar{\mathbf{x}}$	**PC_contribution**	$\mathbf{x}^{\mathbf{NEW}}$	$\hat{\mathbf{x}}$	**Residual**
1	0	0	0	0	0
2	0.8272	-0.6832	0.0817	0.1440	-0.0623
3	1.4788	-0.6564	0.8585	0.8224	0.0361
4	-2.1574	-1.1822	-3.3560	-3.3396	-0.0164
5	-2.1927	-1.2911	-3.4707	-3.4838	0.0131
6	-1.0251	-1.2139	-2.2216	-2.2390	0.0174

Calculate the sum of squares of the residuals using Equation (5.7) and compare to Q_α.

$$Q = (\mathbf{x}^{\mathbf{NEW}} - \hat{\mathbf{x}})'(\mathbf{x}^{\mathbf{NEW}} - \hat{\mathbf{x}}) = 0.0059 \tag{5.7}$$

Since $0.0059(Q) > 3.3280 \times 10^{-17}(Q_\alpha)$ the observation is considered anomalous and the observation is flagged for special attention.

Chapter 6

Experimental Cyberattack to Electric Power Transmission Systems

This research defined a new class of cyberattacks to power systems—malicious modification of network data stored in an accessible database. Data anomalies could be the result of unauthorized access to and modification of data by an intruder or by malicious code that is not quarantined by preventive software. Since modern control centers use state-of-the-art security to prevent cyberintrusion based on recommendations, regulations, and requirements from agencies like NERC, NIST, and DOE [7], operators expect the data stored in protected databases to be secure. A cybercriminal could manage to get around security measures and modify data in a database without the system operator's knowledge.

6.1 Electric Power Transmission System

An electric power transmission system (denoted by green lines in Fig. 6.1) is designed to transport high-voltage electrical energy from power generating plants to electrical substations which are typically located near customer demands. The three major parts of the electric power transmission system referenced here are buses, lines, and generators. The common connection point for the lines, demand sources, and generation sources is referred to as the bus. The actual bus is located in a substation (denoted by blue boxes in Fig. 6.1) along with transformers, switches, and other equipment. The bus (busbar) conducts electricity in a substation. The lines are the structures through which the electrical energy is transported between buses. And, the generators are the machines in a power plant that generate the electrical energy.

Electric transmission systems are complex networks where electricity flows at the speed of light, according to the laws of physics. Multiple input variables and parameters influence

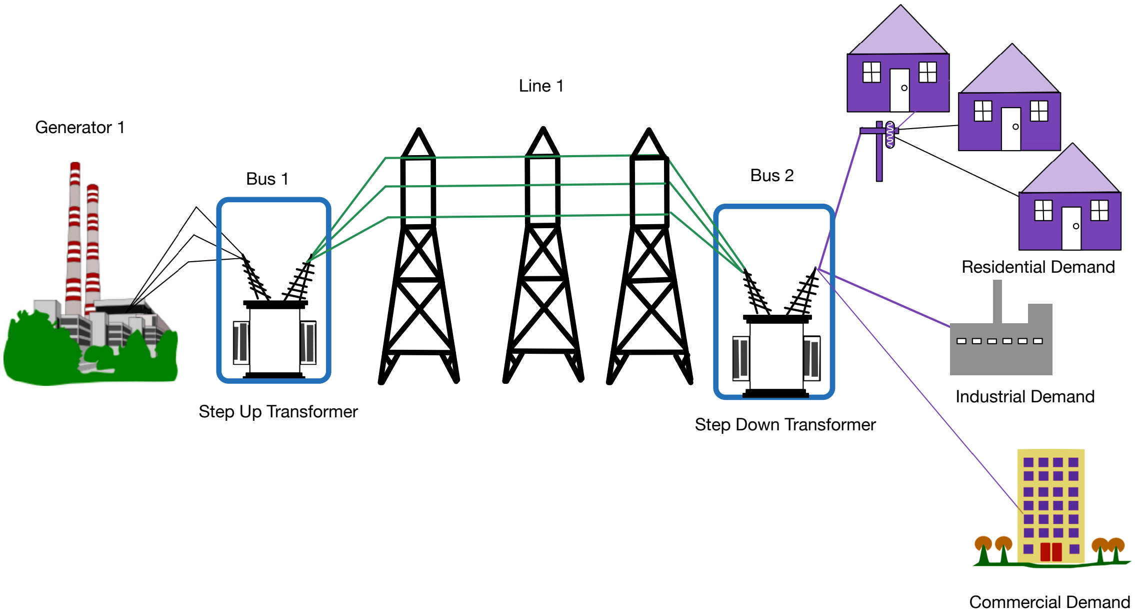


Figure 6.1: Electric System

the flow through the lines and system decision variables voltage angles and amount of generation at each bus. The study of observations from these complicated systems requires the simultaneous analysis of multiple variables at a single point in time.

6.2 Optimal Power Flow (OPF)

Today power is generated by typical steam powered units (fueled by coal, natural gas or nuclear fission), wind turbines, hydroelectric generators, and solar collectors. The decision as to which generators to use at a given point to economically, reliably, and environmentally provide electric power to customers is the goal of the OPF program. Optimization is the basis for economically efficient and reliable electric energy markets. Computationally, the optimizations include nonlinearities and nonconvexities and are difficult to model.

The modern power transmission control center uses multiple instances of the OPF module, often in real-time, to operate the power system as economically as possible while ensuring its reliability despite changes in demand requirements and available resources. The OPF module provides an optimal solution of flows, voltages, and power injections either to the operator or as input to automated generation control (AGC) programs and is commonly executed every 3 minutes with an updated set of values for input parameters [20, 82]. Some of the input values, such as line characteristics and network configuration, typically do not change over a short period of time whereas others, such as customer demand and the set of available generators, vary more often. Therefore, the solution vector of the OPF module changes over time based on the common cause variability in the module's input data.

In the OPF models that follow [53, 80] N^b is the number of buses and N^g is the number of generators in the network.

6.2.1 AC OPF Model

The AC optimal power flow program simultaneously optimizes the reactive and real power flow and its formulation has changed little since it was first discussed by Carpentier in 1962 [83].

Decision Variables.

$$\mathbf{x} = \begin{bmatrix} \boldsymbol{\theta} \\ \mathbf{v} \\ \mathbf{g} \\ \mathbf{q} \end{bmatrix} \tag{6.1}$$

$\boldsymbol{\theta}$ and $\mathbf{v}$ are $N^b \times 1$ vectors of voltage angles and magnitudes, respectively; $\mathbf{g}$ and $\mathbf{q}$ are $N^b \times 1$ vectors of generator real and reactive power injections.

Objective Function.

$$\min_{\boldsymbol{\theta},\mathbf{v},\mathbf{g},\mathbf{q}} \sum_{i=1}^{N^g} C_i g_i \tag{6.2}$$

The objective function sums the polynomial costs, C_i of real power injections for each generator. We particularly focus on the OPF instance whose objective is to find a steady state operation point which minimizes generation cost while enforcing system performance through limits on real and reactive power generator outputs, bus voltage angles and magnitudes, and transmission line flows.

Constraints. Three types of constraints are present in the model—power balance equations, power flow limits and variable limits. Real power balance equations (6.3) ensure that the real power in and out of each bus is equal. Reactive power balance equations (6.4) provide the same assurance for reactive power. The power flow limit equations, (6.5) and (6.6) limit the flow through each transmission line to that which can safely be carried through the line. Equation (6.7) constrains the real power generation at each generator to limits defined by the mechanical capacity of the generating unit and equation (6.8) does the same for reactive power generation. Equations (6.9) and (6.10) restrict the values of voltage magnitudes and voltage angles at each bus to reasonable values. Note that capital letters represent parameters stored in a database.

$$\sum_{i=1}^{N^b} v_i v_j \frac{{R_{ij}}^2}{{R_{ij}}^2 + \chi_{ij}^2} \cos(\theta_i - \theta_j) - \frac{\chi_{ij}}{R_{ij}^2 + \chi_{ij}^2} \sin(\theta_i - \theta_j) - g_i - d_i = 0 \qquad \forall j \in N^b \tag{6.3}$$

$$\sum_{i=1}^{N^b} v_i v_j \frac{{R_{ij}}^2}{{R_{ij}}^2 + \chi_{ij}^2} \sin(\theta_i - \theta_j) - \frac{\chi_{ij}}{R_{ij}^2 + \chi_{ij}^2} \cos(\theta_i - \theta_j) - q_i - d_i = 0 \qquad \forall j \in N^b \tag{6.4}$$

$$\frac{(\theta_i - \theta_j)}{\chi_{ij}} - F_{ij}^{\mathrm{MAX}} \leq 0 \qquad \forall i,j \in N^b \tag{6.5}$$

$$-\frac{(\theta_i - \theta_j)}{\chi_{ij}} - F_{ij}^{\mathrm{MAX}} \leq 0 \qquad \forall i,j \in N^b \tag{6.6}$$

$$G_i^{\mathrm{MIN}} \leq g_i \leq G_i^{\mathrm{MAX}} \qquad \forall i \in N^b \tag{6.7}$$

$$Q_i^{\mathrm{MIN}} \leq q_i \leq Q_i^{\mathrm{MAX}} \qquad \forall i \in N^b \tag{6.8}$$

$$V_i^{\mathrm{MIN}} \leq v_i \leq V_i^{\mathrm{MAX}} \qquad \forall i \in N^b \tag{6.9}$$

$$\Theta_i^{\mathrm{MIN}} \leq \theta_i \leq \Theta_i^{\mathrm{MAX}} \qquad \forall i \in N^b \tag{6.10}$$

The formula for real power flow between bus i and bus j is in equation (6.11)

$$f_{ij}^{P} = \frac{v_i v_j}{\chi_{ij}} \sin(\theta_i - \theta_j) \tag{6.11}$$

and the formula for reactive power flow between bus i and bus j is in equation (6.12).

$$q_{ij} = \frac{v_i \times (v_i - v_j)}{\chi_{ij}} cos(\theta_i - \theta_j) \tag{6.12}$$

6.2.2 DC OPF Model

The DC power flow model can be used to approximate the AC model because of the following observations about high voltage transmission lines:

Observation 1: The resistance of transmission circuits is much less than the reactance.

Observation 2: For most typical operating conditions, the angular separation $(\theta_{\mathbf{k}} - \theta_{\mathbf{j}})$ between two buses, k and j is less than 10-15 degrees, small enough to ensure that $\sin(\theta^{\circ}_{\mathbf{k}} - \theta^{\circ}_{\mathbf{j}}) \approx \theta^{\circ}_{\mathbf{k}} - \theta^{\circ}_{\mathbf{j}}$.

Observation 3: In the per-unit (pu) system, the numerical values of voltage magnitudes, $\mathbf{v_k}$ and $\mathbf{v_j}$ are very close to 1.0, typically ranging from .95 to 1.05.

Decision Variables.

$$\mathbf{x} = \begin{bmatrix} \boldsymbol{\theta} \\ \mathbf{g} \end{bmatrix}, \tag{6.13}$$

where $\boldsymbol{\theta}$ is the $N^b \times 1$ vector of voltage angles and $\mathbf{g}$ is the $N^g \times 1$ vector of generator real power injections.

Objective Function. The objective function sums the cost function $C_i g_i$ of the cost, C_i, of real power injections for each generator, g_i. Focus is on the OPF instance whose objective is to find a steady state operation point which minimizes generation cost while enforcing system performance through limits on real power generator outputs and transmission line flows.

$$\underset{\boldsymbol{\theta},\mathbf{g}}{\text{minimize}} \sum_{i=1}^{N^g} C_i g_i \tag{6.14}$$

Constraints. Three general types of constrains are present in the model. The real power balance equation (6.15) ensures that the power flow into and out of each bus is equal. Line flow limits, equations (6.16) and (6.17) ensure that power flow is limited on a particular line to that which can safely be handled by the line. Equation (6.18) restricts the size of power generation to acceptable values.

$$\sum_{i=1}^{N^b} H_{ij} \frac{(\theta_i - \theta_j)}{\chi_{ij}} + g_i - d_i = 0 \qquad \forall j \in N^b \tag{6.15}$$

$$\frac{(\theta_i - \theta_j)}{\chi_{ij}} - F_{ij}^{\text{MAX}} \leq 0 \qquad \forall i, j \in N^b \tag{6.16}$$

$$-\frac{(\theta_i - \theta_j)}{\chi_{ij}} - F_{ij}^{\text{MAX}} \leq 0 \qquad \forall i, j \in N^b \tag{6.17}$$

$$G_i^{\text{MIN}} \leq g_i \leq G_i^{\text{MAX}} \qquad \forall i \in N^b \tag{6.18}$$

This DC model is a simplification of the AC power flow model. In power transmission, the DC approximation to the AC model is often used for decision making since it is a linear model and provides a more tractable solution. Note that parameters in the model are represented by capital letters.

6.2.3 Solving the OPF

The optimization problem includes parameter input stored in a network accessible database, like measures of a line's reactance, χ_{ij} and a line's flow capacity, F_{ij}^{MAX}, which typically remain the same over time. Customer demand, d_i is a continuous input variable approximated by an hourly sample. The amount of real power generation, g_i, and reactive power generation, q_i, at each bus changes with demand and the operational and reliability limitations of the available generators and the transmission infrastructure. The service status of transmission lines is impacted by weather and maintenance schedules and the service status of generators is impacted by maintenance schedules and the cost and availability of fuel. For research purposes in this paper, the vector of voltage angles, $\boldsymbol{\theta}$, at each bus resulting from the OPF optimization is the variable input to the detection algorithm.

Both the DCOPF and ACOPF functions in MATPOWER use the Matlab Interior Point Solver (MIPS) to solve the OPF.

6.3 The Detection Process

As explained in Chapter 2 no known cyberattacks have been successfully carried out on the U.S. transmission network so it is difficult to identify a typical attack. To study the performance of the detection approach, for each experiment the simulation is run for 672 hours under regular operating conditions that include load variability to establish the model against which the 763rd observation is evaluated. The 2nd through 763rd intervals form the model to evaluate the 764th observation, etc. until the 1440th observation is evaluated from the PCA run on the 678th through 1439th intervals. This method of evaluation is referred to as "rolling" since PCA is run on the previous 762 intervals for each observation that is evaluated.

A cyberattack is represented through the introduction of an anomaly to the observation for evaluation prior to running the OPF module. Incorporated into the rolling method is the recalculation of P— the size of the common cause subspace, Q_α—the threshold, and subsequently Q—the sum of squares of the residual remaining when the observation is analyzed against the PCA model. The process is repeated for the 673rd through 1440th observations–a total of 768 observations. At each hour the same anomaly is injected.

6.4 Case Study, IEEE 24-bus test system

The IEEE 24-bus test system is a well-documented and frequently used medium sized network for testing electric transmission system applications. The test system was designed by a committee of engineers in 1979 to assist researchers in testing transmission and generation reliability. The system as included in MATPOWER provides all of the information necessary to test a DC system. The 24-bus system, pictured in Fig. 6.2 consists of 38 transmission lines, 24 buses of which 17 have demands, and 33 generators. Line, bus and generator data used by the DCOPF program is described in section 6.4.1. For the interested reader complete information relative to the IEEE 24-bus test system is documented in [84].

6.4.1 Data Descriptions

The 24-bus system consists of 24 buses marked with a heavy line in Fig. 6.2 and described in Table 6.1. There are 13 load buses, 10 generator buses and 1 reference bus—13. Buses 1-10 have a nominal rating of 138 KV and buses 11-24 have a nominal rating of 230 KV.

Table 6.1: Bus Descriptions

Bus	**Type**	**baseKV**	**Bus**	**Type**	**baseKV**
1	Gen	138	**13**	Ref	230
2	Gen	138	**14**	Gen	230
3	Load	138	**15**	Gen	230
4	Load	138	**16**	Gen	230
5	Load	138	**17**	Load	230
6	Load	138	**18**	Gen	230
7	Gen	138	**19**	Load	230
8	Load	138	**20**	Load	230
9	Load	138	**21**	Gen	230
10	Load	138	**22**	Gen	230
11	Load	230	**23**	Gen	230
12	Load	230	**24**	Load	230

A Monte Carlo simulation of the power system is coded to generate historical voltage angle data since actual state variable results are difficult to obtain. In the simulation, actual

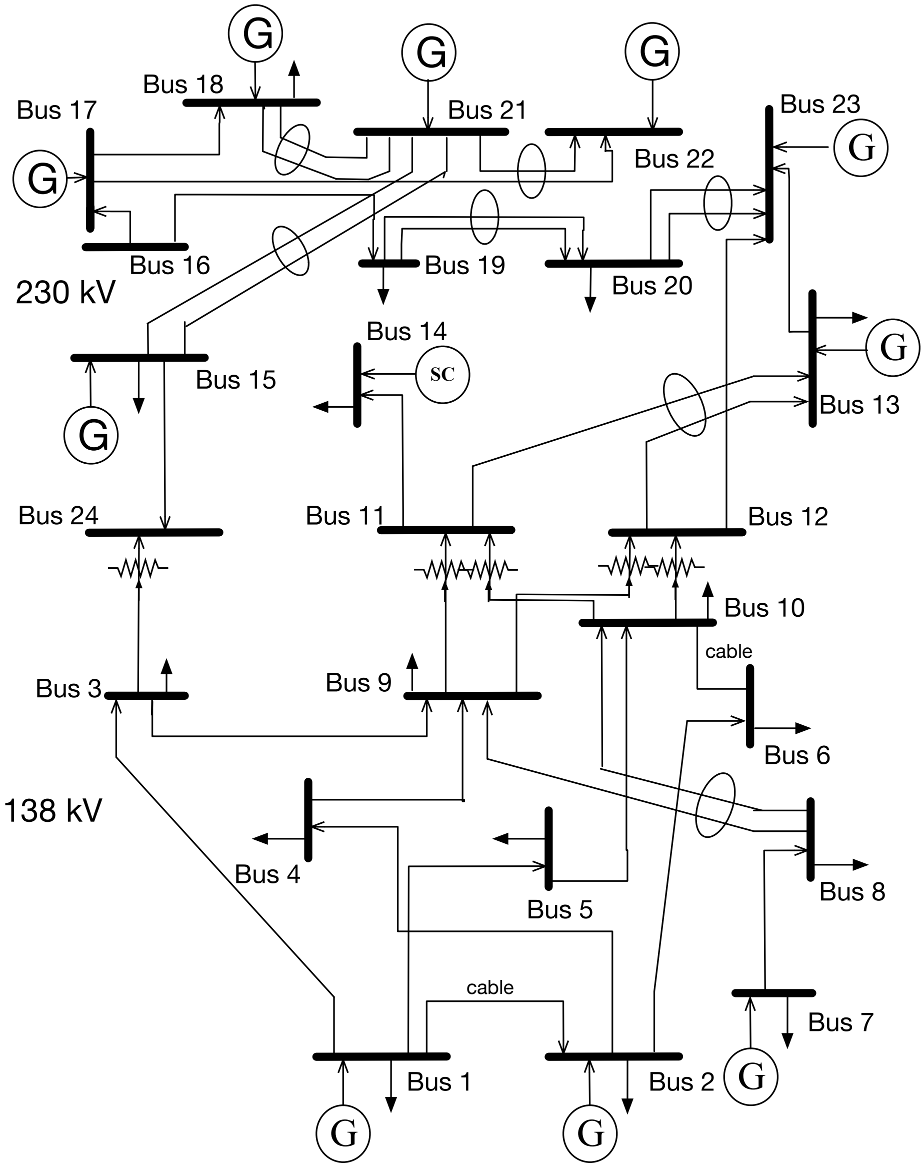


Figure 6.2: One Line Diagram of IEEE 24-Bus Test System

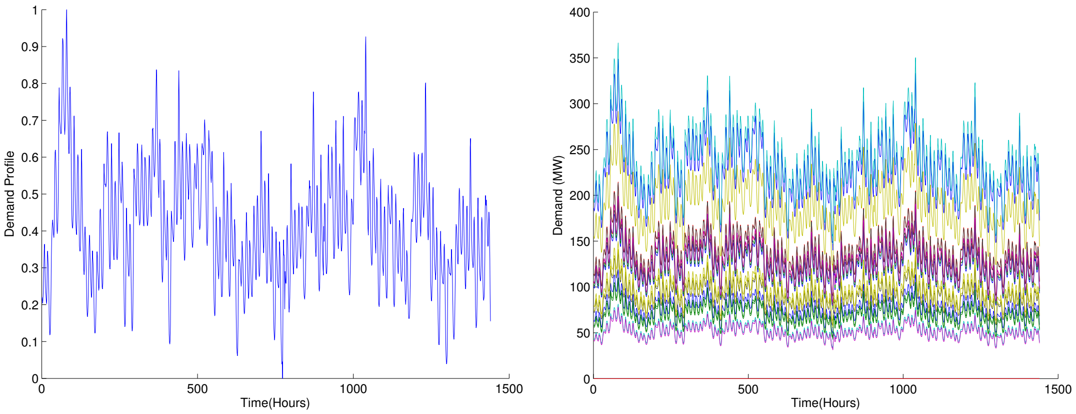
Table 6.2: Line Descriptions

Line	From	To	Reactance	Capacity
	Bus	Bus	(p.u.)	(MW)
1	1	2	0.0139	175
2	1	3	0.2112	175
3	1	5	0.0845	175
4	2	4	0.1267	175
5	2	6	0.1920	175
6	3	9	0.1190	175
7	3	24	0.0839	400
8	4	9	0.1037	175
9	5	10	0.0883	175
10	6	10	0.0605	175
11	7	8	0.0614	175
12	8	9	0.1651	175
13	8	10	0.1651	175
14	9	11	0.0839	400
15	9	12	0.0839	400
16	10	11	0.0839	400
17	10	12	0.0839	400
18	11	13	0.0476	500
19	11	14	0.0418	500
20	12	13	0.0476	500
21	12	23	0.0966	500
22	13	23	0.0865	500
23	14	16	0.0389	500
24	15	16	0.0173	500
25	15	21	0.0490	500
26	15	21	0.0490	500
27	15	24	0.0519	500
28	16	17	0.0259	500
29	16	19	0.0231	500
30	17	18	0.0144	500
31	17	22	0.1053	500
32	18	21	0.0259	500
33	18	21	0.0259	500
34	19	20	0.0396	500
35	19	20	0.0396	500
36	20	23	0.0216	500
37	20	23	0.0216	500
38	21	22	0.0678	500

Table 6.3: Generator Descriptions

Generator	**Bus**	**G$^{\mathbf{MAX}}$ (MW)**	**G$^{\mathbf{MIN}}$ (MW)**	**Cost ($/MWh)**
1	1	20	16	130
2	1	20	16	130
3	1	76	15.2	16.0811
4	1	76	15.2	16.0811
5	2	20	16	130
6	2	20	16	130
7	2	76	15.2	16.0811
8	2	76	15.2	16.0811
9	7	100	25	43.6615
10	7	100	25	43.6615
11	7	100	25	43.6615
12	13	197	69	48.5804
13	13	197	69	48.5804
14	13	197	69	48.5804
15	14	0	0	0
16	15	12	2.4	56.5640
17	15	12	2.4	56.5640
18	15	12	2.4	56.5640
19	15	12	2.4	56.5640
20	15	12	2.4	56.5640
21	15	155	54.3	12.3883
22	16	155	54.3	12.3883
23	18	400	1	4.4231
24	21	400	1	4.4231
25	22	50	10	0.0010
26	22	50	10	0.0010
27	22	50	10	0.0010
28	22	50	10	0.0010
29	22	50	10	0.0010
30	22	50	10	0.0010
31	23	155	54.3	12.3883
32	23	155	54.3	12.3883
33	23	350	140	11.8495

demand data for 1440 intervals from the online PJM database [81] is standardized and scaled so that the value 1.0 corresponds to the peak demand in the data.



(a) Profile

(b) Demand

Figure 6.3: Demand Profile and Simulated Data

The demand at each hour is calculated by applying the demand profile to the demand supplied by the IEEE test system at each bus where there is demand. Although the demand at each bus follows the same time variability pattern, the actual values are distinct at different buses. The demand profile and simulated data for each bus at each hour are plotted in Fig. 6.3.

With the simulated demand data as input, the optimal voltage angle at each bus for each interval is computed by the DCOPF software module from MATPOWER and plotted in Fig. 6.4.

6.4.2 PCA

For illustration purposes, the figures and tables in subsections 6.4.2 and 6.4.3 are based on analysis from the first 672 observations.

The 672 × 24 matrix, **X**, is formed by extracting the first 672 observations of the simulated voltage angle data. The Gleason-Staelin statistic for the PJM data modeled as a 24-bus system, $\Phi = 0.8749$, indicates that just over 87% of the variables are correlated and it is worthwhile to use PCA to analyze the data. PCA is run on the covariance matrix (**S** from

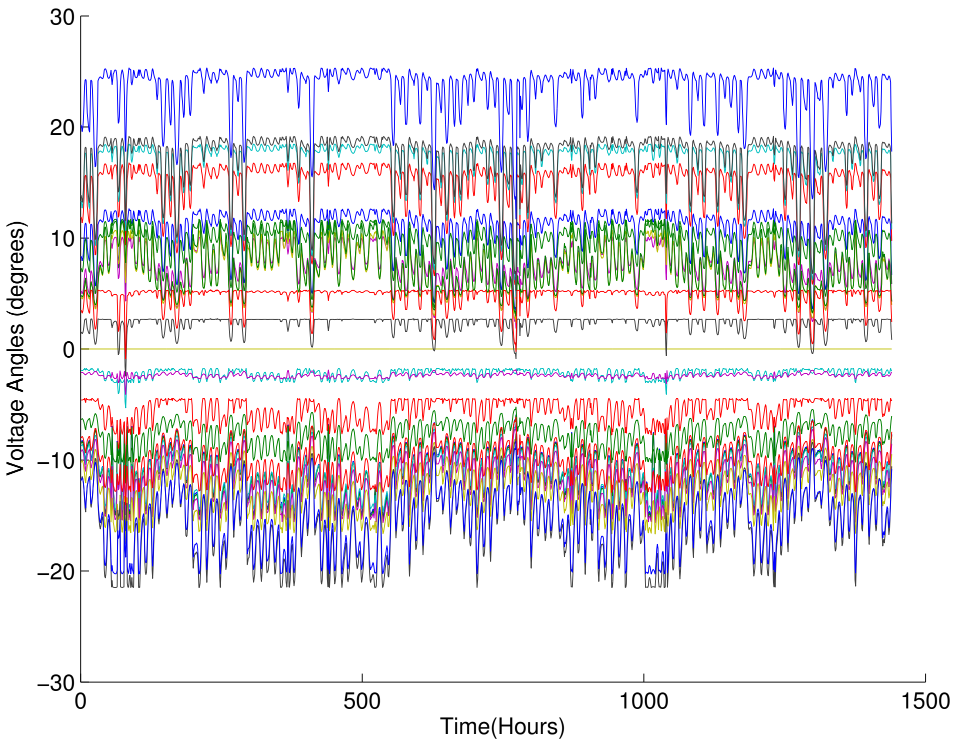


Figure 6.4: Voltage Angle Data

equation 5.5) of mean-centered voltage angle data. The **u** vectors are the PC coefficients and the eigenvectors of **S**. The first PC is formed from equation (6.19).

$$\mathbf{u}_1^T\mathbf{x} = u_{11}\mathbf{x}_1 + u_{12}\mathbf{x}_2 + \ldots + u_{1J}\mathbf{x}_J = \sum_{i=1}^{J} u_{1i}\mathbf{x}_i \tag{6.19}$$

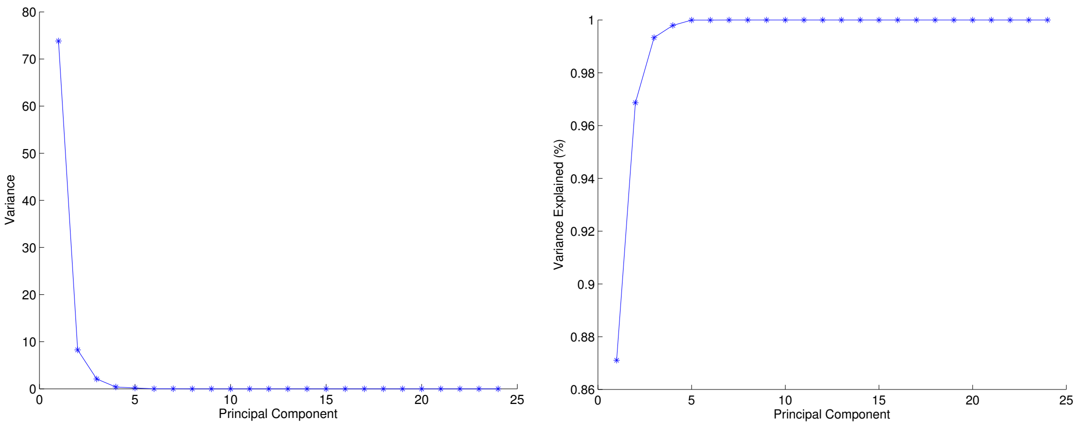
Each of the PCs is orthogonal to all of the other PCs. The first 10 PC coefficients are detailed in Table 6.4. Note that in the first **u** vector the first 12 coefficients are positive, the 13th coefficient is 0 and the remaining 11 coefficients are negative. This represents a contrast in the first 12 and last 11 variables. In Fig. 6.2 note that the bottom portion of the network is made up of 138kV lines and the top portion of the network is made up of 230 kV lines. The first vector of PC coefficients accounts for 87.11% of the total variance and represents this difference.

Table 6.4: First 10 PC coefficients (**U** vectors)

Variable	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**	**PC10**
1	0.24	0.08	0.56	0.18	-0.06	0.12	-0.02	-0.60	0.09	-0.00
2	0.25	0.08	0.56	0.19	-0.06	0.12	-0.02	0.62	-0.04	-0.01
3	0.11	0.21	-0.01	-0.33	0.03	0.00	0.06	-0.14	-0.17	0.01
4	0.21	0.14	0.17	-0.17	0.01	-0.08	0.01	0.30	-0.03	0.00
5	0.23	0.12	0.23	-0.12	-0.01	-0.07	-0.00	-0.33	0.06	0.00
6	0.24	0.16	0.03	-0.33	0.02	-0.20	0.02	0.12	-0.01	0.01
7	0.37	0.18	-0.26	0.54	0.50	0.11	-0.01	-0.00	-0.00	-0.00
8	0.32	0.20	-0.25	0.03	0.30	-0.10	0.01	-0.01	0.00	0.00
9	0.14	0.14	-0.05	-0.26	0.08	-0.12	0.02	0.03	-0.03	0.00
10	0.18	0.13	-0.02	-0.27	0.07	-0.16	0.01	-0.03	0.04	0.00
11	0.04	0.10	-0.03	-0.17	0.06	0.01	-0.02	0.01	0.12	-0.01
12	0.02	0.02	0.04	-0.12	0.13	-0.11	0.01	-0.00	-0.03	0.00
13	0	0	0	0	0	0	0	0	0	0
14	-0.05	0.16	-0.05	-0.23	0.10	0.17	-0.07	0.03	0.33	-0.04
15	-0.18	0.24	0.02	-0.06	0.09	0.49	0.10	-0.01	-0.51	0.01
16	-0.18	0.17	0.05	-0.08	0.15	0.44	-0.13	0.06	0.52	-0.07
17	-0.21	0.32	0.03	0.08	-0.01	-0.07	-0.36	0.01	-0.05	-0.17
18	-0.22	0.40	0.02	0.16	-0.09	-0.33	-0.58	-0.02	-0.20	-0.16
19	-0.19	0.02	0.11	-0.14	0.30	0.09	-0.05	0.02	0.22	-0.03
20	-0.23	-0.13	0.21	-0.08	0.44	-0.16	0.01	-0.01	-0.03	-0.00
21	-0.23	0.38	0.04	0.19	-0.08	-0.22	0.65	0.01	0.12	-0.52
22	-0.22	0.36	0.04	0.14	-0.05	-0.16	0.25	0.02	0.13	0.83
23	-0.25	-0.23	0.29	-0.02	0.52	-0.26	0.04	-0.02	-0.17	0.01
24	-0.07	0.23	0.01	-0.16	0.07	0.30	0.08	-0.06	-0.38	0.01
Variance	73.81	8.27	2.09	0.39	0.17	0.00	0.00	0.00	0.00	0.00
Cum Var %	87.11	96.87	99.34	99.80	99.99	1	1	1	1	1

6.4.3 P and Q_α

The results of Cattell's Scree test in Fig. 6.5(a) indicate that the first 6 PCs should be included in the common cause subspace accounting for a total variance of .999999 and plotted in Fig. 6.5(b). These first 6 PCs are considered the common cause subspace ($P = 6$) and the variance associated with them can be attributed to common causes like the size of the lines connecting the buses and other traits inherent to the voltage angles at each of the buses. The remaining 18 PCs are considered the assignable cause subspace and make up a



(a) Cattell's Scree Test

(b) Cumulative Variance

Figure 6.5: Variance Accounted for by 6 PCs is 0.999999

Table 6.5: Variance and cumulative variance for each principal component

PC	Variance	Cumulative Variance	PC	Variance	Cumulative variance
1	73.8122	0.8712	13	5.3322e-30	1
2	8.2693	0.9687	14	3.5378e-30	1
3	2.0932	0.9934	15	2.7701e-30	1
4	0.3906	0.9980	16	1.8572e-30	1
5	0.1721	0.999996	17	1.3736e-30	1
6	0.0003	0.999999	18	1.0660e-30	1
7	4.2889e-13	1	19	9.5295e-31	1
8	5.5422e-18	1	20	8.2343e-31	1
9	3.1365e-18	1	21	5.6158e-31	1
10	3.9932e-19	1	22	4.9512e-31	1
11	5.4423e-28	1	23	1.1173e-31	1
12	1.5510e-29	1	24	1.4800e-62	1

small percentage of the variance in the data. Table 6.5 gives the variance and cumulative variance for all of the PCs.

6.4.4 Case A

The experiment in Case A assumes that a cyberattacker has accessed the database containing the system data model and has modified the reactance, χ_{ij}. The experiment is run for each line for each interval with six different factors, increased by 80%, increased by 50%, increased by 10%, decreased by 10%, decreased by 50% and decreased by 80%.

The summary in Table 6.6 includes only the lines where not all of the anomalies are recognized. In all of the lines not listed in the table, 76%, all of the simulated anomalies were recognized by the algorithm.

Table 6.6: Results in which less than 100% of the anomalies are found

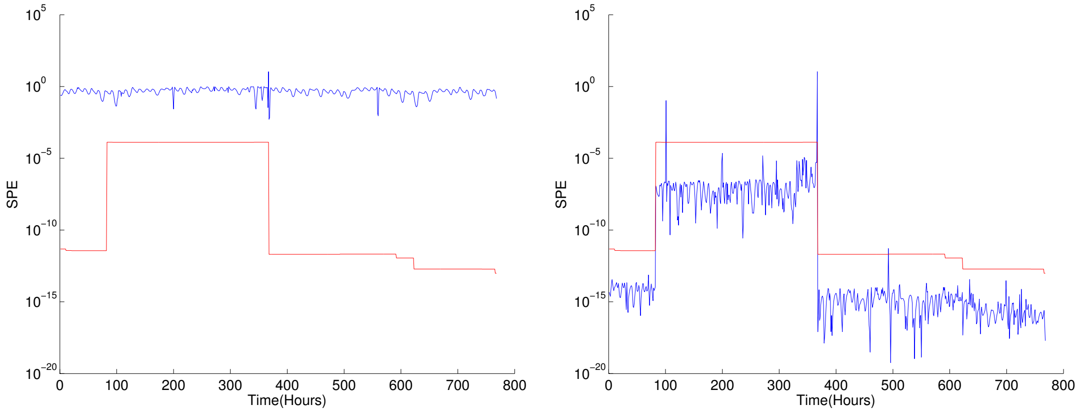
Line	**+80%**	**+50%**	**+10%**	**−10%**	**−50%**	**−80%**	Overall
1	100	100	80	81	100	100	94
4	97	95	81	82	97	99	92
9	100	100	99	99	100	100	100
11	99	98	92	92	98	99	96
22	38	37	37	37	38	54	40
34	98	98	86	89	99	99	95
35	98	98	86	89	99	99	95
36	93	90	80	81	94	97	89
37	93	90	80	81	94	97	89
All	98	98	95	96	98	99	97

Summarizing the results,

- 175104 runs were made on the 24-bus system
- All injected anomalies in 29 of the lines (76%) were detected.
- 97% of all injected anomalies were detected.
- As a result of the injected anomalies 227 OPF (0.13%) runs did not converge. These were counted as detected anomalies since the non-convergence would cause the operator to be made aware of the situation.

Since in the DC model power flow is calculated using voltage angles and reactance, according to the formula in equation 3.2, it is to be expected that modification of reactance would be identified as an anomaly in the majority of instances.

FIg. 6.6 is a comparison of the situation in line 2 when all of the anomalies are found by the algorithm (Fig. 6.6(a)) and the situation when no anomaly is introduced into the system (Fig. 6.6(b)). The changes in Q_α (the red line) are due to the size of the common cause subspace at the time. The size of the subspace varies from 5 to 7 PCs. When all of the anomalies are found, SPE (the blue line) is above Q_α. The three anomalies falsely found are shown in Fig. 6.6(b) and can be identified by the blue line rising above the red line.



(a) Increased reactance on line 2 by 80% finds all anomalies.

(b) No injected anomaly yields 3 false anomalies.

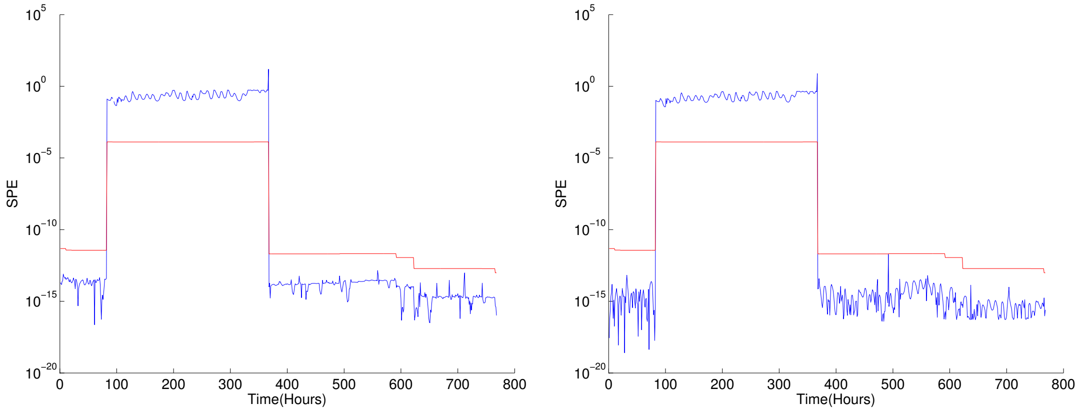
Figure 6.6: Using PCs with Total Cumulative Variance Less Than or Equal to 0.999999

The plots in Fig. 6.7 show the difference in the SPE when the reactance on line 22 is decreased by 10%, Fig. 6.7(a) and the SPE when the reactance on line 22 is increased by 10%, Fig. 6.7(b). The same number of anomalies is found, 285 of 768, but the curves are different since reducing the reactance increases power flow and increasing the reactance decreases power flow.

6.4.5 Case B

The experiment in Case B assumes that a cyberattacker has accessed the database containing the system data model and has modified the line status parameter, indicating that a line is removed from service. The experiment is run for each line for each interval with the results found in Table 6.7.

Overall 94% of the anomalies were found. When line 11 is removed from service the OPF is infeasible. This infeasibility is because line 11 connects bus 7 to bus 8 and is a radial



(a) Decreased reactance on line 22 by 10% finds 285 anomalies.

(b) Increased reactance on line 22 by 10% finds 285 anomalies.

Figure 6.7: Using PCs with Total Cumulative Variance Less Than or Equal to 0.999999

Table 6.7: Anomalies Found When a Line is Removed From Service

Line	Not Converged	Found	Found (%)	Line	Not Converged	Found	Found (%)
1	2	766	100	20	1	767	100
2	14	754	100	21	1	767	100
3	1	767	100	22	0	550	72
4	7	757	99	23	0	768	100
5	0	768	100	24	0	768	100
6	0	768	100	25	5	763	100
7	0	768	100	26	5	763	100
8	13	755	100	27	0	768	100
9	0	768	100	28	0	768	100
10	7	761	100	29	0	768	100
11	DNC			30	11	757	100
12	1	767	100	31	0	768	100
13	8	760	100	32	1	767	100
14	0	768	100	33	1	767	100
15	1	767	100	34	0	767	100
16	0	768	100	35	0	767	100
17	1	767	100	36	1	750	98
18	2	766	100	37	1	750	98
19	0	768	100	38	0	768	100
All	77	27313	94				

line, providing the only source of power to bus 7. Lines 4, 22, 36 and 37 are the other lines where not all of the anomalies are found.

6.4.6 Case C

The simulation in Case C assumes that a cyberattacker has accessed the database containing the system data model and has modified two parameters on a single line: reactance χ_{ij}, generator capacity G_i^{MAX} or line capacity F_{ij}^{MAX}. The experiment is run for each line for each interval with eighteen factors ranging from increased by 90% to decreased by 90%. The choice of which parameters to include and the amount of change applied to the parameters were randomly chosen from the uniform distribution. The results vary according to the parameters changed and the amount of change.

- When reactance is included in the attack and is modified by any factor, 99.57% of the changes are successfully identified as anomalies.
- When generator capacity and line capacity are included together in the attack, the successful identification of the changes as anomalies depends on the change factor.

6.5 Case Study II IEEE 118-bus test system

The IEEE 118-bus test system [80],[85] represents a portion of the American Electric Power System in the Midwestern United States as of December 1962. It is a frequently used larger network for testing transmission system applications. The 118-bus system has 186 transmission lines with 9 transformers; 118 buses of which 99 have loads; and 54 generators, 35 of which are synchronous condensers[1]. Complete data for the system can be found in [86].

The experiments run in this case study are informed by the use of complex network theory [87] to identify line vulnerability. In order to be successful in a cyberattacker's goal of causing damage (i.e. cascading failures) to the power grid it will be necessary for him

[1] A synchronous condenser absorbs or produces reactive power.

to be aware of the relationships among the lines and buses in a large system. A small but strategically placed change to the system could successfully cause a cascading failure.

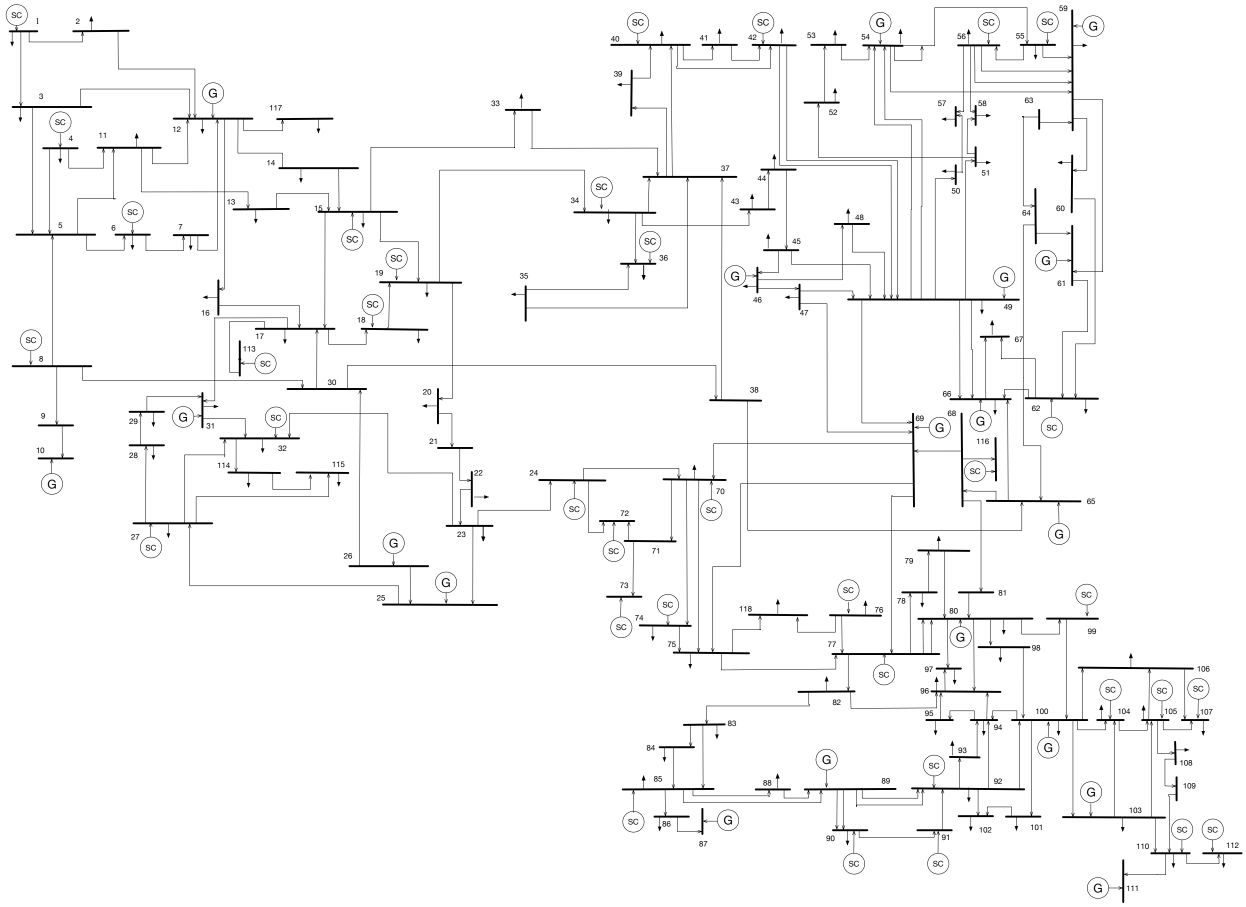


Figure 6.8: One Line Diagram of IEEE 118-bus Test System

6.5.1 Data Descriptions

Because the data for the line rating in [80] is estimated to be 9900 for each line, it is necessary to develop a reasonable line capacity for each line in the 118-bus system. An estimate from [85] uses equation 6.21 when the nominal rating of the buses on either end of the line are the same and equation 6.23 when the ratings are different, indicating a transformer or an ideal line. The formulae were developed by fitting a linear regression model onto the $\log(\chi_{ij}/R_{ij})$ to $\log(F_{ij}^{MAX})$ data of two known systems. The regression model fits the parameters a and k in the function described in equation 6.20 and selects $k = 0.4772$ and

$a = -5.0886$.

$$y = e^a x^k \tag{6.20}$$

$$F_{ij}^{MAX} = \dot{n}e^{-5.0886}\left(\frac{\chi_{ij}}{R_{ij}}\right)^{0.4722} \tag{6.21}$$

where $\dot{n}$ is the common nominal rating for the from bus i and to bus j of the line, χ_{ij} is the reactance of the line, and R_{ij} is the resistance of the line.

The line capacity relates to apparent power as in equation (6.22).

$$g_{ij}^2 + q_{ij}^2 \leq F_{ij}^{MAX^2} \tag{6.22}$$

A reasonable upper bound for $F_{ij}^{MAX^2}$ according to [85] is calculated from equation (6.23).

$$F_{ij}^{MAX^2} = (V_i^{MAX})^2 y_{ij}^2((V_i^{MAX})^2 + (V_j^{MAX})^2 - V_i^{MAX}V_j^{MAX}cos(\theta_i - \theta_j)) \tag{6.23}$$

where V_i^{MAX} and V_j^{MAX} are the maximum voltages at the from bus i and the to bus j, respectively; y_{ij} is the line admittance magnitude; and the voltage angle difference is reasonably estimated to be 15°.

In the 118-Bus system, 35 of the 54 generators are synchronous condensers which only affect reactive power. A new version of the case was located at [86] with updated parameters, including the line flow capacity. This test system has been modernized to improve research using the AC version of the OPF program in MATPOWER. For this reason, the AC version of the OPF program was run in the experiments that follow in this section.

6.5.2 PCA

For illustration purposes, the figures and tables in subsections 6.5.2 and 6.5.3 are based on PCA analysis from the first 672 observations. Section 6.4.2 provides an explanation about the PCA analysis.

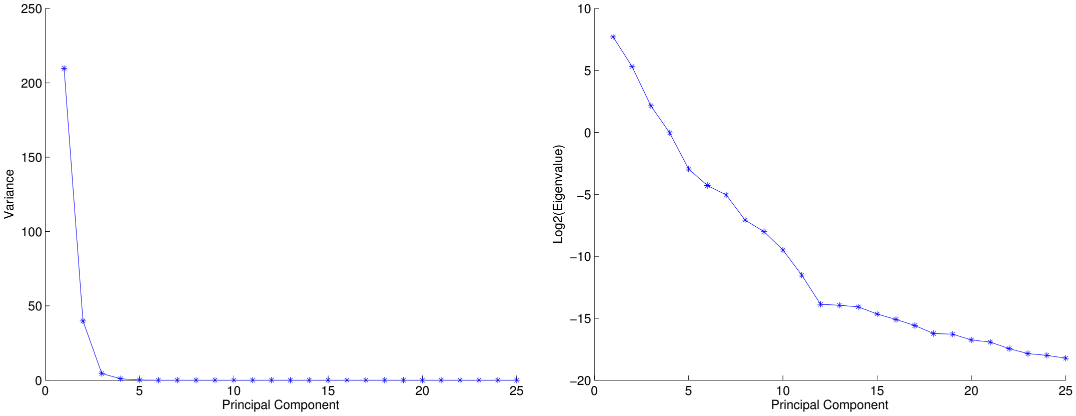
The 672 × 118 matrix, **X**, is formed by extracting the first 672 observations of the simulated voltage angle data. The Gleason-Staelin statistic for the PJM data modeled as a 118-bus system, $\Phi = 0.8333$, indicates that just over 83% of the variables are correlated and it is worthwhile to use PCA to analyze the data. PCA is run on the covariance matrix (**S** from equation 5.5) of mean-centered voltage angle data. The first 10 PC coefficients (**u** vectors) are detailed in Table 6.8.

Table 6.8: First 10 PC coefficients (**U** vectors)

Variable	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}
1	0.137	0.127	0.023	-0.063	-0.143	0.006	0.074	0.009	-0.060	0.012
2	0.132	0.126	0.024	-0.056	-0.151	0.017	0.113	0.021	-0.080	0.018
3	0.133	0.127	0.022	-0.058	-0.137	0.011	0.080	0.003	-0.050	0.011
4	0.117	0.126	0.016	-0.042	-0.104	0.021	0.090	0.002	-0.020	0.009
5	0.115	0.126	0.015	-0.040	-0.100	0.024	0.091	0.002	-0.013	0.005
6	0.126	0.126	0.020	-0.050	-0.131	0.017	0.102	0.002	-0.053	0.021
7	0.127	0.126	0.022	-0.050	-0.141	0.020	0.118	0.003	-0.067	0.027
8	0.093	0.126	0.005	-0.019	-0.049	0.034	0.080	0.005	0.046	-0.015
9	0.093	0.126	0.005	-0.019	-0.049	0.034	0.080	0.005	0.046	-0.015
10	0.093	0.126	0.005	-0.019	-0.049	0.034	0.080	0.004	0.046	-0.015

6.5.3 P and Q_α

The reason for using the Log-Eigenvalue plot becomes more clear when the number of variables is large, 118 in this case. The elbow is much more clear in Fig. 6.9(b) than in Fig. 6.9(a) and indicates that the first 13 PCs should be included in the common cause subspace accounting for a total variance of 0.9999991. These first 13 PCs are considered the common cause subspace ($P = 13$) and the variance associated with them can be attributed to common causes like the size of the lines connecting the buses and other traits inherent to the voltage angles at each of the buses. The remaining PCs are considered the assignable cause subspace and make up a small percentage of the variance in the data.



(a) Cattell's scree test

(b) Log-Eigenvalue plot

Figure 6.9: Variance Accounted for by 13 PCs is 0.9999991.

6.5.4 Line Vulnerability

A power transmission network can be topologically modeled as a graph where buses are considered nodes and transmission lines are considered edges. The field of complex networks [87] provides insight into the relationships among the nodes and edges as well as the robustness of the graph as a whole. The electrical properties of the components involved in the transmission of electric power require enhancements to these topological insights to accurately model the grid. For instance, Steen's definition of the shortest path

Definition 1. *Consider an undirected graph G and two vertices, u, $v \in V(G)$. Let P be a (u, v)-path having minimal weight among all (u, v)-paths in G. The weight of P is known as the (geodesic) distance $d(u, v)$ between u and v. Path P is called a shortest path (u, v)-path, or a geodesic between u and v [87].*

uses geodesic distance to identify the shortest (distance-wise) or most efficient path from one node to another. In many systems modeled as a graph, the shortest path is the one that flow follows, e.g. traffic flow. However, electric power flow is not only determined by the physical node to node relationships, but also by the laws of physics. Electric power flows through all of the available paths, not simply the shortest path.

Line vulnerability in a transmission network has been studied by several researchers in recent years. Since the focus in this research is the transmission line, the research cited focuses only on line vulnerability. Panigrahi considered line reactance as an electrical weight [88] applied to edges to assess power grid vulnerability. In their work, Koc *et al.* proposed a metric to assess network vulnerability against cascading failure due to line overload. The metric, 'electrical node significance' finds the node with the most edges emanating from it and chooses line vulnerability based on the amount of power transmitted through the node [89]. In additional research on network robustness, Koc *et al.* proposed 'effective graph resistance' to identify critical lines [90]. Pepyne applied line loading to grid topology to identify line vulnerability [91]. Cadini, *et al.* introduced the use of 'reliability distances' among network nodes by using the probability of line failure as an edge weight in centrality calculations [92]. Each different electrical weight yields a different set of critical lines.

The technique documented by Koc *et al.* [90] based on effective graph resistance is used in this case study to find line vulnerability where G is defined as a directed graph with N^b nodes and N^e edges.

Effective Resistance. Effective resistance, ρ_{ij}, is the aggregate of reactances between two nodes in the transmission network as explained by equation (6.24).

$$\rho_{ij} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \chi_{ij} \tag{6.24}$$

A second method of calculating effective resistance uses the Laplacian matrix of the graph. The Laplacian matrix with edge weight equal to $1/\chi_{ij}$ in an electric transmission network is the B admittance matrix used in the DC optimal power flow program.

$$L_{ij}^{G} = \begin{cases} \psi_i = \sum_j^{N^b} \frac{1}{\chi_{ij}} & \text{if } i = j \\ -\frac{1}{\chi_{ij}} & \text{if } i \neq j \text{ and } (i,j) \in E \\ 0 & \text{otherwise.} \end{cases} \tag{6.25}$$

In equation (6.25) ψ is considered the strength of node i and E is the set of edges in the graph.

After calculating the Moore Penrose pseudo-inverse of the Laplacian, L^{G+}, the effective resistance between nodes i and j is

$$\rho_{ij} = L_{ii}^{G+} - 2L_{ij}^{G+} + L_{jj}^{G+} \tag{6.26}$$

Effective graph resistance. Effective graph resistance, R_G is the total resistance for all nodes and edges in the network and can be computed in two different ways

- From the aggregate of all effective resistances according to equation (6.27)

$$R_G = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \rho_{ij} \tag{6.27}$$

- From the Laplacian as in equation (6.28) where $\boldsymbol{\mu}$ is the vector of eigenvalues of the Laplacian matrix sorted in order of decreasing size. Note that the last eigenvalue is zero.

$$R_G = N^b \sum_{i=1}^{N^b-1} \frac{1}{\mu_i} \tag{6.28}$$

Vulnerable lines. R_G is calculated for the entire network with all lines in active status and then for the network with (one by one) each line removed. The metric, ΔR_G^l described in equation 6.29 identifies the relative increase in R_G caused by line removal and is used to identify vulnerable lines. An increase in R_G coincides with a decrease in power flow. Less power in a network implies that the network may not have sufficient power available to serve demand. The metric uses the

$$\Delta R_G^l = \frac{R_{G-l} - R_G}{R_G} \tag{6.29}$$

Based on the ΔR_G^l metric, the list of the ten most vulnerable lines in the 118-bus network is documented in Table 6.9.

Table 6.9: 10 most vulnerable lines

Line	From	To	$\Delta R_G^l(\%)$
104	65	68	18.99
96	38	65	15.72
30	23	24	11.00
126	68	81	10.97
54	30	38	10.71
127	80	8	10.62
110	70	71	9.92
37	8	30	8.00
129	82	83	7.54
8	8	5	6.22

6.5.5 Case A

The experiment in Case A assumes that a cyberattacker has accessed the database containing the system data model and has modified the reactance, χ_{ij}. The experiment is run for each of the ten most vulnerable lines for each interval with six different factors, increased by 80%, increased by 50%, increased by 10%, decreased by 10%, decreased by 50% and decreased by 80%.

In a real system, the operator implements the decision variables as calculated from the model with the incorrect reactance. The real power flow results from using the incorrect information from the model. After implementation in the line with the incorrect data, the real power flow is reduced because of a modeled reactance increase and increased because of a modeled reactance decrease.

Table 6.10: Anomalies found on vulnerable lines when reactance is changed

Line	From	To	+80%	+50%	+10%	−10%	−50%	−80%	Overall
104	65	68	100%	100%	100%	100%	100%	100%	100%
96	38	65	100%	100%	100%	100%	100%	100%	100%
30	23	24	100%	100%	100%	100%	100%	100%	100%
126	68	81	100%	99.9%	60.9%	66.2%	100%	100%	87.80%
54	30	38	100%	100%	100%	100%	100%	100%	100%
127	80	8	100%	100%	75.9%	78.4%	100%	100%	92.38%
110	70	71	100%	100%	100%	100%	100%	100%	100%
37	8	30	96.7%	92.6%	65.4%	67.6%	94.8%	98.7%	85.96%
129	82	83	100%	100%	100%	100%	100%	100%	100%
8	8	5	96.4%	88.3%	43.9%	42.7%	95.7%	100%	61.76%

Among the ten lines deemed the most vulnerable to network collapse, 94% of the inserted anomalies are found by the algorithm with 10 false anomalies in the 768 hours.

To determine the effect of the not finding the anomaly, these steps are followed:

1. OPF is run with the reactance changed by the attacker and optimal decision variables (voltage angles, voltage magnitudes real power injections and reactive power injections) are calculated.

2. These decision variables are substituted into the Power Flow (PF) program with the correct reactance.

3. Real power flow is calculated according to equation (6.11) for each line in the network.

4. Reactive power flow is calculated according to equation (6.12).

5. Apparent power flow is calculated according to equation (6.30). Apparent power is used for line use calculations in the AC model since the capacity limits on the line are based on apparent power.

$$f_{ij}^{A} = \sqrt{{f_{ij}^{P}}^{2} + q_{ij}^{2}} \tag{6.30}$$

where f_{ij}^{P} is the real power flow between buses i and j and q_{ij} is the reactive power flow between buses i and j.

6. Line usage is calculated according to equation (6.31) for each line in the network. This is the result that the operator will see. With no overloaded lines, the operator will implement the erroneous power flows.

$$y_{ij} = \frac{f_{ij}^{A}}{F_{ij}^{MAX}} \tag{6.31}$$

Line 126. When the reactance on line 126 is increased by 50%, documented in Fig. 6.10 there is one hour ($t = 72$) when the inserted anomaly is not found.

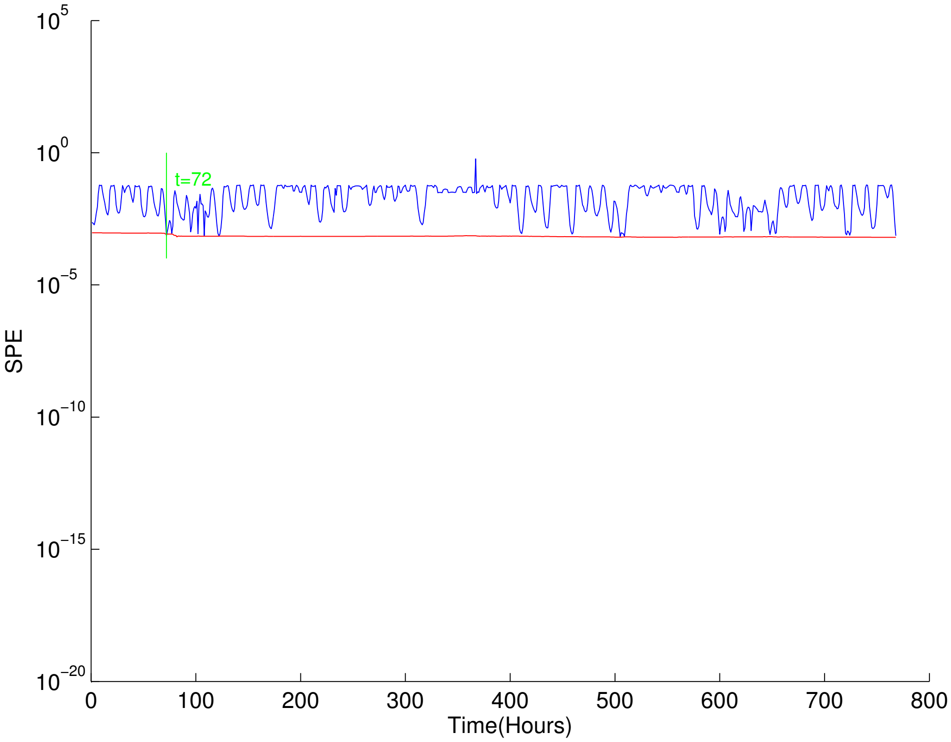
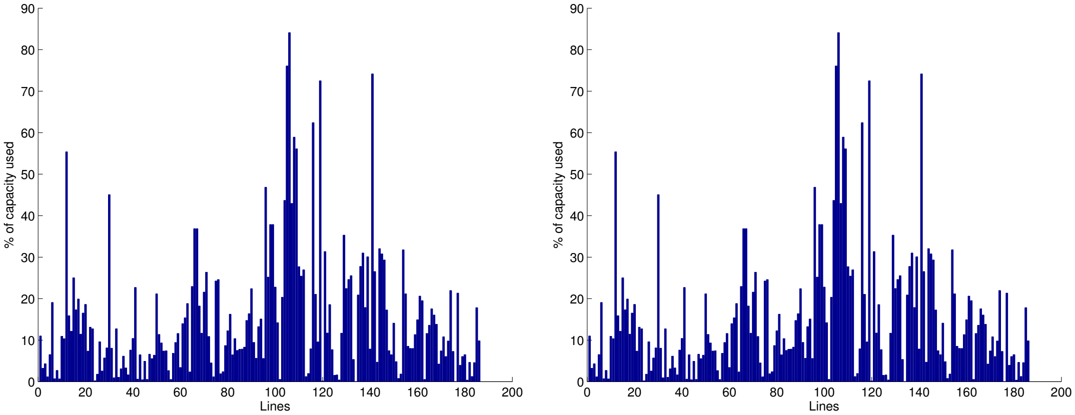


Figure 6.10: Result of Case A when Reactance on Line 126 is Increased by 50%

An examination of the results finds that in line 126 at time $t = 72$ when the reactance increased by 50%, the resulting power flow changes slightly from the power flow without the anomaly, but does not cause any overloads to the system, and therefore will not cause a cascade of failures. The apparent power flow decreased from 37.5366 (calculated with OPF) to 11.5565 (implemented in the real system using PF). The percent of line use in the entire network is presented in Fig. 6.11.



(a) Case A with no anomaly inserted

(b) Case A with reactance on line 126 increased by 50%

Figure 6.11: Line Use Comparison in Line 126

Line 8. Overall, anomaly identification success in line 8 was 78% as seen in Fig. 6.12.

After following the steps documented earlier in this section, the results in Fig. 6.13 shows that the line use changes slightly for all lines with an anomaly in line 8. Reducing the reactance by 10% increased the real power flow in line 8. For instance at time $t = 23$ the real power flow without an anomaly introduced was -0.7418 and implemented with the reactance decreased by 10% it was -0.4891. The results in table 6.11 show the changes to line 8 at time $t = 23$ for the reactance change. The power flow changes are not enough to cause damage to the network. Note that on line 141, both with and without the anomaly in line 8, the line use is 100%.

Line used in all

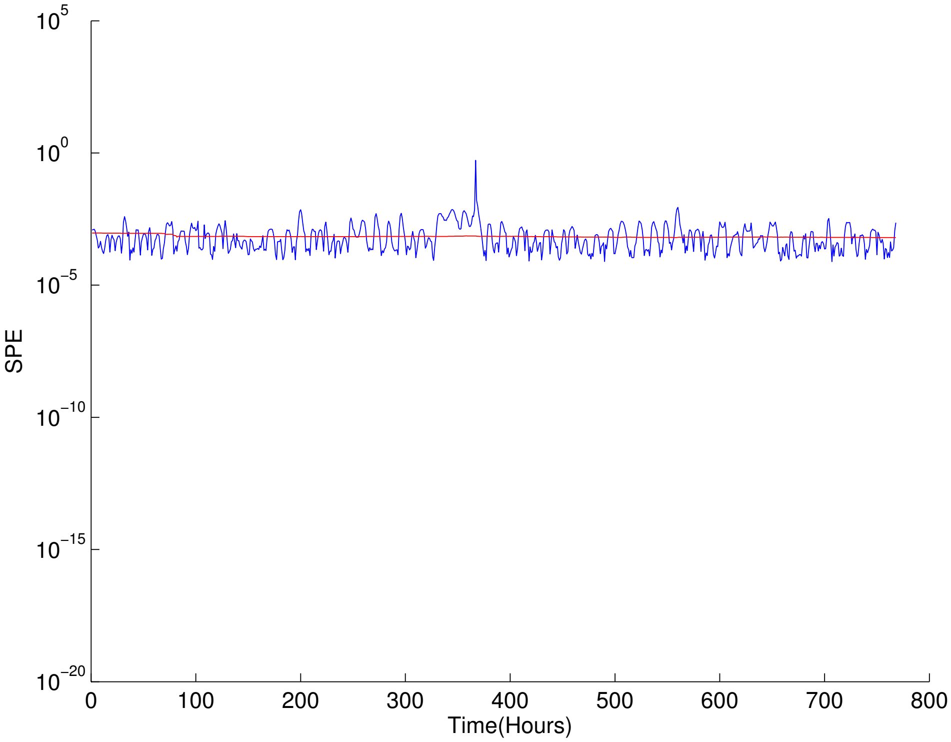
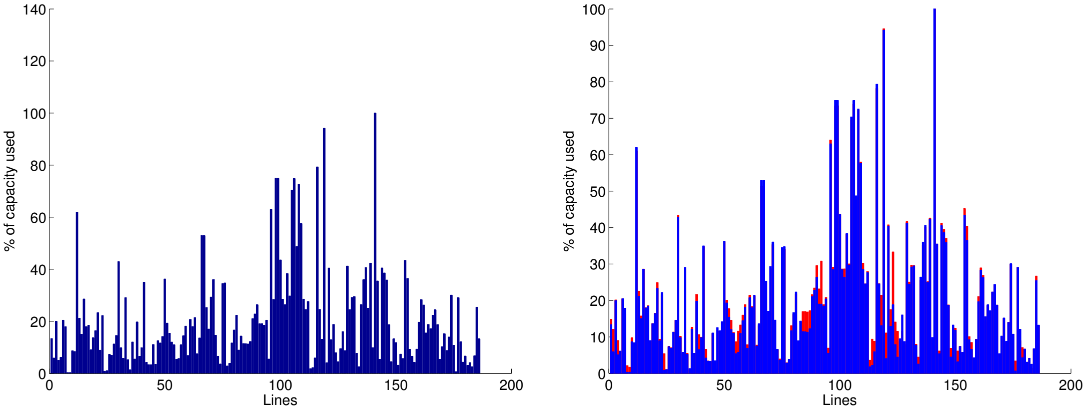


Figure 6.12: Result of Case A when Reactance on Line 8 is Decreased by 10%



(a) No Anomaly Inserted

(b) Reactance in Line 8 Decreased by 10%

Figure 6.13: Line Use Comparison for Line 8

Table 6.11: Changes in Line 8 at $t = 23$

Situation	χ_{ij}	f_{ij}^{P}	q_{ij}	s_{ij}	% line used
Without anomaly	0.0267	-0.7418	-3.7205	3.7937	.36
Implemented	0.0240	-0.4891	23.4970	23.5021	2.14

Line 37. Overall, anomaly identification success in line 37 was 86%. When the reactance was increased by 80%, 96.7% of the anomalies were discovered by the algorithm. In the remainder of this analysis the focus will be on time interval $t = 63$.

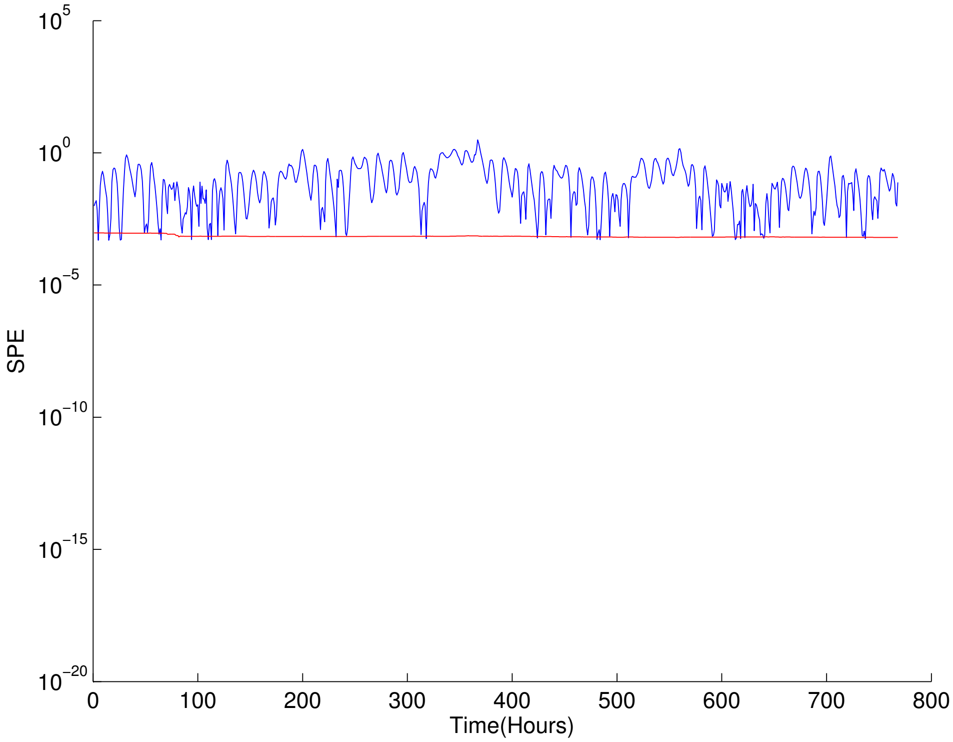
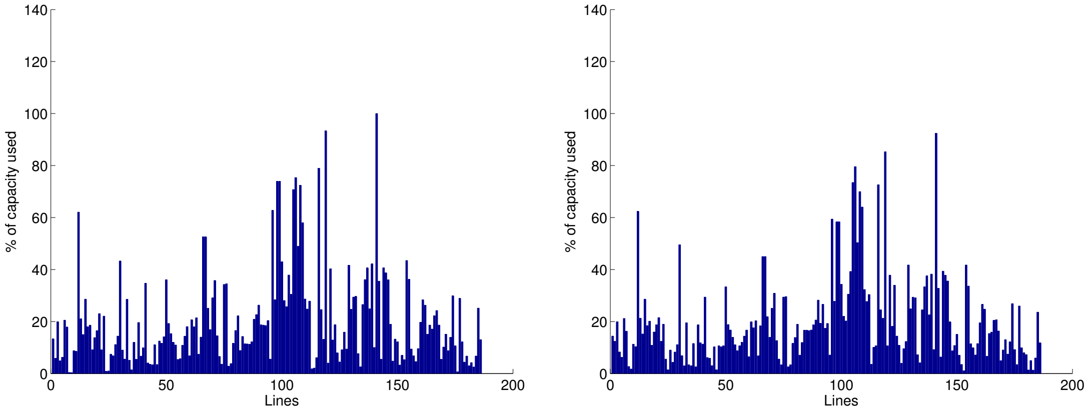


Figure 6.14: Result of Case A when Reactance on Line 37 is Increased by 80%

After following the steps documented earlier in this section, the results in Fig. 6.15 show that the line use for line 37 changes slightly because of the anomaly. Because the reactance was increased the power flow was reduced from 20.5196 to 20.3072 at $t = 63$ on line 37 causing adjustments throughout the network. The power flow changes are not enough to cause damage to the network.

6.5.6 Case B

The experiment in Case B assumes that a cyberattacker has accessed the database containing the system data model and has modified the line status parameter, indicating that a line is removed from service. The experiment is run for each line for each interval and in all of the most vulnerable lines 100% of the anomalies are found.



(a) No Anomaly Inserted

(b) Reactance on Line 37 Increased by 80%

Figure 6.15: Line use Comparison in Line 37 with One Change in the Line

6.5.7 Case C

In Case C the cyberattacker has accessed the database containing the system data model and has modified two parameters each by one of six factors. The algorithm randomly chooses two of three parameters to change–reactance, resistance and MVA rating. The amount of change applied to the parameters is also randomly chosen. The results in the most vulnerable lines, shown in Table 6.12, are successful for 64% of the lines. In any line where reactance is changed the result is 100% successful.

Table 6.12: Case C results

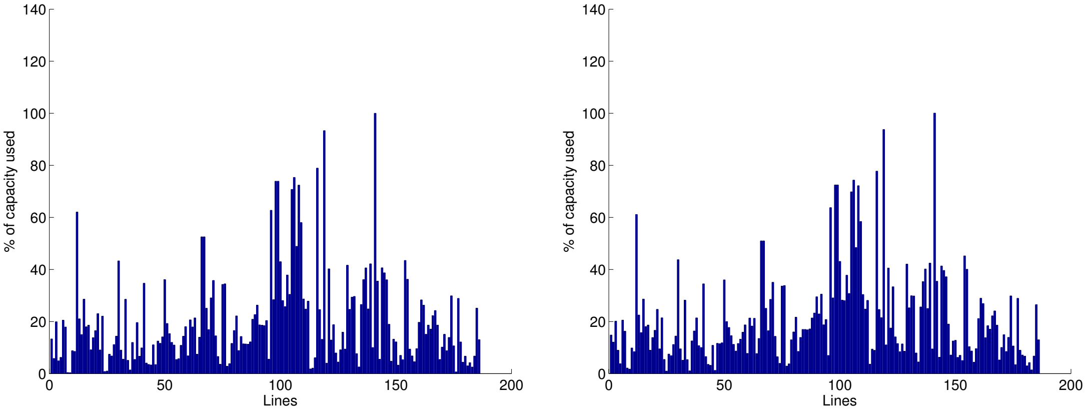
Pipeline	**Parameter 1**	**Changed by**	**Parameter 2**	**Changed by**	**Result (%)**
104	reactance	-30%	resistance	+10%	100
96	resistance	-80%	reactance	-60%	100
30	resistance	+40%	line capacity	+70%	72
126	reactance	+70%	resistance	-90%	100
54	resistance	+40%	reactance	-10%	100
127	line capacity	+30%	resistance	+70%	1
110	line capacity	-40%	reactance	+70%	100
37	line capacity	-80%	resistance	-80%	10
129	resistance	-50%	line capacity	+40%	63
8	line capacity	+70%	resistance	-10%	10

Lines 8 and 127 both have transformers installed on the sending end of the line and line 8 has a synchronous condenser installed at the sending end of the line, bus 8. The synchronous condenser regulates voltage by increasing or decreasing the reactive power.

To better understand the effect of the cyberattack, line 37 will be studied further. Line 37 transmits power from bus 8 to bus 30. The synchronous condenser at bus 8 regulates the voltage on the line by increasing or decreasing the reactive power.

In line 37 the resistance was decreased by 80% and the line capacity was decreased by 80%. At time $t = 7$ Fig 6.16 shows there is no visible change to the line use. Table 6.13 provides the actual values of the pertinent parameters and resulting variables. Notice how the reactive power q_{ij} is decreased.

Although the percentage of line used in line 8 increases significantly, no lines are overloaded.



(a) No Anomaly Inserted

(b) Line 37 Resistance and Line Capacity Decreased by 80%

Figure 6.16: Line use Comparison in Line 37 with Two Changes in a Single Line

Table 6.13: Changes in Line 37 at $t = 7$

Situation	R_{ij}	F_{ij}^{MAX}	f_{ij}^{P}	q_{ij}	f_{ij}^{A}	% line used
Without anomaly	0.00431	580	-1.6955	-4.1861	4.5164	0.41
Implemented	0.00344	116	-1.4562	-12.1593	23.5806	2.15

6.6 Discussion

In both of the transmission networks tested in the electric power industry the algorithm detected more than 90% of the simulated anomalies.

The 24-bus system was modeled using the DC model, a simplification of the AC model often used in planning and some operational situations. The AC model is used in practice in critical operational situations, so its use in the 118-bus system is an important test of the algorithm.

Complex network theory was used in the 118-bus system to identify the most vulnerable transmission lines. These ten lines need to be carefully monitored by the operator. Undetected modification to parameters in the database related to these lines are more likely to result in cascading failure.

Chapter 7

Experimental Cyberattack to Gas Transmission Systems

Natural gas is a nonrenewable fossil fuel composed of hydrocarbon gases, primarily methane. As an energy source, it often fuels heating, cooking, electricity generation, and vehicles. Natural gas is typically found close to petroleum and the two are extracted at the same time in many instances. Before natural gas can be used as a fuel it must go through a treatment process to remove impurities, including water, to meet specifications defined by the market demand.

As seen in Fig. 7.1 the process of getting natural gas to its intended destinations begins with extraction from land wells, offshore wells, liquid natural gas (LNG) tankers, and storage facilities. This gas is gathered in small (gathering) lines and transported to a treatment plant. From the treatment plant the more refined natural gas (85% methane) is transmitted in large pipes (transmission) across thousands of miles to where it is needed. At city gates, the gas is odorized and prepared for distribution through smaller pipelines (distribution) to the end users. Power plants and large industrial customers may be fed directly from the transmission system.

7.1 Gas Transmission System

The transmission system is a complex network of nodes and their connecting pipelines operating under a wide range of pressures. Load nodes are the network points where load is known–this may be customer demand (-) or a supply of gas to the network (+) either from storage, a specific source or from another network. Transit nodes have zero load but represent a change in network topology. Each pipeline can be described as either passive or active. In passive pipelines the gas flows freely according to its properties and the environment. In

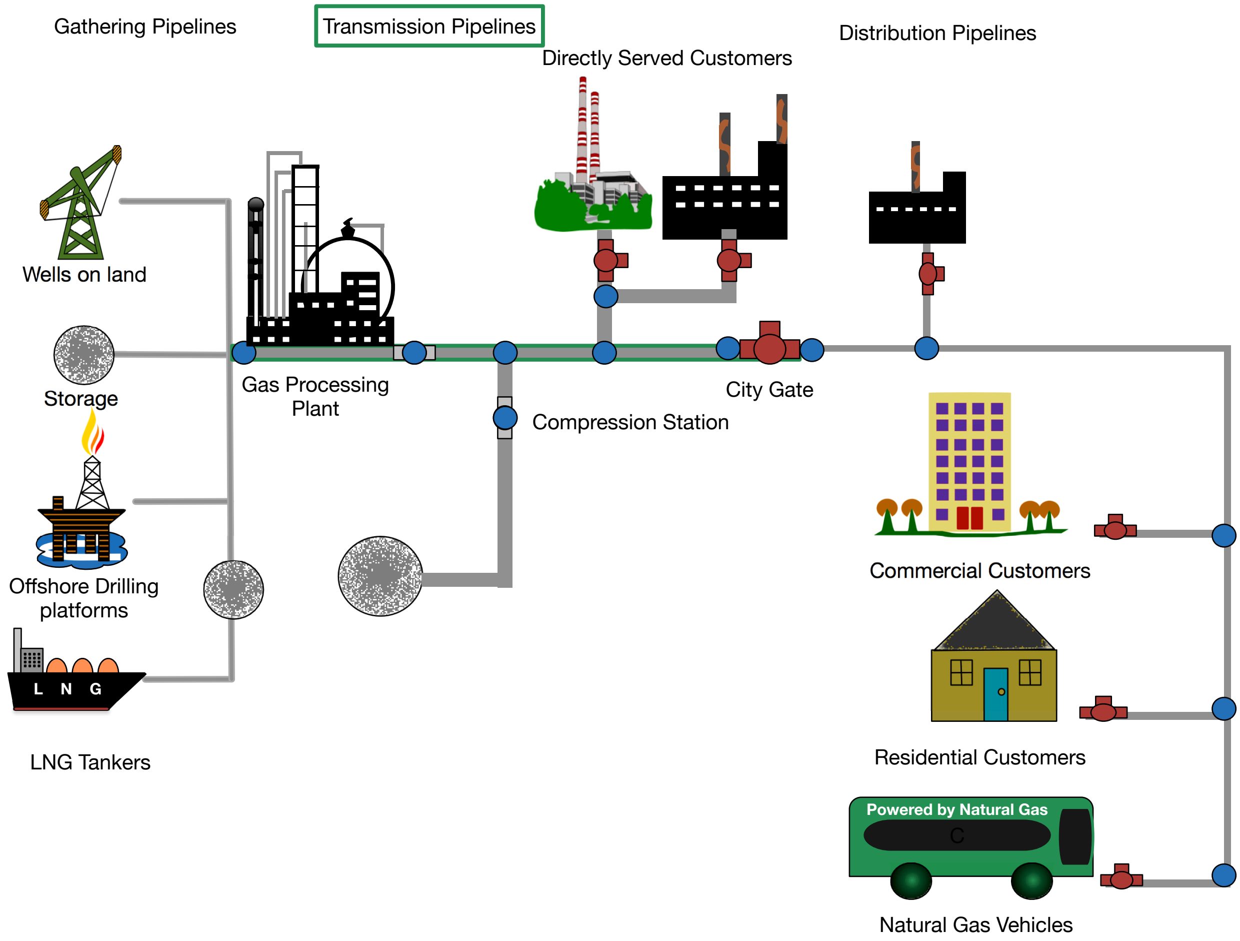


Figure 7.1: Real Natural Gas System

active pipelines equipment either increases (compressor) or decreases (regulator) the pressure at which the gas travels. Transmission pipelines are connected to the nodes in the network and provide for the flow of gas through the network from treatment plants to city gates.

The role of the operator in natural gas networks is changing as more technology is integrated into the process, but slowly. In most networks in the United States, approximations, decompositions, and engineering judgement play a major role in decisions about interstate natural gas flow. For the purpose of this research, the application software used to compute the multivariate data to be analyzed for anomaly detection is the OGF program. Below an OGF model is presented and discussed.

As documented earlier in this paper, no known cyberattacks have been successfully carried out on the U. S. natural gas transmission network so it is difficult to identify a *typical* attack. Line characteristics are changed in the cases described below because this data is accessible, affects the calculation of optimal flows, and changes to line characteristics may not be detected by a system operator.

7.2 Optimal Gas Flow

The flow of natural gas differs significantly from the flow of electricity. Natural gas flows at approximately 30 to 40 miles per hour according to the laws of thermodynamics. Natural gas can be stored in underground facilities until it is needed. Because of these significant differences, the flow of natural gas is not monitored as strenuously as the flow of electricity. The typical natural gas pipeline operator uses the results of sensors with periodic runs of an optimal gas flow program for decision making.

Natural gas transmission pipeline network presents a difficult problem to solve mathematically. Both the objective function and constraint equations are nonlinear and the solution space is non-convex. Researchers have used meta-heuristics such as simulated annealing and GA, dynamic programming, various relaxation methods, and combinations of all of the above [93] to approximate solutions.

In order to make the solution more tractable, in this research these assumptions limit the scope of the problem.

- The problem is in steady-state and flow variables are time-independent.
- Gas flow is considered isothermal. Any heat transfer between the gas and the environment remains constant.
- The flows, f^G_{ij} are unrestricted in sign. If $f^G_{ij} < 0$ then the flow $-f^G_{ij}$ goes from node j to node i.

- Pipelines are strictly horizontal.
- The system is deterministic. All parameters are known in advance.

As in OPF, the solution vector of the OGF module changes over time based on the natural variability in the module's input data.

Researchers in natural gas transmission are concerned with both the mass flow rate and the volumetric flow rate of the gas. Understanding the difference and the relationship between the two is important to the remainder of this chapter.

Volumetric flow of natural gas is defined as "the volume (amount of space occupied) of natural gas that flows through a pipe per unit of time." Volumetric flow is measured in M^3/second or ml/second or ft^3/hour. Mass flow rate of natural gas is defined as "the amount of natural gas that flows through a pipe per unit of time." The two measures are related by equation 7.1 [94].

$$f_{ij}^{G} = \delta \dot{v} \tag{7.1}$$

where δ is the density of natural gas and $\dot{v}$ is the volumetric flow rate of natural gas.

Throughout this research natural gas flow implies mass flow. Any reference to volumetric flow will be specifically noted.

7.2.1 OGF Model

With the de-regulation of natural gas, utility companies that managed the flow of gas through the entire network have been split into several companies with transmission separated from supply and distribution. Since de-regulation much of the literature approximates the optimal flow through the transmission network by minimizing the cost of gas used to power compressors. Important constraints in the model represent the relationships between gas flow in pipelines and pressure at network nodes. The system is modeled as a graph $G = (N, P)$ of nodes and connecting pipelines. The set of nodes is partitioned into N^s

(supply nodes), N^d (demand nodes), and N^t (transshipment nodes). The set of pipelines is divided into P^p (passive pipelines) and P^a (active pipelines with a compressor). A passive pipeline has positive length and connects two distant nodes. An active pipeline has length zero. A device on an active pipeline increases the pressure (compressor) as the gas passes through the device. Let A be the incidence matrix of the graph. The columns of A are partitioned into $A = (A^p, A^a)$. The entire optimization problem is stated below as defined in [95].

$$\underset{f^G,s,p,w}{\text{minimize}} \quad \beta \sum_{(i,j)\in A^a} \frac{1}{0.9\eta_{ij}} w_{ij} \tag{7.2}$$

subject to

$$\sum_{j|(i,j)\in A} f_{ij}^G = \sum_{j|(j,i)\in A} f_{ji}^G + s_i \quad \forall i \in N^n \tag{7.3a}$$

$$sign(f_{ij}^G) f_{ij}^{G^2} = C_{ij}^G (p_i^2 - p_j^2) \quad \forall (i,j) \in A^p \tag{7.3b}$$

$$sign(f_{ij}^G) f_{ij}^{G^2} >= C_{ij}^G (p_i^2 - p_j^2) \quad \forall (i,j) \in A^a \tag{7.3c}$$

$$w_{ij} = \gamma_1 f_{ij}^G \left(\left(\frac{p_j}{p_i} \right)^{\gamma_2} - 1 \right) \quad \forall (i,j) \in A^a \tag{7.3d}$$

$$S_i^{\text{MIN}} \leq s_i \leq S_i^{\text{MAX}} \quad \forall i \in N^n \tag{7.3e}$$

$$P_i^{\text{MIN}} \leq p_i \leq P_i^{\text{MAX}} \quad \forall i \in N^n \tag{7.3f}$$

$$f_{ij}^G \geq 0 \quad \forall (i,j) \in A^a \tag{7.3g}$$

$$\frac{p_j}{p_i} \leq 1.6 \quad \forall (i,j) \in A^a \tag{7.3h}$$

$$w_{ij} \leq W_{ij}^{\text{MAX}} \quad \forall (i,j) \in A^a \tag{7.3i}$$

Decision Variables.

$$\mathbf{x} = \begin{bmatrix} \mathbf{f^G} \\ \mathbf{s} \\ \mathbf{p} \\ \mathbf{w} \end{bmatrix}$$

where $\mathbf{f^G}$ is the $N^f \times 1$ vector of mass gas flow,

$\mathbf{s}$ is the $N^n \times 1$ vector of sources, demand is modeled as a negative source,

$\mathbf{p}$ is the $N^n \times 1$ vector of nodal pressures,

$\mathbf{w}$ is the $N^a \times 1$ vector of energy used by a compressor station.

Objective Function. The objective function minimizes the cost of energy used at the compressor stations in the network. This objective function is realistic in the de-regulated era of natural gas transmission where the operator is concerned with expeditiously taking the gas from suppliers (within limits) and delivering the gas as demanded by customer contract at a minimized transmission cost.

$$\underset{f^G,s,p,w}{\text{minimize}} \quad \beta \sum_{(i,j)\in A^a} \frac{1}{0.9\eta_{ij}} w_{ij}$$

where $\boldsymbol{\eta}$ is the $N^a \times 1$ vector of thermic efficacies of compressor stations,

β is the energy price in K-euro/kW,

and all other variables and parameters are as defined above.

Constraints. Several types of constraints ensure a feasible solution to the OGF.

Nodal balance equations ensure that flows in and out of a node and supply and demand at the node are balanced.

$$\sum_{j|(i,j)\in A} f^G_{ij} = \sum_{j|(j,i)\in A} f^G_{ji} + s_i \quad \forall i \in N^n$$

where $\mathbf{s}$ is the vector of gas supplies(+) and demands(-) at each network node, and all other variables and parameters are as defined above.

The pressure and flow relationship in passive pipelines (equation (7.4)) differs from the pressure and flow relationship in active pipelines (equation (7.5)).

$$sign(f_{ij}^{G}) f_{ij}^{G^2} = C_{ij}^{G}(p_i^2 - p_j^2) \quad \forall (i,j) \in A^p \tag{7.4}$$

$$sign(f_{ij}^{G}) f_{ij}^{G^2} >= C_{ij}^{G}(p_i^2 - p_j^2) \quad \forall (i,j) \in A^a \tag{7.5}$$

$$f_{ij}^{G} \geq 0 \quad \forall (i,j) \in A^a$$

where $\mathbf{C^G}$ is the vector of constants particular to a pipeline beginning at node, i and ending at node, j, and all other variables and parameters are defined above. $\mathbf{C^G}$ is defined by equation (7.6).

$$C_{ij}^{G} = 96.074830 \times 10^{-15} \frac{D_{ij}^5}{\lambda_{ij} Z T L_{ij} \delta} \tag{7.6}$$

where λ_{ij}, defined by equation (7.7) is the friction factor for fully turbulent conditions,

$$\frac{1}{\lambda_{ij}} = \left[2 \log \left(\frac{3.7 D_{ij}}{\epsilon} \right) \right]^2, \tag{7.7}$$

L_{ij}^{G} = length of the pipe in km

D_{ij}=interior diameter of the pipe in mm

T = gas temperature in K (constant at 281.15)

ϵ = absolute rugosity (roughness) of pipe in mm (constant at 0.05)

δ = density of gas relative to air (constant at 0.6106)

Z = gas compressibility factor (constant at 0.8)

The power used by a compressor, w_{ij} as defined by equation (7.8), relates to specific characteristics of the installed compressor. The pressures at the connecting nodes of the active pipeline enforce the energy used by the compressor.

$$w_{ij} = \gamma_1 f_{ij}^G \left(\left(\frac{p_j}{p_i} \right)^{\gamma_2} - 1 \right) \quad \forall (i,j) \in A_a \tag{7.8}$$

$$w_{ij} \leq W_{ij}^{\mathrm{MAX}} \quad \forall (i,j) \in A_a$$

Supply and demand is governed by contractual agreements between the transmission company and suppliers and customers. Unlike in electric power transmission, supply and take amounts typically have a 10% range in either direction.

$$S_i^{\mathrm{MIN}} \leq s_i \leq S_i^{\mathrm{MAX}} \quad \forall i \in N^n$$

Pressures at each node must conform to pressure standards for the node.

$$P_i^{\mathrm{MIN}} \leq p_i \leq P_i^{\mathrm{MAX}} \quad \forall i \in N^n$$

And a pressure ratio of the pressure of gas leaving the compressor to the pressure of gas entering the compressor must remain below 1.6 (a standard industry constant).

$$\frac{p_j}{p_i} \leq 1.6 \quad \forall (i,j) \in A_a$$

7.2.2 OGF Solution

The problem as defined by equations (7.2) and (7.3) is divided into two subproblems. The first problem as modeled in equations (7.9) and (7.10) relaxes the pressure constraints and eliminates the compressors, solving for all flows and supplies under these conditions. The solution to this problem ($\mathbf{f^G}$ and $\mathbf{s}$) is a good starting point for the full problem as defined in equations (7.2) and (7.3).

$$\underset{f^G, s}{\text{minimize}} \sum_{(i,j)\in A} \frac{|f^G_{ij}| {f^G_{ij}}^2}{3C^G_{ij}} \tag{7.9}$$

subject to

$$\sum_{j|(i,j)\in A} f^G_{ij} = \sum_{j|(j,i)\in A} f^G_{ji} + s_i \quad \forall i \in N^n \tag{7.10a}$$

$$S^{\text{MIN}}_i \leq s_i \leq S^{\text{MAX}}_i \quad \forall i \in N^n \tag{7.10b}$$

In this research Matlab's function, *fmincon*, is used to solve both subproblems of the OGF problem.

7.3 Gas Flows

The general flow equation for steady-state gas flow at standard conditions for horizontal pipe is found in equation (7.11).

$$Q_n = C\frac{T_n}{p_n}\sqrt{\frac{(p_1^2 - p_2^2)D^5}{fSL^GTZ}} \tag{7.11}$$

where

$$C = \sqrt{\frac{\pi^2 R_{air}}{64}} \text{ is a constant} \tag{7.12}$$

$T_n = 288$K—the standard temperature,

$p_n \approx 0.1$MPa—the standard pressure,

p_1 is the pressure at the inlet node of the pipeline,

p_2 is the pressure at the outlet node of the pipeline,

D is the inner pipe diameter,

f is the friction factor,

S is the specific gravity of the gas,

L^G is the length of the pipeline,

T is the gas temperature in K,

Z is the gas compressibility factor, and

R is a constant for air.

Three simplified flow equations have been developed for use in the gas industry to model gas flow in transmission pipelines, the Weymouth equation (developed in 1912), the Panhandle A equation (developed in 1940), and the Panhandle B equation (developed in 1956). Each of these equations is based on the general gas equation (7.11) with a different simplification to model the friction in the pipeline. The equations were developed to simplify an analytical solution to flow calculations [96].

The flow equation (7.4) used in this research was derived for use in fully turbulent flow conditions occurring in large pipelines where the gas is transmitted under high pressure and is known as the Weymouth equation [97]. All of the flow equations in use today can be simplified to equation (7.13), effectively the relationship between the mass flow rate of natural gas, f_{ij}^G through the pipeline from i to j and the pressure drop $p_i^2 - p_j^2$ from beginning to end of the pipeline. The constant C_{ij}^G varies according to the gas and pipeline properties (equations 7.6 and 7.7).

$$f_{ij}^G = \sqrt{C_{ij}^G (p_i^2 - p_j^2)} \tag{7.13}$$

Flows are calculated using a MATLAB program modified from [98] that uses the Newton-Raphson method of solving a nonlinear system of equations.

7.4 Case Study: Belgium Gas Transmission Network

The real system used for this experiment is the Belgium Power Flow network as in was documented in the late 1990's. The slightly modified (duplicate lines were removed) Belgian network [99, 100] is a tree shape with 20 nodes and 19 pipelines with 2 compressor stations. The model is presented in Fig. 7.2. The data used by the OGF program is described in section 7.4.1.

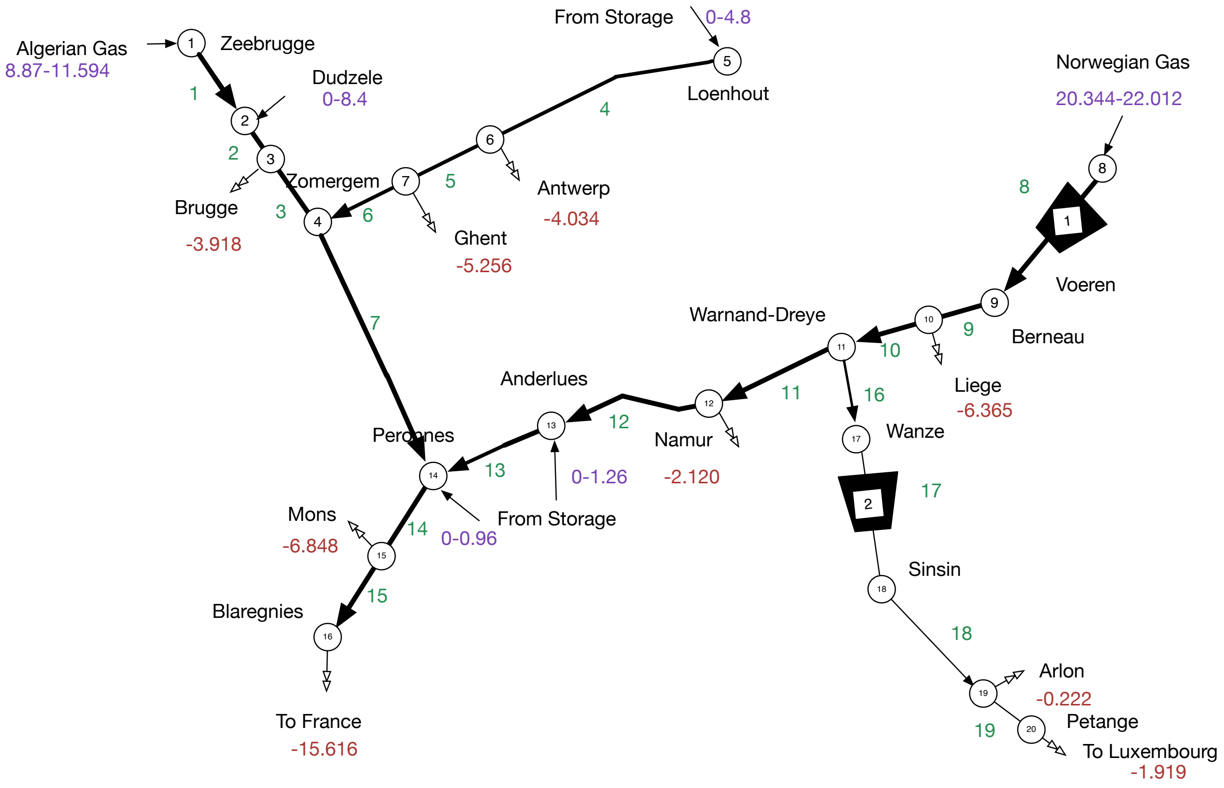


Figure 7.2: Belgium Gas Transmission Network

7.4.1 Data Descriptions

Miscellaneous Constants

Isentropic exponent, $k = 1.287$

Gas Compressibility factor (dimensionless), $Z = 0.8$

Gas Constant, $R^G = 85.2$

Specific gravity, $S_g = 0.6248$

Gas Temperature (K), $T = 281.15$

Gas Density relative to air (dimensionless), $\delta = 0.6106$

Absolute rugosity of pipe (mm), $e = 0.05$

In addition, the following assumptions were made about the two compressors.

- Compressor 1 at Voeren is a turbo compressor so $\eta_1 = .75$ and $\gamma_1 = 0.167$
- Compressor 2 at Sinsin is a moto compressor with $\eta_2 = .80$ and $\gamma_1 = 0.157$.

Table 7.1: Pipeline Descriptions

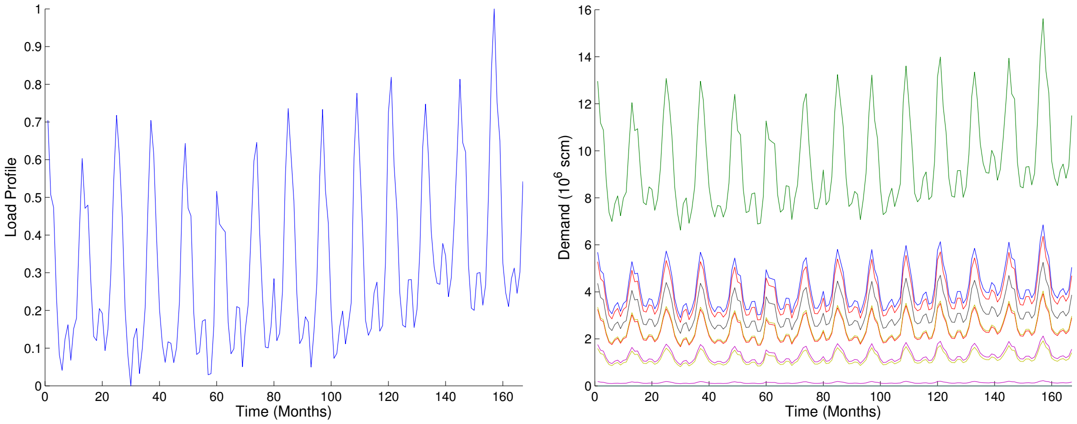
Pipeline	**From**	**To**	**D** **(mm)**	L^G **(km)**	C^G	**Type**	W^{MAX}	γ_1	γ_2	η
1	1	2	890.0	4.0	9.0703	PASSIVE	N/A	N/A	N/A	N/A
2	2	3	890.0	6.0	6.0469	PASSIVE	N/A	N/A	N/A	N/A
3	3	4	890.0	26.0	1.3954	PASSIVE	N/A	N/A	N/A	N/A
4	5	6	590.1	43.0	0.1003	PASSIVE	N/A	N/A	N/A	N/A
5	6	7	590.1	29.0	0.1487	PASSIVE	N/A	N/A	N/A	N/A
6	7	4	590.1	19.0	0.2269	PASSIVE	N/A	N/A	N/A	N/A
7	4	14	890.0	55.0	0.6597	PASSIVE	N/A	N/A	N/A	N/A
8	8	9	890.0	5.0	7.2562	ACTIVE	20888	0.1670	0.2360	0.7500
9	9	10	890.0	20.0	1.8140	PASSIVE	N/A	N/A	N/A	N/A
10	10	11	890.0	25.0	1.4512	PASSIVE	N/A	N/A	N/A	N/A
11	11	12	890.0	42.0	0.8638	PASSIVE	N/A	N/A	N/A	N/A
12	12	13	890.0	40.0	0.9070	PASSIVE	N/A	N/A	N/A	N/A
13	13	14	890.0	5.0	7.2562	PASSIVE	N/A	N/A	N/A	N/A
14	14	15	890.0	10.0	3.6281	PASSIVE	N/A	N/A	N/A	N/A
15	15	16	890.0	25.0	1.4512	PASSIVE	N/A	N/A	N/A	N/A
16	11	17	395.5	10.5	0.0514	PASSIVE	N/A	N/A	N/A	N/A
17	17	18	315.5	26.0	0.0064	ACTIVE	3356	0.1570	0.2360	0.8000
18	18	19	315.5	98.0	0.0017	PASSIVE	N/A	N/A	N/A	N/A
19	19	20	315.5	6.0	0.0278	PASSIVE	N/A	N/A	N/A	N/A

A Monte Carlo simulation of the gas network is coded to generate nodal pressure data since the actual state variable results are difficult to obtain. In the simulation, monthly demand data from the U.S. Energy Information Administration (EIA) [101] beginning in January 2001 and ending in November 2014, 167 months of data, is used. The data is standardized and scaled so that the value 1.0 corresponds to the peak demand in the data. Supplies (that were not demands) were left unchanged. Plots of the profile data and the simulated demand are shown in Fig. 7.3.

Table 7.2: Node Descriptions

Node	**Type**	**Minimum Supply** (10^6 scm)	**Maximum Supply** (10^6 scm)	**Minimum Pressure** (bar[a])	**Maximum Pressure** (bar)	**Demand** (10^6 scm)
1	Supply	8.870	11.594	0	77.0	0.000
2	Supply	0.000	8.400	0	77.0	0.000
3	Demand	-Inf	-3.918	30	80.0	3.918
4	Tranship	0.000	0.000	0	80.0	0.000
5	Supply	0.000	4.800	0	77.0	0.000
6	Demand	-Inf	-4.034	30	80.0	4.034
7	Demand	-Inf	-5.256	30	80.0	5.256
8	Supply	20.344	22.012	50	66.2	0.000
9	Tranship	0.000	0.000	0	66.2	0.000
10	Demand	-Inf	-6.365	30	66.2	6.365
11	Tranship	0.000	0.000	0	66.2	0.000
12	Demand	-Inf	-2.120	0	66.2	2.120
13	Supply	0.000	1.200	0	66.2	0.000
14	Supply	0.000	0.960	0	66.2	0.000
15	Demand	-Inf	-6.848	0	66.2	6.848
16	Demand	-Inf	-15.616	50	66.2	15.616
17	Tranship	0.000	0.000	0	66.2	0.000
18	Tranship	0.000	0.000	0	63.0	0.000
19	Demand	-Inf	-0.222	0	66.2	0.222
20	Demand	-Inf	-1.919	25	66.2	1.919

[a] Bar is the atmospheric pressure at sea level in mm



(a) Profile

(b) Demand

Figure 7.3: Demand Profile and Simulated Data

The simulated demand data is used as input and the optimal nodal pressure at each node for each month is computed by the OGF software module (modified from MATPOWER OPF module).

7.4.2 PCA

For illustration purposes, the figures and tables in subsections 7.4.2 and 7.4.3 are based on analysis from the first 58 observations.

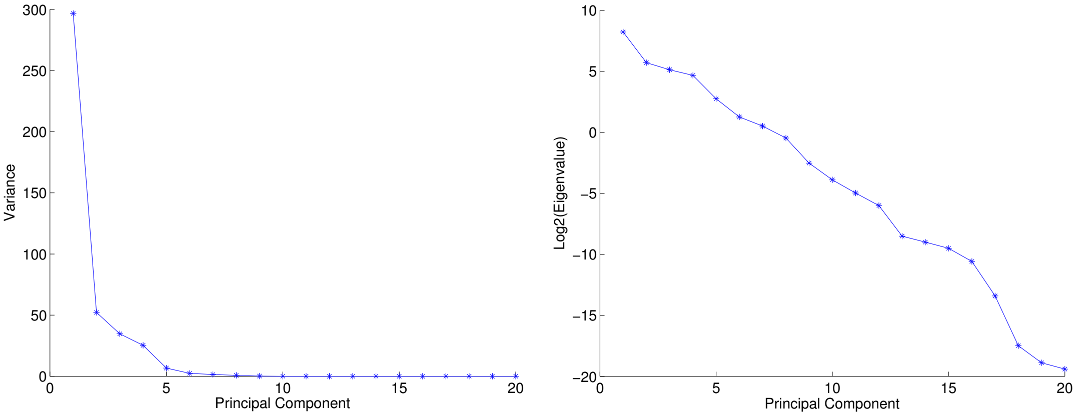
The 58×20 matrix, $\mathbf{X}$, is formed by extracting the first 58 observations of the simulated pressure data. The Gleason-Stalin statistic for the EIA data modeled as a 20-node gas system, $\Phi = 0.7169$ which implies that 71.69% of the variables in the dataset are correlated.

7.4.3 P and Q_α

P is again determined by using the *variance accounted for* method. In this gas network the usual stopping rules are difficult to interpret. As seen below in Fig. 7.4 both the scree test and the LEV test have several minor elbows, but no obvious division to create common cause and assignable cause subspaces. Following the guidelines to include PCs up to and including the beginning of the first elbow, four or five PCs would be in the common cause subspace.

Table 7.3: Variance and Cumulative Variance for each Principal Component

PC	Variance	Cumulative Variance	PC	Variance	Cumulative Variance
1	302.7849	.74	11	0.0318	0.999926504695983
2	46.7182	.86	12	0.0213	0.99997895556989
3	25.6760	.92	13	0.0052	0.999991820327402
4	20.2176	.97	14	0.0014	0.999995181127369
5	5.6006	.9866	15	0.0011	0.999997929652848
6	2.4931	.9927	16	0.0008	0.999999848805467
7	1.9114	.9974	17	0.0001	0.999999973185415
8	0.7438	.9993	18	0.0000	0.999999987610696
9	0.1979	.9997	19	0.0000	0.999999996115087
10	0.0450	.9998	20	0.0000	1



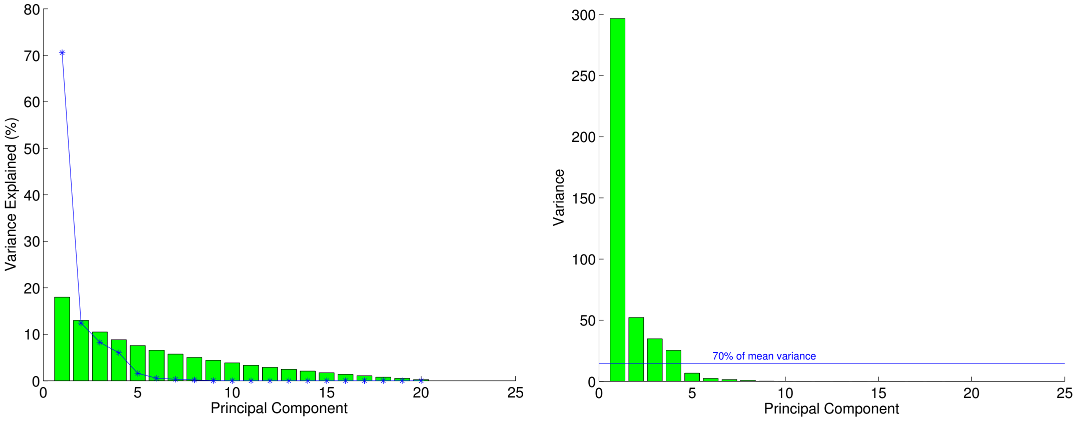
(a) Cattell's Scree Test

(b) Log Eigenvalue Test

Figure 7.4: Scree and Log Eigenvalue Tests Suggest Keeping 5 or 6 PCs

The broken stick test shown in Fig. 7.5(a) suggests only one PC should be kept since only one is not attributeable to chance alone. And, the Kaiser Gutmann test with Jolliffe's modification in Fig. 7.5(b) indicates that four PCs fall above 70% of the mean variance.

Experience indicates that choosing P based on the *amount of variance accounted for* works relatively well for anomaly identification in gas networks. In this study choosing .99998 as the variance accounted for by the common cause subspace results in $P = 12$.



(a) Broken Stick Test

(b) Kaiser Gutmann Test

Figure 7.5: Broken Stick and Kaiser Gutmann Tests Suggest Keeping 1 or 4 PCs

As seen in Fig. 7.6 twelve false anomalies are recorded when the algorithm is run with no anomaly inserted in the database. Note that Q_α changes over time based on the number of PCs ($P = 11$ or 12) necessary to represent 99.998% of the variation.

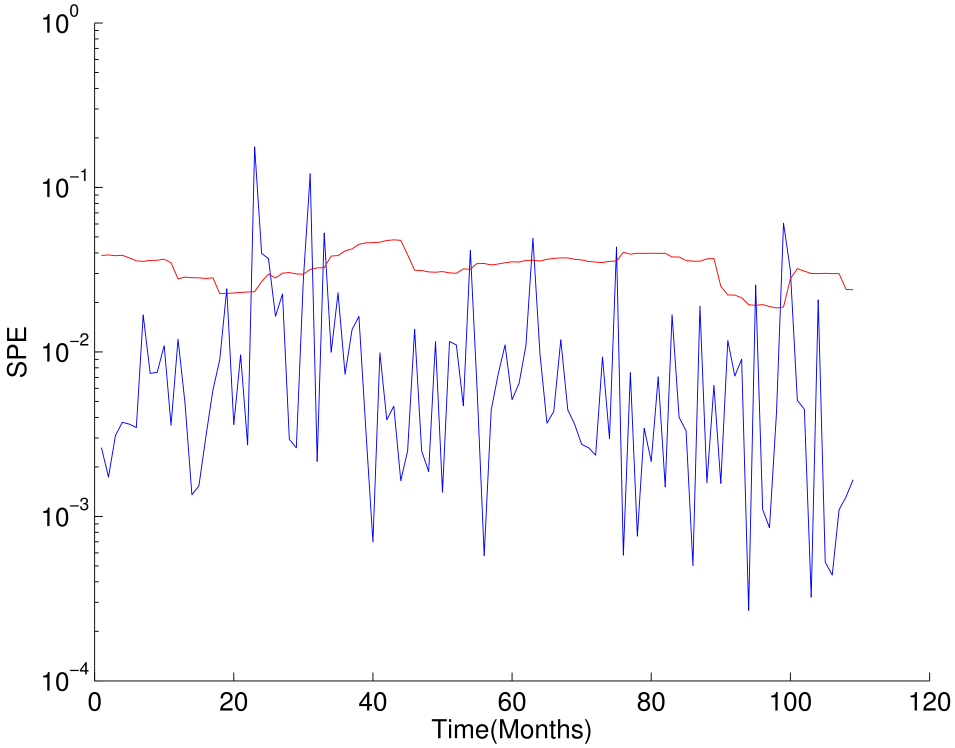


Figure 7.6: False Anomalies

7.4.4 Case Study

In this experiment it is assumed that a cyberattack has compromised the parameter for diameter that is stored in a database and, without the knowledge of the operator, input to the OGF module affecting the state variables—gas flow, source (+) and demand (-) amounts, nodal pressures and energy used by compressor stations–is erroneous. If the operator implements the incorrect state variables, damage could affect the network.

The Matlab function, *fmincon* is used to find the solution to the first problem with the vector, **b**, set to the maximum supply and ignoring the compressor stations. The results from problem 1 (equations (7.9) and (7.10)) provide the starting point for problem 2 (equations (7.2) and (7.3)) that is also solved using *fmincon*.

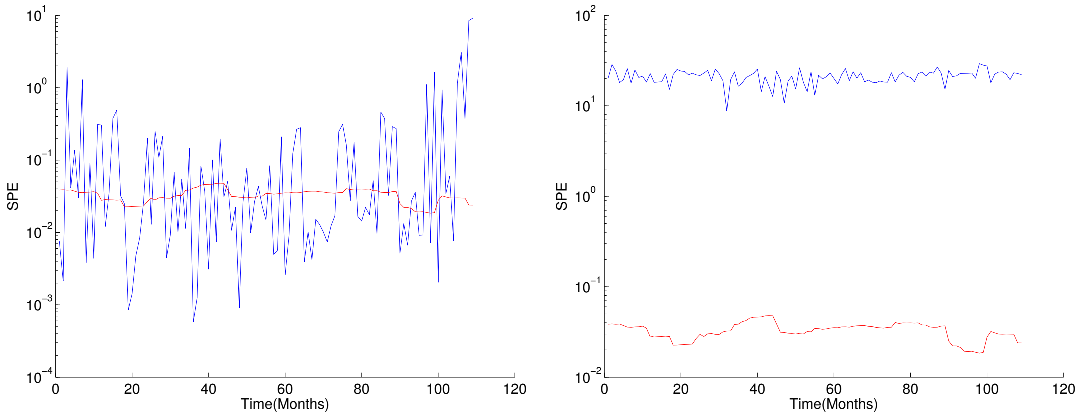
The diameter was reduced by 80%, 50% and 20% in each line (separately) and the results analyzed. Overall, 85% of the anomalies were found and over 95% of the anomalies were found on 10 of the 19 pipelines. The results in table 7.4 are for the 9 lines where less than 98% of the anomalies were found.

Table 7.4: Sample of Anomalies Detected when Diameter is Reduced

line	−80% (%)	−50% (%)	−20% (%)	Overall(%)
1	100	100	83.49	94.50
3	99.08	94.50	76.15	89.91
4	47.71	63.30	29.36	46.79
5	70.64	35.78	14.68	40.37
6	100	94.50	34.86	76.45
7	97.25	89.91	73.39	86.85
8	79.82	47.71	46.79	58.10
15	100	100	36.70	78.90
17	60.55	38.53	22.02	40.37

Fig. 7.7(a) shows that the SPE (blue line) is only greater than Q_α (red line) for 32 months, indicating the found anomalies in pipeline 4. Notice in Fig. 7.2 that the demand at

nodes 6 and 7 can be supplied from pipelines 6 and 5 as well as pipeline 4. In Fig. 7.7(b) where all anomalies are found in pipeline 11, the SPE is greater than Q_α for all 109 months.



(a) Decreased diameter on line 4 by 80% (b) Decreased Diameter on Line 11 by 50%

Figure 7.7: Results of Diameter Changes in Gas Pipelines 4 and 11

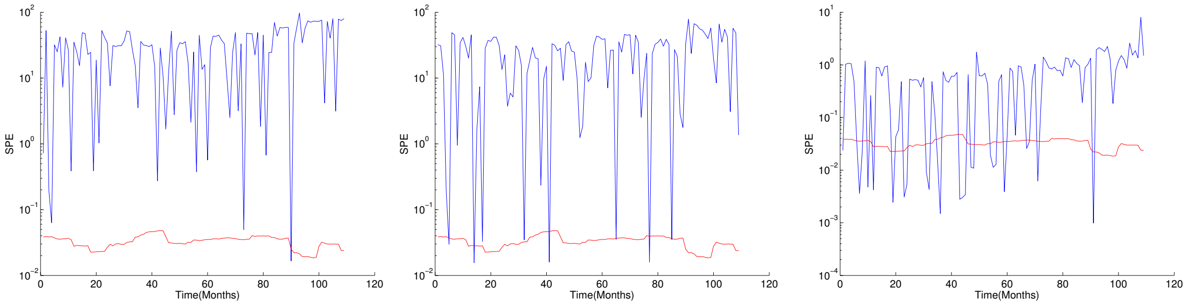
To determine the effect of the inserted anomalies that are not found by the algorithm, a sampling of the results were studied further. Line capacity in this research is approximated by equation (7.14) by using the gas flow equation (7.4) with maximum pressure at the inlet to each pipeline and minimum pressure at the outlet from each pipeline.

$$f_{ij}^{G^{MAX}} \approx \sqrt{C_{ij}^{G}(p_i^2 - p_j^2)} \quad \forall (i,j) \in A_p \cup A_a \tag{7.14}$$

Line use is subsequently calculated for each line from equation (7.15).

$$l_{ij}^{USE} = \frac{f_{ij}^{G}}{f_{ij}^{G^{MAX}}} \quad \forall (i,j) \in A_p \cup A_a \tag{7.15}$$

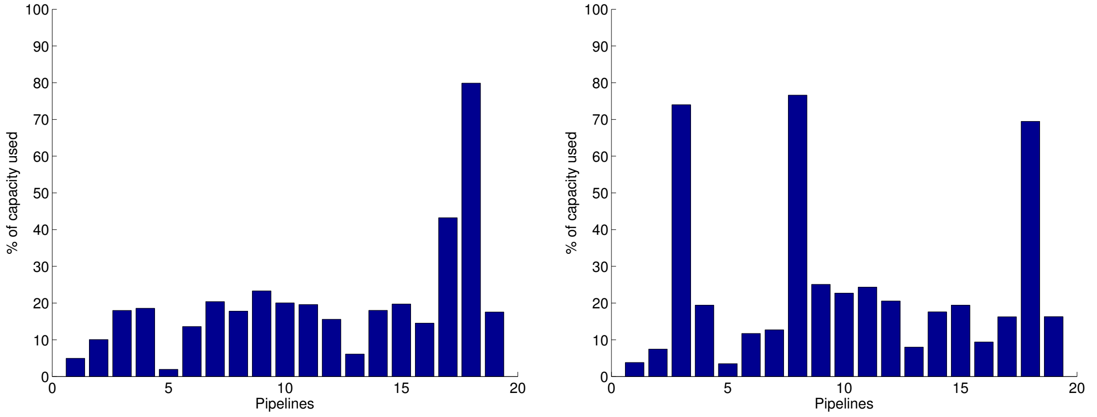
Line 3. Pipeline 3 runs from node 3 (a demand node) to node 4 (a transshipment node) and is included in the only path from node 1 to node 4. The results in Fig. 7.8 show that 108 of the anomalies are found when the pipeline diameter is decreased by 80%, 103 of the anomalies are found when the pipeline diameter is decreased by 50%, and 83 anomalies are found when the pipeline diameter is decreased by 20%.



(a) Decreased Diameter by 80% (b) Decreased Diameter by 50% (c) Decreased Diameter by 20%

Figure 7.8: Results from Pipeline 3

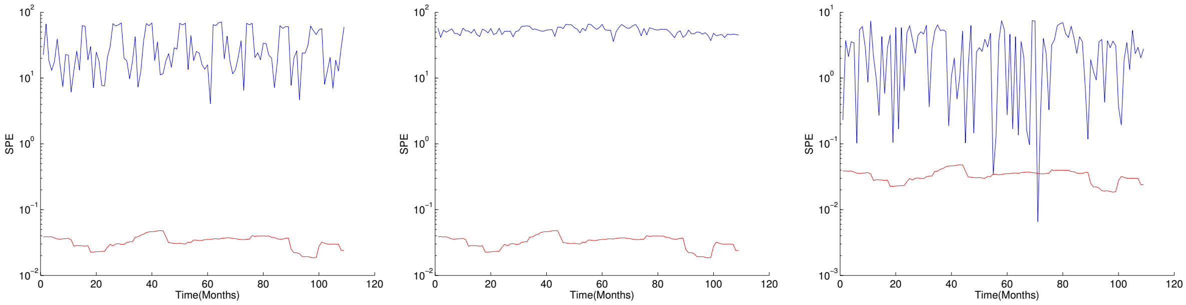
To illustrate the effect of not discovering anomalies, in Fig. 7.9 the pipeline use at time $t = 5$ is presented for each pipeline in the network when the diameter is reduced by 50% in pipeline 3. The pipeline pressures all change as do the flows through the network; however, the pipeline use shows that no pipeline is overloaded when the incorrect information is implemented by the operator.



(a) No Anomaly Inserted (b) Diameter on Pipeline 3 Decreased by 50%

Figure 7.9: Line Use Comparison in Pipeline 3

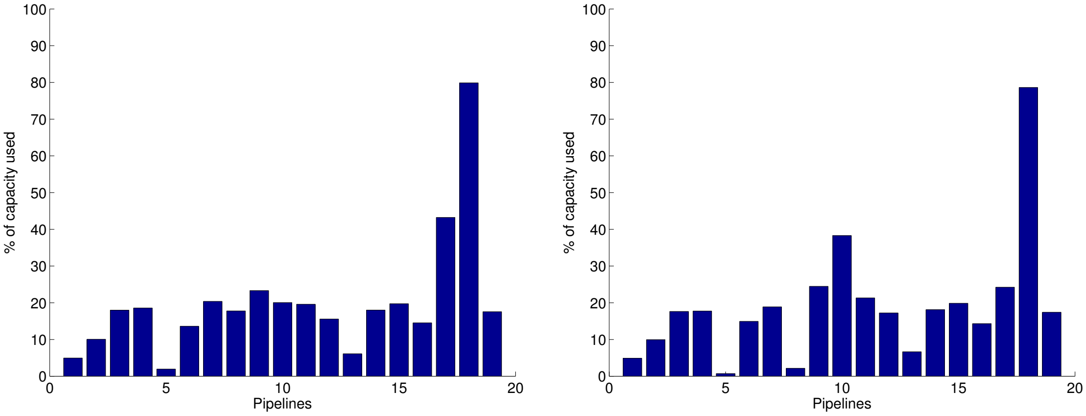
Line 10. Line 10 runs from node 10 (a demand node) to node 11 (a transshipment node). The results in Fig. 7.10 show that 109 of the anomalies are found when the diameter is decreased by 80%, 109 of the anomalies are found when the diameter is decreased by 50%, and 107 anomalies are found when the diameter is decreased by 20%.



(a) Decreased Diameter by 80% (b) Decreased Diameter by 50% (c) Decreased Diameter by 20%

Figure 7.10: Results from Pipeline 10

To further illustrate the effect of not discovering anomalies, in Fig. 7.11 the pipeline use at time $t = 55$ is presented for each pipeline in the network when the diameter is reduced by 20% in pipeline 10. The pipeline pressures all change as do the flows through the network; however, the pipeline use shows that no pipeline is overloaded when the incorrect information is implemented by the operator.



(a) No Anomaly Inserted (b) Diameter on Pipeline 10 Decreased by 80%

Figure 7.11: Line Use Comparison

7.4.5 Discussion

The small Belgium gas transmission network is a tree shaped network with no loops. In practice, this situation is unlikely to occur since redundancy is important for pigging (cleaning) the pipelines, line packing situations, and reliability of the network. Most of the situations where the optimization program did not converge are caused by infeasibility.

Reducing the diameter of the pipeline by 80% could reduce the optimized supply calculation to less than the minimal supply allowed by constraints.

A sampling of the anomalies that were not found by the algorithm revealed that in those samples the flow of gas through the pipeline changed on every pipeline in the network, but no line was overloaded. Whereas the flow changes, the constraints are still met and the anomaly is unlikely to cause failure in the network.

Chapter 8

Conclusion

Sophisticated cyberterrorists will have sufficient technical computer and critical system knowledge to devise an attack through the Internet which could compromise critical infrastructure in the United States. This research described a new class of cyberattacks to power systems–malicious modification of network data stored in an accessible database. The algorithm developed to address these cyberattacks uses the results of principal component analysis to detect data anomalies resulting from this class of attack in multiple systems.

The generic algorithm was evaluated through comprehensive testing on two well-known test cases from the power transmission industry and a test case from the natural gas transportation industry. Parameters associated with transmission lines and natural gas pipelines were changed and the application processor in each case produced state variables. PCA was used to compare the trending state variable data with current observations and an alarm notified the operator if anomalous data was encountered.

The algorithm was successful in detecting introduced anomalies at various severity levels with a reasonable number of false alarms. Overall the anomalies were detectable more than 90% of the time.

Engineers, system operators and government officials know that cybercrime prevention is not sufficient to protect critical infrastructure. Detection algorithms strategically placed in critical infrastructure management systems will certainly increase reliability.

This new algorithm adds a dimension of protection for critical infrastructure that has not previously been addressed in the literature.

Bibliography

[1] Wilsh. (2012, April) Cybersecurity: Threats Impacting the Nation. [Online]. Available: http://www.gao.gov/assets/600/590367.pdf

[2] B. Obama. (2013, February) Executive Order 13636: Improving Critical Infrastructure Cybersecurity. [Online]. Available: https://www.whitehouse.gov/the-press-office/2013/02/12/executive-order-improving-critical-infrastructure-cybersecurity

[3] (2014, February) Framework for Improving Critical Infrastructure Cybersecurity. [Online]. Available: http://csrc.nist.gov/publications/nistpubs/800-41-Rev1/sp800-41-rev1.pdf

[4] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing Network-Wide Traffic Anomalies," in *Proceedings of ACM Conference of the Special Interest Group on Data Communications (SIGCOMM).* ACM, 2004.

[5] R. Dunia, S. J. Qin, T. F. Edgar, and T. J. McAvoy, "Identification of Faulty Sensors Using Principal Component Analysis," *AIChE Journal*, vol. 42, pp. 2797–2812, 1996.

[6] W. H. Woodall, D. J. Spitzner, D. C. Montgomery, and S. Gupta, "Using Control Charts to Monitor Process and Product Quality Profiles," *Journal of Quality Technology*, vol. 36, pp. 309–320, 2004.

[7] D. Dolezilek and L. Hussey, "Requirements or Recommendations? Sorting Out NERC CIP, NIST, and DOE Cybersecurity," in *2011 64th Annual Conference for Protective Relay Engineers*, 2011.

[8] J. Valenzuela, J. Wang, and N. Bissinger, "Real-Time Intrusion Detection in Power System Operations," *Power Systems, IEEE Transactions on*, vol. 28, no. 2, pp. 1052–1062, May 2013.

[9] [Online]. Available: https://ics-cert.us-cert.gov/content/overview-cyber-vulnerabilities

[10] [Online]. Available: http://www.nerc.com/Pages/default.aspx

[11] [Online]. Available: http://www.velaw.com/uploadedFiles/VEsite/Resources/SummaryCIPVersion5Standards2014.pdf

[12] [Online]. Available: http://www.nist.gov/itl/csd/launch-cybersecurity-framework-021214.cfm

[13] [Online]. Available: http://energy.gov/sites/prod/files/oeprod/DocumentsandMedia/21_Steps_-_SCADA.pdf

[14] K. Scarfone and P. Hoffman. (2009, September) Recommendations of the National Institute of Standards and Technology. National Institute of Standards and Technology. [Online]. Available: http://csrc.nist.gov/publications/nistpubs/800-41-Rev1/sp800-41-rev1.pdf

[15] G. C. Wilshusen. (2009, May) Cyber Threats and Vulnerabilities Place Federal Systems at Risk. [Online]. Available: http://www.gao.gov/new.items/d09661t.pdf

[16] B. Wingfield. (2012, January) Power-Grid Cyber Attack Seen Leaving Millions in Dark for Months. [Online]. Available: http://www.bloomberg.com/news/2012-02-01/cyber-attack-on-u-s-power-grid-seen-leaving-millions-in-dark-for-months.html

[17] R. McMillan, "Siemens: Stuxnet Worm Hit Industrial Systems," *PCWorld*, 2010.

[18] F. F. Wu, K. Moslehi, and A. Bose, "Power System Control Centers: Past, Present, and Future," *Proceedings of the IEEE*, vol. 93, pp. 1890–1907, 2005.

[19] Integrated Topology Processing: a Breakthrough in Power System Software Unification. [Online]. Available: http://www.powerworld.com/products/IntegratedTP.asp

[20] J. A. Momoh, R. J. Koessler, M. S. Bond, B. Stott, D. Sun, A. Papalexopoulos, and P. Ristanovic, "Challenges to Optimal Power Flow," *IEEE Transactions on Power Systems*, vol. 12, pp. 444–447, 1997.

[21] F. C. Schweppe, J. Wildes, and D. B. Rom, "Power System Static State Estimation, Parts, i, ii, and iii," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-89, pp. 120–135, 1970.

[22] H. D. Merrill and F. C. Schweppe, "Bad Data Suppression in Power System Static State Estimation," *IEEE Transactions on Power Apparatus and Systems*, vol. 90, pp. 2718–2725, November/December 1971.

[23] E. Handschin, F. C. Schweppe, J. Kohlas, and A. Fechter, "Bad Data Analysis for Power System State Estimation," *IEEE Transactions on Power Apparatus and Systems*, vol. 94, pp. 329–337, 1975.

[24] L. Mili, T. 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. PAS-103, no. 11, pp. 3239–3252, 1984.

[25] A. Monticelli, F. F. Wu, and M. Yen, "Multiple Bad Data Identification for State Estimation by Combinatorial Optimization," *IEEE Transactions on Power Delivery*, vol. 1, pp. 361–369, 1986.

[26] V. H. Quintana, A. Simoes-Costa, and M. Mier, "Bad Data Detection and Identification Techniques Using Estimation Orthogonal Methods," *IEEE Transactions on Power Apparatus and Systems*, vol. 101, pp. 3355–3364, 1982.

[27] A. Simoes-Costa and V. H. Quintana, “An Orthogonal Row Processing Algorithm for Power System Sequential State Estimation,” *IEEE Tranactions on Power Apparatus and Systems*, vol. 100, no. 2, pp. 3791–3800, February 1981.

[28] ——, “A Robust Numerical Technique for Power System State Estimation,” *IEEE Transactions on Power Apparatus and Systems*, vol. 100, pp. 691–698, 1981.

[29] N. Vempati and R. R. Shoults, “Sequential Bad Data Analysis in State Estimation using Orthogonal Transformations,” *IEEE Transations on Power Systems*, vol. 6, no. 1, pp. 157–162, February 1991.

[30] Y. Liu, P. Ning, and M. K. Reiter, “False Data Injection Attacks against State Estimation in Electric Power Grids,” in *ACM Conference on Computer and Communications Security*. Associated Computer Machinery, 2009, pp. 21–32.

[31] 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 *Workshop on Secure Control Systems*, 2010.

[32] H. Sandberg, A. Teixeira, and K. H. Johansson, “Stealth Attacks and Protection Schemes for State Estimators in Power Networks,” in *1st Workshop Secure Control Systems (CPSWEEK)*, 2010.

[33] G. Dan and H. Sandberg, “Stealth Attacks and Protection Schemes for State Estimators in Power Systems,” in *2010 First IEEE International Conference on Smart Grid Communications (SmartGridComm)*, Oct. 2010, pp. 214–219.

[34] O. Kosut, L. Jia, R. J. Thomas, and L. Tong, “Malicious Data Attacks on the Smart Grid,” *IEEE Transactions on Smart Grid*, vol. 2, pp. 645–658, 2011.

[35] M. A. Rahman and H. Mohsenian-Rad, “False Data Injection Attacks Against Nonlinear State Estimation in Smart Power Grids,” in *Power and Energy Society General Metting (PES), 2013 IEEE*, 2013.

[36] Q. Yang, J. Yang, W. Yu, D. An, N. Zhang, and W. Zhao, “On False Data-Injection Attacks against Power System State Estimation: Modeling and Countermeasures,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, pp. 717–729, 2014.

[37] L. Xie, Y. Mo, and B. Sinopoli, “Integrity Data Attacks in Power Market Operations,” *IEEE Transactions on Smart Grid*, vol. 2, no. 4, pp. 659–666, December 2011.

[38] A. Giani, E. Bitar, M. Garcia, M. McQueen, P. Khargonekar, and K. Poolla, “Smart Grid Data Integrity Attacks: Characterizations and Countermeasures,” in *IEEE SmartGridComm*, 2011.

[39] T. T. Kim and H. V. Poor, “Strategic Protection Against Data Injection Attacks on Power Grids,” *IEEE Transactions on Smart Grid*, vol. 2, no. 2, pp. 326–333, June 2011.

[40] S. Mousavian, J. Valenzuela, and J. Wang, "Probabilistic Risk Mitigation Model for Cyber-Attacks to PMU Networks," *Power Systems, IEEE Transactions on*, vol. 30, pp. 156–165, 2015.

[41] Y. Fujita, T. Namerikawa, and K. Uchida, "Cyber Attack Detection and Faults Diagnosis in Power Networks by Using State Fault Diagnosis Matrix," in *2013 European Control Conference, ECC 2013*, 2013.

[42] S. Mousavian, J. Valenzuela, and J. Wang, "Real-time Data Reassurance in Electrical Power Systems Based on Artificial Neural Networks," *Electric Power Systems Research*, vol. 96, pp. 285–295, 2013.

[43] V. Ravi, "Detection of Cyber Attacks in Power Distribution Energy Management Systems," Master's thesis, Arizona State University, 2014.

[44] A. Abur and A. G. Exposito, *Power System State Estimation: Theory and Implementation*. Marcel Dekker, Inc., 2004.

[45] L. Huang, X. Nguyen, M. Garofalakis, M. Jordan, A. Joseph, and N. Taft, "In-Network PCA and Anomaly Detection," UC Berkeley, Tech. Rep., January 2007.

[46] (2009). [Online]. Available: https://ics-cert.us-cert.gov/sites/default/files/recommended_practices/Defense_in_Depth_Oct09.pdf

[47] Site last accessed March 2015. [Online]. Available: https://www.digitalbond.com/tools/quickdraw/

[48] A. A. Cardenas, S. Amin, Z.-S. Lin, Y.-L. Huang, C.-Y. Huang, and S. Sastry, "Attacks Against Process Control Systems: Risk Assessment, Detection, and Response," in *ASIACCS '11*, 2011, pp. 355–366.

[49] Y. Mo, S. Weerakkody, and B. Sinopoli, "Physical Authentication of Control Systems: Designing Watermarked Control Inputs to Detect Counterfeit Sensor Outputs," *IEEE Control Systems Magazine*, pp. 93–109, 2015.

[50] M. Clayton. (2012, May) Alert: Major Cyber Attack Aimed at Natural Gas Pipeline Companies. [Online]. Available: http://www.csmonitor.com/USA/2012/0505/Alert-Major-cyber-attack-aimed-at-natural-gas-pipeline-companies

[51] (2014). [Online]. Available: https://ics-cert.us-cert.gov/sites/default/files/Monitors/ICS-CERT_Monitor_%20Jan-April2014.pdf

[52] (2012, October). [Online]. Available: http://www.ingaa.org/file.aspx?id=19143

[53] A. J. Wood and B. F. Wollenberg, *Power Generation, Operations, and Control*, 2nd ed. John Wiley & Sons, 1996.

[54] K. Pearson, "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philosophical Magazine*, vol. 2, no. 11, pp. 559–571, 1901.

[55] H. Hotelling, “Analysis of a Complex of Statistical Variables Into Principal Components,” *Journal of Educational Psychology*, vol. 24, pp. 417–441 and 498–520, 1933.

[56] H. Ringberg, J. Rexford, A. Soule, and C. Diot, “Sensitivity of PCA for Traffic Anomaly Detection,” in *Signetrics 2007*, 2007.

[57] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed., ser. Springer Series in Statistics. Springer, 2002.

[58] T. C. Gleason and R. Staelin, “A Proposal for Handling Missing Data,” *Psychometrika*, vol. 40, pp. 229–252, 1975. [Online]. Available: http://dx.doi.org/10.1007/BF02291569

[59] J. E. Jackson, *A User’s Guide to Principal Components.* John Wiley, 2003.

[60] ——, “Stopping Rules in Principal Components Analysis: A Comparison of Heuristic and Statistical Approaches,” *Ecology*, vol. 74, pp. 2204–2214, 1993.

[61] R. Cangelosi and A. Goriely, “Component Retention in Principal Component Analysis with Application to cDNA Microarray Data,” *Biology Direct*, vol. 2, p. 2, 2007. [Online]. Available: http://www.biology-direct.com/content/2/1/2

[62] H. F. Kaiser, “The Application of Electronic Computers to Factor Analysis,” *Educational and Psychological Measurement*, vol. 20, pp. 141–151, 1960.

[63] J. Stevens, *Applied Multivariate Statistics for the Social Sciences.* Lawrence Erlbaum Associates, 1986.

[64] R. B. Cattell, “The Scree Test for the Number of Factors,” *Multivariate Behavioral Research*, vol. 1(2), pp. 245–276, 1966.

[65] R. B. Cattell and J. Jaspers, “A General Plasmode (no. 30-10-5-2) for Factor Analytic Exercises and Research,” *Multivariate Behavioral Research Research Monographs*, vol. 67, pp. 1–212, 1967.

[66] R. MacArthur, “On the Relative Abundance of Bird Species,” in *Proceedings of the National Academy of Science USA*, vol. 43, 1957, pp. 293–295, pMC free article, PubMed.

[67] S. Frontier, “Study of the Decay of Values in a Principal Component Analysis: Comparison with the Model of the Broken Stick,” *Journal of Experimental Marine Biology and Ecology*, vol. 25, pp. 67–75, 1976.

[68] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly Detection: A Survey.” *ACM Computing Surveys*, vol. 41, no. 3, July 2009. [Online]. Available: http://dl.acm.org/citation.cfm?doid=1541880.1541882

[69] F. Y. Edgeworth, “On Discordant Observations,” *Philosophical Magazine*, vol. 23, pp. 364–375, 1887.

[70] S. J. Roberts, “Extreme Value Statistics for Novelty Detection in Biomedical Signal Processing,” in *First International Conference on Advances in Medical Signal and Information Processing*, 2000.

[71] T. Fawcett and F. Provost, “Activity Monitoring: Noticing Interesting Changes in Behavior,” in *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, A. Press, Ed., 1999, pp. 53–62.

[72] M. Desforges, P. Jacob, and J. Cooper, “Applications of Probability Density Estimation to the Detection of Abnormal Conditions in Engineering,” in *Proceedings of the Institute of the Mechanical Engineers*, vol. 212, 1998, pp. 687–703.

[73] X. Song, M. Wu, C. Jermaine, and S. Ranka, “Conditional Anomaly Detection,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, pp. 631–645, 2007.

[74] E. B. Martin, A. J. Morris, and J. Zhang, “Process Performance Monitoring Using Multivariate Statisitical Process Control,” *IEE Proceedings - Control Theory and Applications*, vol. 143, pp. 132–144, 1996.

[75] A. Ferrer, “Multivariate Statistical Process Control Based on Principal Component Analysis (mspc-pca): Some Reflections and a Case Study in an Autobody Assembly Process,” *Quality Engineering*, vol. 19, pp. 311–325, 2007.

[76] W. A. Shewhart, *Statistical Method from the Viewpoint of Quality Control*, W. E. Deming, Ed. Dover Publications, Inc., 1986.

[77] Y. Liu, L. Zhang, and Y. Guan, “Sketch-Based Streaming PCA Algorithm for Network-Wide Traffic Anomaly Detection,” in *Distributed Computing Systems (ICDCS), 2010 IEEE 30th International Conference on*, june 2010, pp. 807 –816.

[78] J. E. Jackson and G. S. Mudholkar, “Control Procedures for Residuals Associated with Principal Component Analysis,” *Technometrics*, vol. 21, pp. 341–349, 1979.

[79] Matlab student version release r2014a. The MathWorks, Inc. Massachusetts, United States.

[80] R. D. Zimmerman, C. E. Murillo-Sanchez, and R. J. Thomas, “Matpower: Steady-State Operations, Planning and Analysis Tools for Power Systems Research and Education,” *IEEE Transactions on Power Systems*, vol. 26, no. 1, pp. 12–19, February 2011.

[81] PJM Operational Data. [Online]. Available: http://www.pjm.com

[82] M. B. Cain, R. P. O’Neill, and A. Castillo. (2012, December) History of Optimal Power Flow and Formulations. Federal Energy Regulatory Commission. [Online]. Available: http://www.ferc.gov/industries/electric/indus-act/market-planning/opf-papers/acopf-1-history-formulation-testing.pdf

[83] J. Carpentier, “Contribution e l’etude do Dispatching Economique,” *Bulletin Society Francaise Electriciens*, vol. 3, 1962.

[84] IEEE RTS Task Force of APM Subcommittee, “IEEE Reliability Test System,” *IEEE Transactions on Power Apparatus and Systems*, vol. 98, no. 6, pp. 2047–2054, 1979.

[85] C. Coffrin, D. Gordon, and P. Scott. (2015, May) NESTA: The Nicta Energy System Test Case Archive. [Online]. Available: http://arxiv.org/pdf/1411.0359v3.pdf

[86] C. Coffrin. Updated Test Cases for use with MatPower. [Online]. Available: https://github.com/nicta/nesta

[87] M. van Steen, *Graph Theory and Complex Networks: An Introduction.* Maarten van Steen, 2010.

[88] P. Panigrahi, “Topological Analysis of Power Grid to Identify Vulnerable Transmission Lines and Nodes,” Master’s thesis, National Institute of Technology, Rourkela in Odisha, Inda, 2013.

[89] Y. Koc, M. Warnier, R. E. Kooij, and F. M. T. Brazier, “A Robustness Metric for Cascading Failures by Targeted Attacks in Power Networks,” in *Networking, Sensing and Control (ICNSC), 2013 IEEE International Conference on*, April 2013, pp. 48–53.

[90] Y. Koc, M. Warnier, R. e. Kooij, and F. M. T. Brazier. (2013, December) Structural Vulnerability Assessment of Electric Power Grids. [Online]. Available: arXiv:1312.6606v1[physics.soc-ph]

[91] D. L. Pepyne, “Topology and Cascading Line Outages in Power Grids,” *Journal of Systems Science Systems Engineering*, vol. 16, pp. 202–221, 2007.

[92] F. Cadini, E. Zio, and C. A. Petrescu, *Critical Information Infrastructure Security.* Springer Berlin Heidelbert, 2009, ch. Using Centrality Measures to Rank the Importance of the Components of a Complex Network Infrastructure, pp. 155–167.

[93] A. A. Jamshidifar, “Optimization of Natural Gas Transmission Network Using Genetic Algorithm,” in *11th International Conference on Intelligent Systems Design and Applications*, 2011.

[94] S. Moaveni, *Engineering Fundamentals: An Introduction to Engineering.* Cengage Learning; ’005 edition (January 1, 2015), 2014.

[95] B. Bakhouya and D. D. Wolf, “Solving the Gas Transmission Problem with Consideration of the Compressors,” HEC Ecole de Gestion de l’Universite de Liege (ULG), Liege, Belgium, Tech. Rep., 2008.

[96] R. Z. Rios-Mercado and C. Borraz-Sanchez, “Optimization Problems in Natural Gas Transportation Systems: A State-of-the-Art review,” *Applied Energy*, vol. 147, pp. 536–555, 2015.

[97] A. J. Osiadacz, *Simulation and Analysis of Gas Networks.* D & F.N. Spon Ltd, 1987.

[98] A. D. Woldeyohannes, M. A. A. Majid, C. F. Chyuan, and A. T. Baheta, "Matlab Based Performance Evaluation of Natural Gas Transmission System due to Corrosion," *Journal of Petroleum Science Research*, vol. 3, pp. 16–23, 2014.

[99] D. De Wolf and Y. Smeers, "Optimal Dimensioning of Pipe Networks with Application to Gas Transmission networks," *Operations Research*, vol. 44, pp. 596–608, 1996.

[100] ——, "The Gas Transmission Problem Solved by an Extension of the Simplex Algorithm," *Management Science*, vol. 46, pp. 1454–1465, 2000.

[101] [Online]. Available: http://www.eia.gov