Data Analytics for Statistical Learning in Healthcare and Manufacturing

Tomilayo Komolafe

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Industrial and Systems Engineering

Jaime A. Camelio, Committee Chair
Pablo A. Tarazaga
Zhenyu (James) Kong
Raj M. Ratwani
Srijan Sengupta

December 6, 2018
Blacksburg, VA

Keywords: Advanced manufacturing, Anomaly detection, Cyber-physical attacks, Electro-mechanical impedance, Instrumented fixture, Machine learning, Social determinants of health

Data Analytics for Statistical Learning in Healthcare and Manufacturing

Tomilayo Komolafe

ABSTRACT

The prevalence of big data has rapidly changed the usage and mechanisms of data analytics within organizations. Big data is a widely-used term without a clear definition. The difference between big data and traditional data can be characterized by four V's: *velocity* (speed at which data is generated), *volume* (amount of data generated), *variety* (the data can take on different forms), and *veracity* (the data may be of poor/unknown quality). As many industries begin to recognize the value of big data, organizations try to capture it through means such as: side-channel data in a manufacturing operation, unstructured text-data reported by healthcare personnel, various demographic information of households from census surveys, and the range of communication data that define communities and social networks.

Big data analytics generally follows this framework: first, a digitized process generates a stream of data, this raw data stream is pre-processed to convert the data into a usable format, the pre-processed data is analyzed using statistical tools. In this stage, called 'statistical learning of the data', analysts have two main objectives (1) develop a statistical model that captures the behavior of the process from a sample of the data (2) identify anomalies in the process. Figure 1 illustrates the general framework of big data analytics.

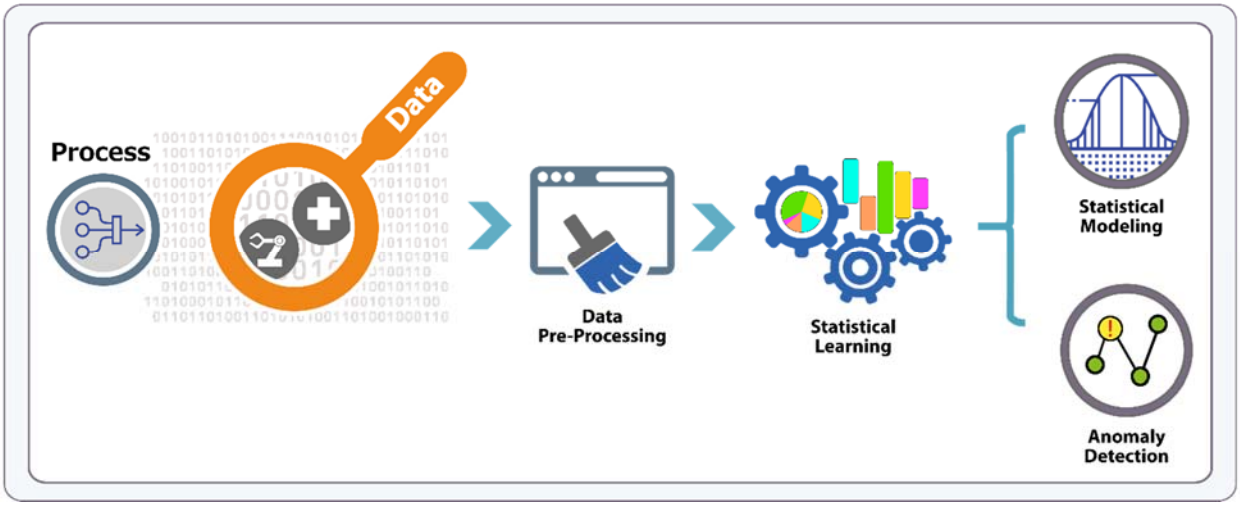

Figure 1. Framework describing the mechanisms of big data analytics

However, several open challenges still exist in this framework for big data analytics. Recently, data types such as free-text data are also being captured. Although many established

processing techniques exist for other data types, free-text data comes from a wide range of individuals and is subject to syntax, grammar, language, and colloquialisms that require substantially different processing approaches. Once the data is processed, open challenges still exist in the statistical learning step of understanding the data.

Statistical learning aims to satisfy two objectives, (1) develop a model that highlights general patterns in the data (2) create a signaling mechanism to identify if outliers are present in the data. Statistical modeling is widely utilized as researchers have created a variety of statistical models to explain everyday phenomena such as predicting energy usage behavior, traffic patterns, and stock market behaviors, among others. However, new applications of big data with increasingly varied designs present interesting challenges. Consider the example of free-text analysis posed above. There's a renewed interest in modeling free-text narratives from sources such as online reviews, customer complaints, or patient safety event reports, into intuitive themes or topics. As previously mentioned, documents describing the same phenomena can vary widely in their word usage and structure.

Another recent interest area of statistical learning is using the environmental conditions that people live, work, and grow in, to infer their quality of life. It is well established that social factors play a role in overall health outcomes, however, clinical applications of these social determinants of health is a recent and an open problem. These examples are just a few of many examples wherein new applications of big data pose complex challenges requiring thoughtful and inventive approaches to processing, analyzing, and modeling data.

Although a large body of research exists in the area of anomaly detection increasingly complicated data sources (such as side-channel related data or network-based data) present equally convoluted challenges. For effective anomaly-detection, analysts define parameters and rules, so that when large collections of raw data are aggregated, pieces of data that do not conform are easily noticed and flagged

In this work, I investigate the different steps of the data analytics framework and propose improvements for each step, paired with practical applications, to demonstrate the efficacy of my methods. This paper focuses on the healthcare, manufacturing and social-networking industries, but the materials are broad enough to have wide applications across data analytics generally. My main contributions can be summarized as follows:

- In the big data analytics framework, raw data initially goes through a pre-processing step. Although many pre-processing techniques exist, there are several challenges in pre-processing text data and I develop a pre-processing tool for text data.
- In the next step of the data analytics framework, there are challenges in both statistical modeling and anomaly detection
    - I address the research area of statistical modeling in two ways:
        - There are open challenges in defining models to characterize text data. I introduce a community extraction model that autonomously aggregates text documents into intuitive communities/groups
        - In health care, it is well established that social factors play a role in overall health outcomes however developing a statistical model that characterizes these relationships is an open research area. I developed statistical models for generalizing relationships between social determinants of health of a cohort and general medical risk factors
    - I address the research area of anomaly detection in two ways:
        - A variety of anomaly detection techniques exist already, however, some of these methods lack a rigorous statistical investigation thereby making them ineffective to a practitioner. I identify critical shortcomings to a proposed network based anomaly detection technique and introduce methodological improvements
        - Manufacturing enterprises which are now more connected than ever are vulnerably to anomalies in the form of cyber-physical attacks. I developed a sensor-based side-channel technique for anomaly detection in a manufacturing process

Data Analytics for Statistical Learning in Healthcare and Manufacturing

Tomilayo Komolafe

GENERAL AUDIENCE ABSTRACT

The prevalence of big data has rapidly changed the usage and mechanisms of data analytics within organizations. The fields of manufacturing and healthcare are two examples of industries that are currently undergoing significant transformations due to the rise of big data. The addition of large sensory systems is changing how parts are being manufactured and inspected and the prevalence of Health Information Technology (HIT) systems in healthcare systems is also changing the way healthcare services are delivered. These industries are turning to big data analytics in the hopes of acquiring many of the benefits other sectors are experiencing, including: reducing cost, improving safety, and boosting productivity. However, there are many challenges that exist along the framework of big data analytics, from pre-processing raw data, to statistical modeling of the data, and identifying anomalies present in the data or process. This work offers significant contributions in each of the aforementioned areas and includes practical real-world applications.

Big data analytics generally follows this framework: first, a digitized process generates a stream of data, this raw data stream is pre-processed to convert the data into a usable format, the pre-processed data is analyzed using statistical tools. In this stage, called 'statistical learning of the data', analysts have two main objectives (1) develop a statistical model that captures the behavior of the process from a sample of the data (2) identify anomalies or outliers in the process.

In this work, I investigate the different steps of the data analytics framework and propose improvements for each step, paired with practical applications, to demonstrate the efficacy of my methods. This work focuses on the healthcare and manufacturing industries, but the materials are broad enough to have wide applications across data analytics generally. My main contributions can be summarized as follows:

- In the big data analytics framework, raw data initially goes through a pre-processing step. Although many pre-processing techniques exist, there are several challenges in pre-processing text data and I develop a pre-processing tool for text data.
- In the next step of the data analytics framework, there are challenges in both statistical modeling and anomaly detection
    - I address the research area of statistical modeling in two ways:

- There are open challenges in defining models to characterize text data. I introduce a community extraction model that autonomously aggregates text documents into intuitive communities/groups
- In health care, it is well established that social factors play a role in overall health outcomes however developing a statistical model that characterizes these relationships is an open research area. I developed statistical models for generalizing relationships between social determinants of health of a cohort and general medical risk factors

- I address the research area of anomaly detection in two ways:
  - A variety of anomaly detection techniques exist already, however, some of these methods lack a rigorous statistical investigation thereby making them ineffective to a practitioner. I identify critical shortcomings to a proposed network based anomaly detection technique and introduce methodological improvements
  - Manufacturing enterprises which are now more connected than ever are vulnerably to anomalies in the form of cyber-physical attacks. I developed a sensor-based side-channel technique for anomaly detection in a manufacturing process

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation (NSF) grant number 1446804 - *CPS: Synergy: Collaborative Research: Cyber-Physical Approaches to Advanced Manufacturing Security*. However, any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

In addition, I want to acknowledge the efforts of all the co-authors I have worked with on different publications while being enrolled in the PhD program at the Grado Department of Industrial and Systems Engineering: Mr. Allan Fong, Dr. Sarah Parker, Ms. Ella Franklin (Fong *et al., 2017*), Dr. Wenmeg Tian, Dr. Gregory T. Purdy, Dr. Mohammad Albakri, Dr. Pablo Tarazaga (Komolafe *et al., 2017*), Dr. Srijan Sengupta, Dr. Leanna House, Dr. Alan Lattimer (Komolafe *et al., 2018*), Ms. Prakriti Gupta, Mr. Aditya Bharadwaj, (Prakriti *et al., 2018*), Ms. A. Valeria Quevedo, Dr. William H. Woodall (Komolafe *et al., 2018*), Mr. Zachariah Dirazonian, Dr. Jaime Camelio, Dr. Raj M. Ratwani (Komolafe *et al., 2018*), Mr. Chenang Liu, Dr. Zhenyu (James) Kong (Komolafe *et al., 2018*).

I would also like to thank my advisor, for offering me the opportunity to earn a PhD degree from Virginia Tech: my remaining committee members, for their helpful advice and valuable comments - and all my colleagues in the Center of Innovation-based Manufacturing (CIbM). Also, this work could not have been done without the efforts of Andwele Grant and Randy Waldron who were instrumental in designing and manufacturing the fixture used in this work. Finally, I would like to thank Allan Fong, Dr. Lee Wells, Heather Scott, MedStar Institute of Innovation and Socially Determined, for providing guidance and feedback on this work.

Contents

List of Tables

List of Figures

1 INTRODUCTION

The prevalence of big data has rapidly changed the usage and mechanisms of data analytics within organizations (DeSmit, Elhabashy et al. , Savage, Zhang et al. 2014, DeSmit 2016). Big data is a widely-used term, without a clear definition, but the difference between big data and traditional data can be characterized by four V's: *velocity* (speed at which data is generated), *volume* (amount of data generated), *variety* (the data can take on different forms), and *veracity* (the data may be of poor/ unknown quality). As many industries begin to recognize the value of big data, organizations try to capture it through means such as: side-channel data in a manufacturing operation, unstructured text-data reported by healthcare personnel, various demographic information of households from census surveys, and the range of communication data that define communities and social networks.

Data analytics combined with the power of big data has far-reaching possibilities. For example, in manufacturing, big data could be used to detect the presence of cyber-physical attacks that alter a part or process by analyzing an array of side-channel data sources. In healthcare, big-data in the form of the millions of unstructured text data captured annually can be analyzed to understand the frequency and types of medical errors. Demographic information of a neighborhood can be examined to predict health outcomes. In network monitoring, social networks can be scanned to identify a terrorist cell (DeSmit, Elhabashy et al. , Wells and Camelio 2013, Savage, Zhang et al. 2014, Wells, Camelio et al. 2014, Vincent, Wells et al. 2015, DeSmit 2016, Mazrae Farahani, Baradaran Kazemzadeh et al. 2016).

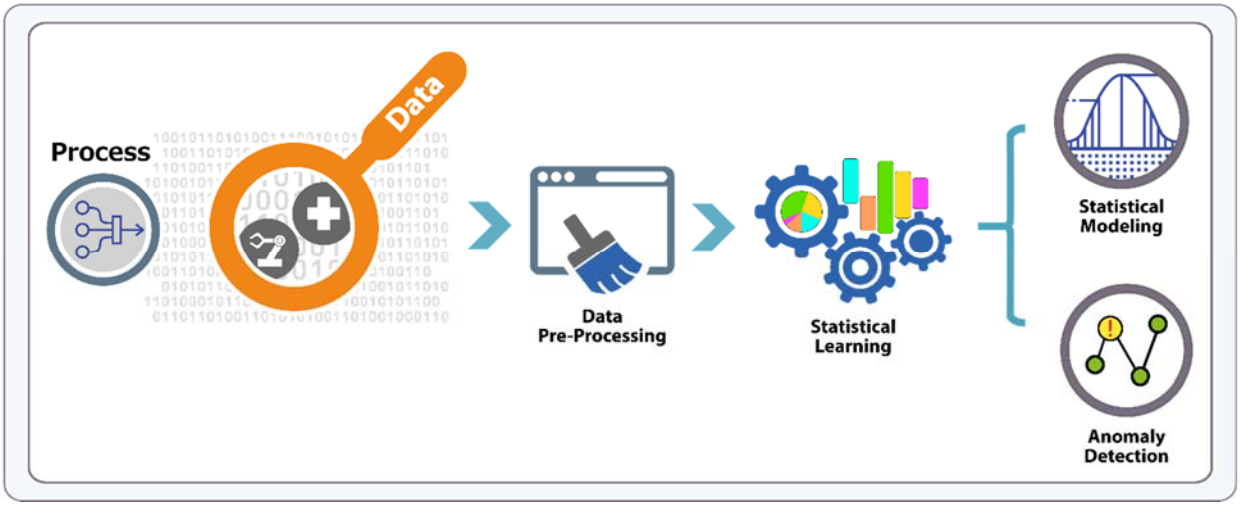

Figure 2. Framework describing the mechanisms of big data analytics

Big data analytics will often follow a similar pattern (depicted visually in Fig 1): first, a digitized process generates a stream of data, then this raw data stream is pre-processed to convert the data into a usable format, and finally the pre-processed data is analyzed using statistical tools. In this stage, analysts have two main objectives: (1) develop a statistical model that extracts patterns from a subsample of the data and subsequently generalizes to the whole statistical population (2) identify anomalies in the data.

1.1 Research area in data pre-processing

However, several open challenges still exist throughout the steps of big data analytics. For example, the recent practice of capturing free-form text data deals with major issues where meanings and phrases within text could be convoluted or idiomatic. This work addresses the problem of word-sense disambiguation in free text narratives written by front-line staff in the healthcare sector. Ambiguous words in healthcare documents hinder the applicability of many natural language processing (NLP) tools (Pakhomov, Pedersen et al. 2005, Moon, McInnes et al. 2015). While many other data types have established pre-processing standards, because free-text data comes from a wide range of individuals it is subject to their individual syntax, grammar, language, and colloquialisms. All of these require substantially different pre-processing approaches (Pakhomov, Pedersen et al. 2005, Chasin, Rumshisky et al. 2014, Moon, McInnes et al. 2015).

Textual analysis requires that the terms within the texts have clear, defined meanings. The interpretations of these terms are often difficult to glean from ambiguous communications shared between colleagues within healthcare facilities. An autonomous method for identifying context and meaning of ambiguous terms within electronic health organizations (EHRs) could improve document retrieval, enhance NLP analytics in healthcare, and open a standardized communication line between different healthcare organizations (Naessens 2017).

1.2 Research area in statistical learning

After data pre-processing, the next step in big data analytics is statistical learning. As previously mentioned, one objective of statistical learning is to develop a model that highlights patterns in the data. Researchers have applied statistical modeling to uncover patterns that explain phenomena such as energy use behavior, traffic patterns, and stock market trends. However, new applications of big data with increasingly varied designs present interesting challenges. Consider the example of free-text analysis posed above, there's a renewed interest in analyzing free-text

narratives from sources such as online reviews, customer complaints, or patient safety event reports in a hospital.

The healthcare sector is rapidly digitizing its processes in order to improve patient outcomes while reducing costs. As a result, many hospitals are integrating Health Information Technologies (HIT), thereby producing a cache of big data. Big data analytics in healthcare spans a wide range of processes, from storing radiology lab results, to patient demographic information, to clinical notes on patients. Healthcare organizations, in particular, could benefit from analyzing free-text narratives contained in their HIT systems. Patient safety event reports (PSEs), for example, are free text reports that aim to account for any concerns that a staff notices during patient care. In PSEs, front-line staff submit a report about unsafe conditions, hazards, or serious safety events observed during healthcare delivery. These reports are critical. Recent data has shown that medical errors that occur during the administration of healthcare related services are currently the 3rd leading cause of death in the U.S. (Makary and Daniel 2016, Kavanagh, Saman et al. 2017, Naessens 2017). Therefore, quick and thorough analysis of patient safety event text could highlight patterns in many of these serious safety events and provide some path to reducing errors.

Many healthcare centers have procedures for documenting and storing PSEs. Since there can be thousands or even millions of files, having dedicated staff members read through every single PSE is both cumbersome and unrealistic. However, contained in these PSEs is information that, when aggregated, could be helpful in: (1) identifying types of errors that frequently occur at the facility, (2) identifying errors that typically lead to serious illnesses or death (3) capturing performance of certain units over time for various healthcare-related concerns (4) creating accountability for departments or units that should be held responsible for certain errors. All of these potential benefits are much easier to realize when PSEs are aggregated through connecting text patterns.

Another open challenge in statistical modeling is predicting individual behavioral patterns based on a social context. For example, Netflix, a large distributor of movies and TV shows, recommends TV shows to its consumers based on social cues. Google, the world's largest internet search engine, directs specific ads to its users based on their social behavioral patterns. In healthcare, there is a similar challenge. Recently, researchers are exploring the feasibility of using statistical models to predict an individual's health outcome based on the social conditions of his or her neighbors. This area of study is called the social determinants of health. Social determinants

of health (SDOH), are the social conditions in the environments in which people live, learn, work, play, worship, and age that affect a wide range of health, quality-of-life outcomes, and risks (Koh, Piotrowski et al. 2011).

SDOH span a wide range of social conditions, from economic well-being to food insecurity, housing, and transportation among others. In many SDOH studies, researchers pair a particular SDOH with a specific set of health outcomes (Kuruvilla, Schweitzer et al. 2014, Klinenberg 2016). However, negative living conditions affect more than a limited set of health conditions. For example, low income impacts more than just a likelihood of heart disease. Transportation barriers limit the ability to have regular preventative care, for all patients regardless of whether they have a heart or lung disease. Similarly, food insecurity or poor housing affects a range of health concerns. Therefore, a study that focuses on pairing SDOH with general health outcomes, rather than specific set of diseases, is beneficial. I investigate the effect of SDOH on medical risk factors and behavioral patterns such as emergency department (ED) utilization and body mass index (BMI). These risk factors aim at capturing both the at-risk population and the possible range of health complications that can arise from SDOH. Looking at health outcomes through general features brings together different aspects of the health care system.

Another common challenge with measuring SDOH is that, many data sources in existing SDOH research are wholly comprised of participant responses to survey questions. The data and results are reliant on the recall and/or attentiveness of a participant. Furthermore, individual survey measurements are costly and difficult to obtain. Rather, it would be beneficial to use data sources which are publicly validated, credible, and statistically sound. Lastly, most research areas like (Pickett and Pearl 2001, Sampson, Martins et al. 2018) use census blocks or administrative districts to define these neighborhood boundaries, however, these methodologies have some limitations. For example, census sampling areas (census tracts) in the U.S. are drawn to encapsulate between 4000-8000 households (Schlossberg 2003). However, this methodology of sampling is not designed for human interaction with the neighborhood. This approach crosscuts natural neighborhood boundaries and infrastructure systems and does not account for the local interactions within neighborhoods and their community. Rather, measuring SDOH effects at the appropriate neighborhood boundary benefits the statistical learning process because those areas are more likely to have shared infrastructures and SDOH limitations that more accurately depict patterns.

1.3 Research area in anomaly detection

The second aim of statistical learning is to create a signaling mechanism that identifies anomalies in the data. Although a large body of research exists in the area of anomaly detection, increasingly complicated data sources (such as side-channel related data or network-based data) can confound traditional methodologies. For effective anomaly-detection, analysts define parameters and rules, so that when large collections of raw data are aggregated, pieces of data that do not conform are easily noticed and flagged.

Because a range of phenomena can be represented as networks, anomaly detection in network monitoring spans a wide range of fields and applications. For example, network monitoring can be applied to identifying clandestine terrorist cells in a social network, mutations in gene transcription, or underutilized power stations in a power-grid network (Procter, Thompson et al. 2010, Cer, Bruce et al. 2011, Cer, Bruce et al. 2012). Fittingly, many anomaly detection techniques have been developed. Typically, anomaly detection techniques focus on defining what conditions constitute a normal network and discriminating between anomalous nodes and non-anomalous nodes (Papadimitriou, Kitagawa et al. 2003, Miller, Beard et al. 2015).

Networks can be static, where we have a single snapshot of the system, or dynamic, where we have network snapshots at several points in time. Anomalies can have different meanings in these two scenarios (Papadimitriou, Kitagawa et al. 2003, Miller, Beard et al. 2015). In dynamic networks, an anomaly typically corresponds to a group of nodes behaving in a manner that is significantly different from past behavior. The general approach for detecting such anomalies is to extract some features of the network (such as centrality measures, degree distribution, etc.), and monitor these features over time, and raise a signal when these observed features cross a pre-determined threshold. A rich class of anomaly detection techniques have been developed for dynamic networks, for example, density-based techniques (Papadimitriou, Kitagawa et al. 2003), clustering-based techniques (Wang, Xie et al. 2012), distribution-based techniques (Šaltenis 2004, Akoglu, McGlohon et al. 2010, Akoglu, Tong et al. 2015), and scan methods (Priebe, Conroy et al. 2005). On the other hand, the goal of anomaly detection in a static network is to detect a subgraph that is significantly different from the overall network (Miller et al. 2015). Some popular approaches include network analysis at the egonet level (Akoglu, Tong et al. 2015), spatial autocorrelation (Ranshous, Shen et al. 2015), and modularity maximization (Haveliwala 2003, Sun, Qu et al. 2005, Newman 2016). In this study, I restrict my analysis to static networks.

Anomaly detection technique consists of computing a particular network metric (in statistical terms, the *test statistic*) and comparing its value to a benchmark distribution (in statistical terms, the *null distribution*) which represents the distribution of the metric in absence of an anomaly. If the value of the metric exceeds a determined threshold, obtained from the benchmark distribution, an anomaly is signaled. There can be two kinds of errors: false alarms that happen when the value of the metric exceeds the threshold although there is no anomaly, and detection failures when the value of the metric is below the threshold in spite of an anomaly. A principled statistical evaluation is therefore critically important to systematically study whether the related assumptions are satisfied for a wide range of scenarios: 1) the network metric should closely follow the benchmark distribution when there is no anomaly, 2) the probability of false alarms is low and close to target values, and 3) the probability of detection failures is low and close to target values. However, there has been relatively little work in such evaluation of anomaly detection techniques and I address these critical shortcomings in this work along with introducing some methodological improvements.

This work also extends the application of anomaly detection to a manufacturing setting. The rise of Industry 4.0 has led to a digitization of many manufacturing processes (Dr. Ralf C. Shlaepfer 2014, Vincent, Wells et al. 2015, DeSmit 2016). There is now a stream of data, via side-channels, that accompany a part as it moves through a manufacturing plant. Parameters such as the temperature of the part during processing, feed rate, feed speed of the cutter, vibrations metrics, among many other parameters are being monitored (Dr. Ralf C. Shlaepfer 2014, Vincent, Wells et al. 2015, DeSmit 2016). However, the rise of the connected manufacturing plant has also increased the vulnerability of a manufacturing enterprise to cyber-physical attacks. These are attacks that aim to alter, maliciously, a part or process and a digitized factory lends itself to a myriad of attack possibilities.

One approach to improving the resilience of manufacturing enterprises is to integrate additional sensors that can augment current non-destructive evaluation (NDE) techniques. Piezoelectric transducers (PZT) which act as both actuators and sensors, have been used for structural health monitoring of bridges, gears, and avionics components since the 1990s (Sun, Chaudhry et al. 1995, Raju, Park et al. 1999, Park, Cudney et al. 2001). For a given part, a PZT generates a unique signature that captures the inherent properties of that part in the form of its mass, stiffness, or damping properties. Unlike traditional inspection techniques, which focus on a

geometric feature or a set of geometric features to determine if a part is conforming or not, PZTs capture the interactions between different features. This approach to NDE is more robust to capturing alterations in a part or process because it encompasses information about the overall part.

However, there are many challenges that must be addressed for this integration of PZT sensors in manufacturing to be successful. First, as a manufacturing process includes physical parts, a sensor based technique should take into account the environment in which the physical part is being created including how operators will interact with the part. It is also critical that an accompanying quality loss metric is developed that can distinguish between conforming parts and non-conforming parts.

1.4 Motivation

The fields of manufacturing and healthcare are two examples of industries that are currently undergoing significant transformations due to the rise of big data. The rise of connected systems is changing the way networks (manufacturing or healthcare) are monitored, the addition of large sensory systems is changing how parts are being manufactured and inspected, and the prevalence of HIT in healthcare systems is also changing the way healthcare services are delivered. These industries are turning to big data analytics in the hopes of acquiring many of the benefits other sectors are experiencing, including: reducing cost, improving safety, and boosting productivity. However, there are many challenges that exist along the framework of big data analytics, from pre-processing raw data, to statistical modeling of the data, and also identifying anomalies present in the data or process. This work offers significant contributions in each of the aforementioned areas and includes practical applications.

1.5 Research objectives

This work investigates different steps of the data analytics framework and proposes improvements for each step, paired with practical applications, to demonstrate the efficacy of new methods. This work focuses on the healthcare and manufacturing industries, but the techniques are broad enough to have wide applications across data analytics generally. My main contributions can be summarized as follows:

- In the big data analytics framework, raw data initially goes through a pre-processing step. Although many pre-processing techniques exist, there are several challenges in pre-processing text data and in this work, a pre-processing tool for text data is developed. Machine learning

is applied to the problem of word-sense disambiguation to pre-process text data and a web application is created to demonstrate the utility of the proposed methodology.

- In the next step of the data analytics framework, there are challenges in both statistical modeling and anomaly detection
- I address the research area of statistical modeling in two ways:
    - There are open challenges in defining models to characterize text data. I introduce a community extraction model that autonomously aggregates text documents into intuitive communities/groups. I compare the effectiveness of my unsupervised document aggregation technique to a corpus of documents that are self-tagged by front-line staff. I show that my technique is both intuitive in how documents are aggregated, and effective in collating types and frequency of medical errors in a healthcare organization.
    - It is well established that social factors play a role in overall health outcomes, however developing a statistical model that characterizes these relationships is an ongoing predicament. I developed statistical models for generalizing relationships between social determinants of health of a population and general medical risk factors. Furthermore, I use a sophisticated hexagonal binning approach to build intuitive local neighborhoods. This methodology ensures that the defined neighborhoods have similarly shared infrastructure and SDOH limitations. I apply machine learning algorithms, Spearman correlation, and other statistical tools to identify SDOH variables and the general health outcomes with which they are paired. This way, the models generated are not just retrospective, but prospective.
    - I address the research area of anomaly detection in two ways:
        - A variety of anomaly detection techniques already exist, however some of these methods lack a rigorous statistical investigation. This makes them ineffective to a practitioner as there is no appropriate signaling measure for detecting when an anomaly is present. In other cases, the methods were applied to only a small subset of networks that is not representative of the diverse representations of networks. Therefore, I evaluate the statistical properties of some of these network monitoring anomaly detection techniques while providing methodological improvements as these would bolster their effectiveness to a

practitioner. In this work, I carry out a systematic statistical evaluation of a suite of spectral methods for anomaly detection in static networks.

- Manufacturing enterprises are vulnerable to anomalies in the form of cyber-physical attacks. I developed a sensor-based side-channel technique for anomaly detection in a manufacturing process. I also create an appropriate quality-loss metric that properly distinguishes between conforming parts and anomalous parts. Furthermore, I develop an instrumented fixturing device that demonstrates the efficacy of my approach.

2 RELATED WORK

2.1 Pre-processing text data using machine learning algorithms

In the big data analytics framework, raw data initially goes through a pre-processing step. Although many pre-processing techniques exist, there are several challenges in pre-processing text data. An open research topic area is how to determine the meaning of ambiguous words or terms. The aim of word sense disambiguation is to autonomously contextualize an ambiguous word in a document or text (Pakhomov, Pedersen et al. 2005, Aronson and Lang 2010, Garla and Brandt 2012, Moon, McInnes et al. 2015). As the field of text mining increases and with increasing interest in applying Natural Language Processing (NLP) tools to different domains, the ability to automatically decipher the proper meaning of ambiguous words in these domains becomes more relevant (Chasin, Rumshisky et al. 2014). This is because, for complete textual analysis, it is necessary to understand the interpretation of words as they occur in their respective context (Pakhomov, Pedersen et al. 2005, Aronson and Lang 2010, Moon, McInnes et al. 2015).

The presence of ambiguous words in healthcare related documents is a challenging problem and hinders the applicability of many NLP tools (Pakhomov, Pedersen et al. 2005, Moon, McInnes et al. 2015). Free text in electronic health records (EHR) contain acronyms or abbreviations, informal sentence structure including typographical errors, variations in formatting between different organizations, or frequent use of sentence fragments (Pakhomov, Pedersen et al. 2005, Chasin, Rumshisky et al. 2014, Moon, McInnes et al. 2015). This is because the primary intent of these EHRs is for sharing between other healthcare practitioners and not for future textual analysis (Chasin, Rumshisky et al. 2014, Moon, McInnes et al. 2015). But to perform textual analysis, it is critical to first identify the proper meaning of words in the text. Specifically,

identifying the proper context of ambiguous terms in these EHRs could improve document retrieval, NLP analysis, and allow for sharing of EHR data across different healthcare centers or between experts in different domains (Naessens 2017).

Work has been done to disambiguate ambiguous words that occur in clinical text such as medical history reports, radiology notes, hospital visit records, and even text from clinical trials (Pakhomov, Pedersen et al. 2005, Garla and Brandt 2012, Chasin, Rumshisky et al. 2014, Moon, McInnes et al. 2015). This is primarily due to the existence of publicly accessible clinical text as well as compendiums of biomedical terminologies such as the Multi-parameter Intelligent Monitoring in Intensive Care (MIMIC II) database, Unified Medical Language System (UMLS), MEDLINE, among many others (Kim, Ohta et al. 2003, Chasin, Rumshisky et al. 2014). Pakhomov et al., introduced the use of publicly available information, such as Google's API, to disambiguate acronyms and abbreviations that occur in clinical text (Pakhomov, Pedersen et al. 2005). Specifically they identified 8 common acronyms and demonstrated that it was possible to determine the senses of these acronyms using publicly available tagged documents (Pakhomov, Pedersen et al. 2005). Chasin et al., compared a supervised machine learning approach (graphs and trees) with an unsupervised approach (topic modeling using clustering algorithms). Specifically, they used graph-based word sense disambiguation methods such as path-based and page-rank-based approaches that use the UMLS with accuracy between 40% - 50% and compare this approach to the Bayesian topic-modeling approach which performed better, reaching an accuracy of 66.9% (Chasin, Rumshisky et al. 2014). Savova et al., built and evaluated an open source natural language processing (NLP) application called the clinical Text Analysis and Knowledge Extraction System (cTAKES) (Savova, Masanz et al. 2010). The application can determine senses of words in clinical texts, is able to recognize nouns in a clinical text, and furthermore classify clinical texts into different disease groups among other tasks (Savova, Masanz et al. 2010).

Moon et al., compared multiple feature selection approaches as well as bag-of-words (BoW) for building supervised machine learning models to be applied in clinical text disambiguation (Moon, McInnes et al. 2015). They showed that the relatively simple technique of using bag-of-words performs comparatively well (Chasin, Rumshisky et al. 2014, Moon, McInnes et al. 2015). Also, studies have shown that the collocation and co-occurrences are stronger contributors to the performance of a disambiguation model than part-of-speech tagging (Pedersen 2000). Because of the effectiveness of these simple techniques in disambiguating clinical text, we

employ the same approaches to our patient safety event reports data set. For a more in depth discussion of current and previous approaches to word sense disambiguation of clinical text, refer to (Moon, McInnes et al. 2015).

2.2 Statistical modeling of text data using community extraction algorithms

Once the data is processed, open challenges still exist in the statistical learning step. One of the objectives of statistical learning is to develop a model that highlights general patterns in the data. Statistical modeling is widely utilized as researchers have created a variety of statistical models to explain everyday phenomena. However, new applications of big data with increasingly varied designs present interesting challenges. Consider the example of free-text analysis. There's a renewed interest in analyzing free-text narratives from sources such as online reviews, customer complaints, or patient safety event reports in a hospital.

In healthcare, there is an increase in the integration of Health Information Technology (HIT) in hospitals. Data collected in healthcare could therefore span a wide range of processes, from storing radiology lab results, to patient demographic information, to clinical notes on patients. In particular, big data analytics can be applied to free-text narratives contained in these HIT systems. Patient safety event reports (PSEs), captured in HIT systems are free text reports that aim to account for any concerns that a staff notices during the delivery of care to patients. In PSEs, front-line staff can submit a report about unsafe conditions, hazards, or serious safety events.

However, autonomously aggregating text documents into intuitive groups is an open area of research. Some researchers employ Latent Dirichlet Allocation (LDA), which is based on probability distributions of words in documents, to determine the groups documents belong to (Mehrotra, Sanner et al. 2013). Another area of research is using Latent Semantic Analysis (LSA) which relies on the spectral properties of a term document matrix derived from the documents to group these documents into similar groups (Cheng, Yan et al. 2014). Other methods rely on clustering algorithms whereby the corpus is also converted to a term document matrix and then projected to a lower dimension. Then a range of clustering algorithms, from K-Means (Steinbach, Karypis et al. 2000), Density-based spatial clustering of applications with noise (DB-SCAN) (Ertöz, Steinbach et al. 2003), High density-based spatial clustering of applications with noise (HDB-SCAN) (Jackson, Qiao et al.), and other variations of these clustering methods are applied. Hierarchal topic modeling approaches also exist that group closely related documents at each rung, producing a tree like link between similar document groups.

An assumption inherent in these topic modeling approaches is that a majority of the documents should belong to a group, and outliers are a few, if any. In my data exploration, I noticed that a range of documents, particularly in PSEs, describe unique events that do not belong to group. Therefore applying the topic modeling approaches that exist would force these documents to fit into one of the many communities it generates, resulting in communities that are not intuitive in interpretation. Community extraction, though, can handle these set of unique problems in classifying documents into communities. In community extraction, communities are extracted from the network via an iterative process (Chen and Saad 2012). In each step of community extraction, groups of nodes which have an intra connection density that is higher than their inter connection to the rest of the network are extracted in that iteration. The extraction process continues till there is no subset of the network with intra connection densities significantly higher than their inter connection densities.

2.3 Statistical modeling of general health outcomes using social factors

Social determinants of health (SDOH), are the social conditions in the environments in which people are born, live, learn, work, play, worship, and age that affect a wide range of health, quality-of-life outcomes and risks (Koh, Piotrowski et al. 2011). SDOH are also sometimes referred to as socioeconomic factors (Ferrer 2018) (Smith, Griffiths et al. 2018) (Pink and Allbon 2008). Organizations such as the MacArthur Foundation Network on Socioeconomic Status and Health (Williams, Costa et al. 2008), the Robert Wood Johnson Foundation (RWJF) Commission to Build a Healthier America (Foundation, Braveman et al. 2008), and the 2008 commission from the World Health Organization (WHO) (Organization 2008), have identified SDOH as critical to the health and well-being of a population. The WHO report has been instrumental in furthering the discussion of SDOH in recent years (Braveman, Egerter et al. 2011).

Prior to this report, it was well established that differing levels of poverty are linked to different health outcomes in a population (Marmot 2005, Organization 2008, Marmot, Friel et al. 2009, Braveman, Egerter et al. 2011, Marmot and Allen 2014). That is, a gradient of health outcomes exist from the worst-off in society to the most privileged (Marmot and Allen 2014). However, a variety of cases exist where these variations between populations far exceed effects applicable to poverty or genetic differences (Willett 2002, Wilkinson and Marmot 2003, Marmot 2005, Anderson, Robson et al. 2016). For example, while Costa Rica has a much lower gross national product (GNP) per capita, (< $10,000) than the U.S. (over $34,000), the country has a

similar life expectancy to the U.S., 77.9 years to 76.9 years respectively (Marmot, Friel et al. 2009). Similar trends are observable in Cuba and Greece, which both have a lower GNP measures than the U.S. yet comparable or higher life expectancy (Marmot, Friel et al. 2009).

Supporting the concepts of SDOH, other studies have shown that when people emigrate from their countries, they eventually imitate health risk outcomes of their new residence (Willett 2002). The Australian Institute of Health and Welfare in 2017 produced the Health Performance Framework Report which attributes a 34% gap in health outcomes between indigenous populations and non-indigenous population to SDOH (Pink and Allbon 2008, Anderson, Robson et al. 2016, Smith, Griffiths et al. 2018).

The academic study of SDOH has a long history. Conventional wisdom in health care has been that SDOH play a major role in health outcomes. As far back as the 17th century, pioneers such as John Graunt, Edwin Chadwick, and Friedrich Engels in England, Rudolf Virchow in Germany, among others explored the associations between certain living conditions and mortality rates (Ferrer 2018). However, few studies were published in this area until 1991 and ever since, there has been growing interest in understanding SDOH (Braveman, Egerter et al. 2011).

Although, research into clinical applications of SDOH are more recent. The WHO 2008 commission identified that SDOH are not sufficiently being accounted for in our health care system (Organization 2008, Organization 2008, Marmot, Friel et al. 2009). The commission identified several social conditions that account for disparities in health outcomes. In total, ten major SDOH were identified: social gradient, stress, early life opportunities, social exclusion, occupation, unemployment, social support, addiction, food security, and transportation (Marmot 2005, Organization 2008, Marmot, Friel et al. 2009, Marmot, Allen et al. 2012).

In conjunction with the report, recent policies in the U.S. have added to the growing interest in SDOH research for clinical applications. In the U.S, the 2010 Affordable Care Act incentivized providers and payers in the healthcare industry to move towards value based care (Alley, Asomugha et al. 2016). Furthermore, Accountable Care Organizations (ACOs) were created to address patient population health problems in a comprehensive way. Other major players in the U.S. health care market also played a role in driving SDOH research. The Centers for Medicare and Medicaid Services (CMS), which looks into developing innovative payment and service models, added a financial imperative that providers reduce cost and utilization (Alley, Asomugha et al. 2016). Institutions are now given greater latitude on how this quality of care is achieved

(Gillam 2008) (Oberlander and Laugesen 2015). In addition, the prevalence of electronic health records (EHRs), has provided health care institutions with increased data capacity for measuring SDOH, spurring the research further.

Previous researchers have aimed at creating a framework for capturing the breadth of SDOH factors in human society by addressing the problem from societal factors, and/or geographical extent. Authors in (Ferrer 2018) and (Braveman, Egerter et al. 2011) describe an upstream and downstream framework that encapsulates SDOH at an institutional level and their effects on individual's health and mortality. Social inequities (which are based on personal attributes such as gender, sex, immigration status) have an effect on access to economic and institutional resources. These factors indirectly influence social environments such as transportation infrastructure resources, availability of food options, and these then shape risk behaviors such as smoking, poor nutrition, and so forth. Furthermore, these risk behaviors over time will have an impact on a population's overall propensity to certain chronic diseases and/or injuries which then affect mortality rates (Ferrer 2018). Authors in (Smith, Griffiths et al. 2018) focus on the impact of various SDOH from a temporal perspective, i.e., effect of early life support, to education in the youth years, to housing, and income as the individual matures into an adult, and access to welfare at later stages in life on an individual's health outcome.

In many SDOH studies, researchers pair a particular SDOH with a specific set of health outcomes. Authors in (Kuruvilla, Schweitzer et al. 2014) showed that low educational opportunities translate to higher infant mortality rates. Researchers in (Klinenberg 2016) showed that married men are likely to live longer in comparison to their single counterparts and suffer from fewer chronic health issues. Authors in (Kushel, Gupta et al. 2006) connect housing instability as well as food insecurity to delays in seeking care, possibly leading to over-utilization of the emergency department. Authors in (Meyers, Cutts et al. 2005) pair effects of housing and food insecurity to under-nutrition among young children in the household. Authors in (Meltzer and Schwartz 2016) show that higher out-of-pocket expenses is tied to individuals postponing medical examinations for financial reasons. Researchers in (Gregory and Coleman-Jensen 2017) pair food insecurity SDOH with 10 major chronic conditions such as hypertension, coronary heart disease, hepatitis, stoke, cancer, asthma, diabetes, arthritis, chronic obstructive pulmonary disease, among others. Fowler et. al., show that food insecurity as measured by lack of supermarkets in high Supplemental Nutrition Assistance Program (SNAP) and Woman Infants and Children (WIC)

utilization neighborhoods results in decreased physical activity and depressive symptoms in that population. The organization, Feeding America, linked food insecurity to higher levels of diabetes and obesity (America 2011). Duncan et al., show the effects of early childhood poverty to specific health outcomes. Researchers in (Syed, Gerber et al. 2013) correlated decrease in ownership of a driver's license to frequency of missed appointments.

As we see, there has been extensive research into the effect of SDOH on health outcomes. However, in these research studies, SDOH are typically paired with a particular disease or group of communicable or non-communicable disease(s) (Kuruvilla, Schweitzer et al. 2014) (Klinenberg 2016) (Meyers, Cutts et al. 2005, America 2011, Gregory and Coleman-Jensen 2017). However, focusing on a specific set of health outcomes has some major limitations in SDOH application. For example, studies which pair food insecurity with hypertension (Gregory and Coleman-Jensen 2017) , housing with asthma (McNamara, Balaj et al. 2017), income with heart related problems, malnutrition, depression (Meyers, Cutts et al. 2005), or transportation barriers with diabetes (Locatelli, Sharp et al. 2017), ignore the larger at-risk populations, who might not have that particular disease but are also affected by SDOH. Inherent in the definition of SDOH are spatially shared, collective experiences that impact a community's health. Focusing observations on diagnosed cases of negative health outcomes overlooks nuances implied by the network of activities and arrangements, localized to specific neighborhoods that result in these locationally-significant health phenomena. Just as individual behaviors impact individual health, neighborhood attributes impact a community's health.

2.4 Statistical evaluation of an anomaly detection technique in network monitoring

A network consists of nodes which represent individual entities and relationships between nodes are represented as edges (Bader and Madduri 2008, Woodall, Zhao et al. 2017). Investigators in recent years have demonstrated that many phenomena can be represented as networks (Woodall, Zhao et al. 2017). These phenomena can span a multitude of fields such as: the power grid (Albert, Albert et al. 2004, Dahan, Sela et al. 2017, Woodall, Zhao et al. 2017) where nodes are power stations and edges the transmission lines, or a social network where nodes are individuals and interactions between individuals depicted as edges (Savage, Zhang et al. 2014, Mazrae Farahani, Baradaran Kazemzadeh et al. 2016), or gene sequencing where the nucleotides that make up DNA and RNA during transcription are represented as network motifs (Raulf-Heimsoth, Chen et al. 1998, Procter, Thompson et al. 2010, Cer, Bruce et al. 2011, Cer, Bruce et

al. 2012). Networks are therefore capable of visually and mathematically representing a myriad of fields (Woodall, Zhao et al. 2017).

For this reason, methodologies that can be applied to networks to identify abnormalities in these various fields is of significant importance. This is termed the anomaly detection problem where the primary aim is to identify a subset of a network that is behaving outside of normal conditions (Miller, Beard et al. 2015, Woodall, Zhao et al. 2017). For example, it is useful to a practitioner to identify over-burdened power plants in a power grid network (Dahan, Sela et al. 2017 , Woodall, Zhao et al. 2017), or for a security agency to identify a clandestine operation such as a terrorist cell in a large social network (Savage, Zhang et al. 2014, Mazrae Farahani, Baradaran Kazemzadeh et al. 2016), or to identify a series of abnormal proteins in a gene transcription process (Procter, Thompson et al. 2010, Cer, Bruce et al. 2011, Cer, Bruce et al. 2012). Anomaly detection therefore provides a tool-set that allows practitioners to detect unusual behavior in a wide variety of fields (Akoglu, McGlohon et al. 2010, Dahan, Sela et al. 2017, Woodall, Zhao et al. 2017). Typically, anomaly detection techniques focus on defining what conditions constitute a normal network and discriminating between anomalous nodes and non-anomalous nodes (Dahan, Sela et al. 2017, Woodall, Zhao et al. 2017).

As mentioned in the introduction, networks can either be *static* or *dynamic* and this designation affects the type of anomaly detection technique that can be used. This work focuses on *static* networks although future research can look into extending the benefits of this work to dynamic networks. The critical factors to consider in an anomaly detection problem are the size of the network, the size of the anomalous subgraph to be detected, and the types of anomalies that are of interest (Miller, Beard et al. 2015). For example, a small anomalous subgraph is harder to detect than a large anomalous subgraph in the same network (Miller, Beard et al. 2015). Also, the type of anomalous subgraphs to detect will significantly affect the efficacy of the proposed method (Miller, Beard et al. 2015). Anomaly detection techniques that are robust to these critical factors are, therefore, highly sought after by practitioners. Recently, investigators (Miller, Bliss et al. 2010, Singh, Miller et al. 2011, Miller, Beard et al. 2015) developed a suite of anomaly detection methods for static networks based on spectral properties (i.e., eigenvalues and eigenvectors) that are robust to these critical factors. In particular, the authors demonstrated the applicability of their methods for detecting different types of anomalous subgraphs that are in some instances smaller than 1% of the network size.

In (Miller, Beard et al. 2015) three spectral methods were proposed, namely the chi-square algorithm, the L_1 norm algorithm, and the Sparse Principal Component Analysis (PCA) algorithm. Of these, the Sparse PCA method has some significant limitations in its implementation. This method requires estimating the sparse matrix of an eigenspace which is an NP hard problem (Miller, Beard et al. 2015). Additionally, there was no significant improvement in performance in comparison to the other two methods (Miller, Beard et al. 2015). In the interest of time and space, this work restricts its analysis to the chi-square and L_1 norm algorithms, and the Sparse PCA method is not covered as it offers no benefits compared to the other two.

2.5 Side-channel based anomaly detection method with application in manufacturing

Current manufacturing enterprises are moving to become more interconnected with the rise of Industry 4.0 (Ralston, Graham et al. 2007, Dr. Ralf C. Shlaepfer 2014, Wells, Camelio et al. 2014). These systems continue to connect resources using the Internet of Things (IoT) which increases the vulnerability space for cyber-physical attacks. Occurrences such as the Stuxnet malware attack on Iran's nuclear facility (Albright, Brannan et al. 2010) and the cyber-physical attack on a steel mill in Germany (Lee, Assante et al. 2014) reflect successful cyber-physical attacks on production systems. The danger to these systems is the ease of malware spreading across the entire manufacturing enterprise. An infected USB stick was responsible for the spread of the Stuxnet attack and a compromised business network for the attack on the German steel mill (Albright, Brannan et al. 2010, Falliere, Murchu et al. 2011, Dr. Ralf C. Shlaepfer 2014, Wells, Camelio et al. 2014). In both cases, manufacturing operations were brought to a halt. These attacks reflect the urgent need for more secure interconnected manufacturing systems and robust inspection techniques to detect potential malicious tampering.

The interconnected manufacturing facility requires a means for part and process authentication as well as process shift detection to improve on traditional statistical quality control (SQC) monitoring methodologies (Wells, Camelio et al.). Piezoelectric transducers (PZT), composed of lead zirconate titanate, are impedance based non-destructive evaluation (NDE) methodologies that provide the benefits of in-process inspection with part authentication. The creation of unique identifiers for parts based on their electromechanical impedance signature profiles, provide the added part authentication and process verification steps needed for additional security measures in manufacturing (Sun, Chaudhry et al. 1995, Park, Sohn et al. 2003, Divsholi and Yang 2014).

PZTs have been used for non-destructive evaluation (NDE) of critical infrastructure including bridges, aerospace components, and precision parts (Sun, Chaudhry et al. 1995, Wong, Du et al. 2015, Köhler, Gaul et al. 2016). This is due to the coupled electromechanical characteristics of PZTs. It's application relies on using impedance signatures to detect the presence of a structural change (Park, Sohn et al. 2003). The general framework behind impedance based NDE methodologies involves the excitation of a structure via an actuator. The structure resists this excitation based on its physical characteristics such as its stiffness, mass, or damping properties (Park, Sohn et al. 2003, Wong, Du et al. 2015). The measure of its resistance is its impedance (Wong, Du et al. 2015). During impedance based NDE, the bonded actuator and sensors typically have a geometry and weight that is negligible on the observed signature (Park, Cudney et al. 2000, Wong, Du et al. 2015). Therefore, the observed signature is entirely determined by the physical characteristics of the structure, including defects or non-conformities that may exist (Vincent, Wells et al. 2015, Wong, Du et al. 2015).

The configuration of a PZT bonded to a structure can be viewed as spring mass damper system (Park, Cudney et al. 2000, Wong, Du et al. 2015). Equations are derived for a given frequency range by assuming that the structure under test can be approximated as a single-degree-of-freedom system, whereas the PZT wafer is handled as an axially deforming continuous system (Park, Cudney et al. 2000, Wang, Tehranipoor et al. 2008). Solving the wave equation for the spring mass dampener system results in the electrical admittance, Y(ω), shown in equation (1) (Sun, Chaudhry et al. 1995). This is the frequency dependent equation of the electrical admittance of the PZT. It highlights the relationship between the frequency range of excitement, structural impedance, and the resulting electrical admittance that is measured with the impedance analyzer (Sun, Chaudhry et al. 1995, Wong, Du et al. 2015).

$$Y(\omega) = i\omega\alpha(\varepsilon_{33}^{T}\,(1 - i\delta) - \frac{Z_s(\omega)}{(Z_s(\omega)+Z_a(\omega)}\, d_{3x}^{2} Y'^{E}_{xx}) \qquad (1)$$

where: Y is the electric admittance (inverse of impedance), Z_a the mechanical impedance of the PZT, Z_s is the mechanical impedance of the structure, Y_{xx}^{E} is the Young's modulus of the PZT at zero electric field, d_{3x} is the piezoelectric coupling constant in the x direction at zero stress, ϵ_{33}^{T} is the dielectric constant at zero stress, δ is the dielectric loss tangent of the PZT, and α is the geometric constant of the PZT.

An important aspect of the equation is the relationship of the mechanical impedance of the structure Z_s and mechanical impedance of the PZT, Z_a, to the electric admittance. When a PZT is

bonded to a structure, its mechanical impedance Z_a remains constant so it can be ignored (Park, Cudney et al. 2000). The equation above shows that a change in the mechanical impedance of the structure will directly affect the observed electric admittance and the resulting signature generated with the analyzer reflects the characteristics of the structure. Equation (1) also shows that this is dependent on the frequency of excitation ω (Sun, Chaudhry et al. 1995, Na 2017).

The approximation of PZTs as a spring mass damper system with applications in monitoring systems for NDE of a part was introduced by Liang et al. in 1994 and implemented by Sun et al. in 1995 (Park, Sohn et al. 2003). In these applications, a baseline signature profile is established corresponding to the structure's initial state (Bray and Stanley 1996, Vincent, Wells et al. 2015, Wong, Du et al. 2015). The structure is interrogated over time and the signature profile is compared to the baseline profile (Bray and Stanley 1996, Vincent, Wells et al. 2015). This allows observers to monitor the degradation of a structure over time. Sun et al. applied the concept to monitoring bridges by showing that PZTs were able to detect the looseness of bolts on a large prototype (Sun, Chaudhry et al. 1995). Park et al. and Paul et al. demonstrated that at high frequencies, it is possible to detect small or emerging damages in structures such as concrete slabs or beams (Park, Cudney et al. 2000, Paul and Jayaguru 2016, Na 2017). For larger structures, Divsholi and Yang showed that PZTs can be aggregated together for non-destructive evaluation (Divsholi and Yang 2014). Lalande et al. were able to detect abrasive wear and bending fatigue on complex precision parts such as gears (Lalande, Rogers et al. 1996). In all of these applications, the PZT is directly bonded to the structure being tested using a high stiffness adhesive (Park, Sohn et al. 2003). This approach limits the use of PZTs for NDE in a manufacturing process because: (1) direct bonding of a PZT to a part is undesirable as the removal of the PZT adds additional manufacturing operations: (2) the bonding agent could damage some parts: and (3) it leads to reproducibility concerns as requiring a new PZT for interrogating each individual part leads to additional variation between measurements.

One potential solution is to mount the PZT to a device that allows for the repeated use of the same PZT without damaging the part. Other researchers have also looked at alternative mounting methods. Na et al. and Silveira et al. used two magnets attached to each side of a beam via magnetic attraction. The excitation energy from the PZT is then transferred via the magnet to the structure and the impedance of the structure is measured. Na et al. observed sensitivity loss when these magnets were used instead of direct mounting on the structure. Although, they

concluded the loss in sensitivity is minimal if the part being interrogated is relatively thin as with some automobile parts or aircraft wings (Na, Tawie et al. 2012, da Silveira, Campeiro et al. 2017).

Researchers have also investigated the use of steel wires or metal foils to extend the reach of an excitation signal. The steel wires are bonded to the part using a stiff adhesive and the wires transfer the excitation energy from the transducers to the part and reflect back the impedance of the part. Although the investigators noticed a reduction in sensitivity consistent with the magnetic mounting approach (da Silveira, Campeiro et al. 2017). Furthermore, due to the requirement to bond the steel wires to the interrogated part, this methodology faces the same limitations as directly bonded PZTs when used in a manufacturing setting.

2.6 Existing gaps and proposed research areas

In this work, I investigate the different steps of the data analytics framework and propose improvements for each step, paired with practical applications, to demonstrate the efficacy of my methods. This work focuses on the healthcare and manufacturing industries, but the materials are broad enough to have wide applications across data analytics generally. My main contributions can be summarized as follows:

1. **Created a web-application that pre-processes text data using machine learning**: In the big data analytics framework, raw data initially goes through a pre-processing step. Although many pre-processing techniques exist, there are several challenges in pre-processing text data and I develop a pre-processing application that addresses some of these challenges. I apply machine learning to the problem of word-sense disambiguation to pre-process text data and create a web application to demonstrate the utility of my methodology.
2. In the next step of the data analytics framework, there are challenges in both statistical modeling and anomaly detection
    a. I address the research area of statistical modeling in two ways:
        i. **Designed a community extraction methodology for topic modeling of patient safety event reports:** There are open challenges in defining models to characterize text data. I introduce a community extraction model that autonomously aggregates text documents into intuitive communities/groups. I compare the effectiveness of my unsupervised document aggregation technique to a corpus of documents that are self-

tagged by front-line staff. I show that my technique is both intuitive in how documents are aggregated, and effective in collating types and frequency of medical errors in a healthcare organization.

ii. **Modeled overall health outcomes using newly defined methods for social determinants of health:** It is well established that social factors play a role in overall health outcomes however developing a statistical model that characterizes these relationships is an open research area. I developed statistical models for generalizing relationships between social determinants of health of a population and general medical risk factors. Furthermore, I use a sophisticated hexagonal binning approach to create intuitive local neighborhoods. This methodology ensures that the defined neighborhoods are more intuitive as they have similarly shared infrastructure and SDOH limitations. I apply machine learning algorithms, Spearman correlation and other statistical tools to identify SDOH variables and the general health outcomes with which they are paired. This way, the models generated are not just retrospective, but prospective.

b. I address the research area of anomaly detection in two ways:

i. **Investigated a sensitive spectral method for anomaly detection, identified critical shortcomings, and made improvements:** A variety of anomaly detection techniques exist already, however, some of these methods lack a rigorous statistical investigation thereby making them ineffective to a practitioner. I investigate a suite of spectral methods for anomaly detection. However these methods have limitations that impact their use in real-world applications. For example, there is no appropriate signaling measure for detecting when an anomaly is present. Also, these methods were applied to only a small subset of networks that is not representative of the diverse representations of networks. Therefore, I evaluate the statistical properties of some of these network monitoring anomaly detection techniques while providing methodological improvements as these would bolster their effectiveness to a practitioner.

ii. **Developed a side-channel technique for anomaly detection in a manufacturing process**: Manufacturing enterprises which are now more connected than ever are vulnerably to anomalies in the form of cyber-physical attacks. I developed a sensor-based side-channel technique for anomaly detection in a manufacturing process. I also create an appropriate quality loss metric that properly distinguishes between conforming parts and anomalous parts. Furthermore, I develop an instrumented fixturing device to assist in integrating this sensor based anomaly detection methodology to a manufacturing setting.

3 CREATED A WEB APPLICATION THAT PRE-PROCESSES TEXT DATA USING MACHING LEARNING

3.1 Overview of chapter

The goal of this study was to apply and evaluate multiple supervised learning approaches to disambiguating words in patient safety event reports. Although work has been done to disambiguate clinical texts, it is unclear if the same approaches apply to patient safety event (PSE) reports since PSEs have distinctive characteristics that separate them from other healthcare related texts. We identified 11 ambiguous words or terms from a corpus containing 69,000 patient safety event (PSE) reports. Terms were culled via a two-step review process. Over 2000 PSE reports were coded for the training of the machine learning models. Three machine learning algorithms were evaluated: Multinomial Logistic Regression, Random Forest, and Support Vector Machines. We incorporated two approaches when building our machine learning models: (1) Bag-of-words and (2) inclusion/exclusion of stop-words. It should be noted that a paper on this topic has been submitted to the Journal of Healthcare Informatics (Komolafe *et. al., 2018*).

3.2 Significance of research into word-sense disambiguation

There are significant challenges associated with disambiguating terms in PSEs. The informal style of the text means we have to account for typographical errors, use of sentence fragments, as well as variations in style between front-line staff reporting on an incident. As other researchers have shown that simple techniques such as bag-of-words are quite effective in word sense disambiguation, we incorporated this approach to our data set. We believe that the models

we develop will provide sufficient performance in disambiguating ambiguous words and can be a valuable source for improving textual analysis of PSE reports.

3.3 Patient safety event reports

Patient safety event reports are free text reports that aim to account for any concerns that a staff notices during the delivery of care to patients. In patient safety reports, front-line staff can submit a report about unsafe conditions, hazards, or serious safety events. Therefore, analyzing patient safety event text could highlight the causes of many of these unsafe conditions or serious safety events. In fact, recent data has shown that medical errors that occur during the administration of healthcare related services are currently the 3rd leading cause of death in the U.S. (Makary and Daniel 2016, Kavanagh, Saman et al. 2017, Naessens 2017). Furthermore, because death certificates in the U.S. do not include human factor or system related errors, the mortality figure above is likely under-representative of the true impact of medical errors (Makary and Daniel 2016).

Fortunately, there is a large corpus of patient safety event data that has been gathered from HITs around the U.S. However, patient safety reports present particular challenges in word sense disambiguation that are not typical of clinical documents. This is because patient safety event reports are generally less formal, less structured, and more prone to variability because the intent of these reports is to give a personal account of a healthcare service delivery event. The topics in these reports range from describing system issues, communication topics, or even concerns with fellow colleagues. Furthermore, PSE reports are written by different front-line staff (e.g., physicians, nurses, technicians) which adds an extra layer of variability.

The following short example from our patient safety event database which contains reports obtained from the MedStar Union Memorial Hospital HIT illustrate some typical problems such as lexical ambiguity:

> This **am**, **resident** sustained injury to **rh**. **NP** stayed by **pt** waiting on **er**.

In the simple example above: the term, '**am**', could either refer to the term ante-meridian or the verb to-be or an abbreviation of the word, 'American': the word '**resident**' can mean either resident-patient or resident-doctor: '**rh**' here means right-hand but could be interpreted as right-half: '**NP**' can be interpreted as no-problem or nurse-practitioner: '**pt**' could mean physical-therapist or patient: and '**er**' can mean emergency-room or a type of extended-release drug.

Any analysis on these PSE reports will benefit from a more accurate representation of the report. This work investigates the application of a simple word sense disambiguation technique to

PSE reports, bag-of-words (Moon, McInnes et al. 2015). We apply supervised machine learning techniques to a corpus containing 2,000 patient safety reports. Some of the main contributions in this study are (1) Identify common ambiguous words that occur in patient safety reports and (2) Compare and evaluate the performances of three machine learning algorithms to this corpus, Multinomial Logistic Regression, Random Forest, and Support Vector Machine (3) Demonstrate the effectiveness of a simple WSD technique to PSEs. This work is significant as any task performed on PSEs such as automatically aggregating similar types of incidents and determining root cause of events that relies on machine learning methods stands to directly benefit from a more accurate representation of the reports. Additionally, it will allow for the proper textual analysis of reports by experts in other domains.

The rest of this work is organized as follows: We discuss the PSE dataset used to train and test our machine learning models. We then elaborate on the feature extraction methods critical in developing our machine learning models. We show the results of the different machine learning models used and explain their drawbacks, if any. The performance of our best suited model shows promise in using machine learning models to disambiguate patient safety reports.

3.4 Method

3.4.1 Dataset

2,000 patient safety reports from a large multi-hospital healthcare system sampled from a three year period was used in this study. Each patient safety event report includes structured data elements (e.g., department, event date, severity of event) and a brief free text description of the incident. The analysis in this study will focus on the free text portion of these reports.

3.4.2 Identifying ambiguous words

We first identified ambiguous words or terms and narrowed down which terms to disambiguate using a two-step review process. First, Taber's Online Medical Dictionary was used as a resource to generate a list of medical abbreviations that could have multiple meanings. For instance, a term found in Taber's Medical Dictionary is the abbreviation 'IM' which is shorthand for 'intramuscular' (Venes 2017). 'IM' is also found in the PSE data as a mis-spelling of the term 'I'm' which is the contracted form of the phrase 'I am'. We also ensured that we captured acronyms and abbreviations in all their available forms, including with and without periods, slashes, and spaces. Taber's Online Medical Dictionary was chosen as a source for retrieving ambiguous medical related words because it is a respected online freely available resource (Venes 2017).

Secondly, we included ambiguous terms that we encountered when manually reading through PSEs. In total, 18 ambiguous terms from the 69,000 PSE reports were identified in the first stage of the review process, 10 from Taber's Online Medical Dictionary: *am, co, er, hr, im, or, pa, pm, pt, yo,* and 8 more from manually identifying ambiguous terms in our PSE reports: *cd, dc, fu, left, ir, resident, right, sat*.

In the second stage of the review process, reports containing these 18 ambiguous terms were selected. For each ambiguous term, we retrieved at least 100 reports where the proper sense of the term can be identified. Seven terms were either not sufficiently represented in the original 69,000 PSE reports or had only one interpretation and were therefore removed from the final term set. The 11 identified terms and their ambiguous meanings are shown in Table 1 along with their percent usage in the 2,000 sense tagged reports.

Terms identified fell into two categories: (1) two letter acronyms or abbreviations (7 terms) or (2) manually curated terms from the PSE corpus that could have multiple meanings (4 terms). The 'Reports Analyzed' column of Table 1 includes a count of how many documents were coded for a particular term. Some coded documents had overlapping terms.

Table 1. Ambiguous terms and their respective interpretations based on conversations with healthcare professionals

	Terms	**Reports Analyzed**	**Meanings and Percentage Usage (%)**
1	e.r.	212	emergency_room (87.3), extended_release (12.7)
2	pt	191	patient_person (86.9), physical_therapist (13.1)
3	resident	190	patient_person (28.9), resident_doctor (71.1)
4	or	197	operating_room (20.3), or (79.7)
5	am	197	anter_meridiem (71.1), verb_to_be (28.9)
6	hr	190	heart_rate (20.0), hour (80.0)
7	sat	188	o2_sat (41.0), to_sit (59.0)
8	left	199	left_side (49.7), to_leave (50.3)
9	right	202	correct (5.0), right_now (22.2), right_side (72.8)
10	dc	138	discharge_medication (20.3), discharge_patient_person (73.2), district_of_columbia (6.5)

11	co	132	carbon_oxygengas (12.1), co_person (9.1), complaining_of (78.8)

3.4.3 Pre-processing

We pre-processed the corpus to convert any permutation of dates, times, or dosages into the words “date”, “time”, and “dose” respectively. This is done to ensure that during the feature extraction phase, these words are understood by the algorithm as a common feature rather than distinct dates, times, or doses. We also remove special characters and convert all other instances of numbers appearing in the report into the word "num". We also compare the performances of each of the models with English stop-words included or excluded from the text. Stop-words from the R-package tm were used with the exception of terms that also belonged in the list of ambiguous words (Feinerer and Hornik 2012). Preprocessing and modeling were performed using the R software package (Team 2014).

3.4.4 Algorithm design and evaluation

In our methodology, we apply three common supervised machine learning algorithms to disambiguating words in free text data: Multinomial Logistic Regression, Support Vector Machines, and Random Forest algorithms. These three algorithms were specifically chosen as they have been used in multiple classification applications and have literature that documents their performance (Xu and Jelinek 2004 , Adler, De Alfaro et al. , Tufféry 2011,Moon, 2015 #1034, Moon, McInnes et al. 2015). The general framework for supervised machine learning algorithms in WSD is illustrated in Figure 3 (Moon, McInnes et al. 2015). First the senses of ambiguous terms in 2,000 reports were identified. From the 2,000 coded reports, we progress to the feature selection stage as in Figure 3 where the aim is to identify which features are important for building the machine learning classification models. Important features in our study are words neighboring the term of interest. The hypothesis being that words within a certain distance of the term would be relevant in determining the sense of the word. As there is no exact agreement on what this distance should be, the number n of neighboring words is varied from 1 neighboring word, to 5 neighboring words.

For example, in the text below:

This **am**, **resident** sustained injury to **rh**. **NP** stayed by **pt** waiting on **er**.

If the term of interest is **resident** and we include stop-words, then for $n = 1$, we would include the words am and sustained when building out our feature matrix. If $n = 3$, the terms *this, am* as well as *sustained, injury, to* will be included in the feature matrix. When $n > 5$, the performance of the models is similar to using the entire sentence and for completeness, this study also includes using the entire sentence in feature selection. For each n a different model was created and the following sections report on the performances of the different models. We also employ the bag-of-words approach as previous studies have shown this simple approach is quite effective in word sense disambiguation (Moon, McInnes et al. 2015). We ignore using syntactic parsing or word embedding in this study as our aim is to compare the effectiveness of the bag-of-words approach to a PSE dataset.

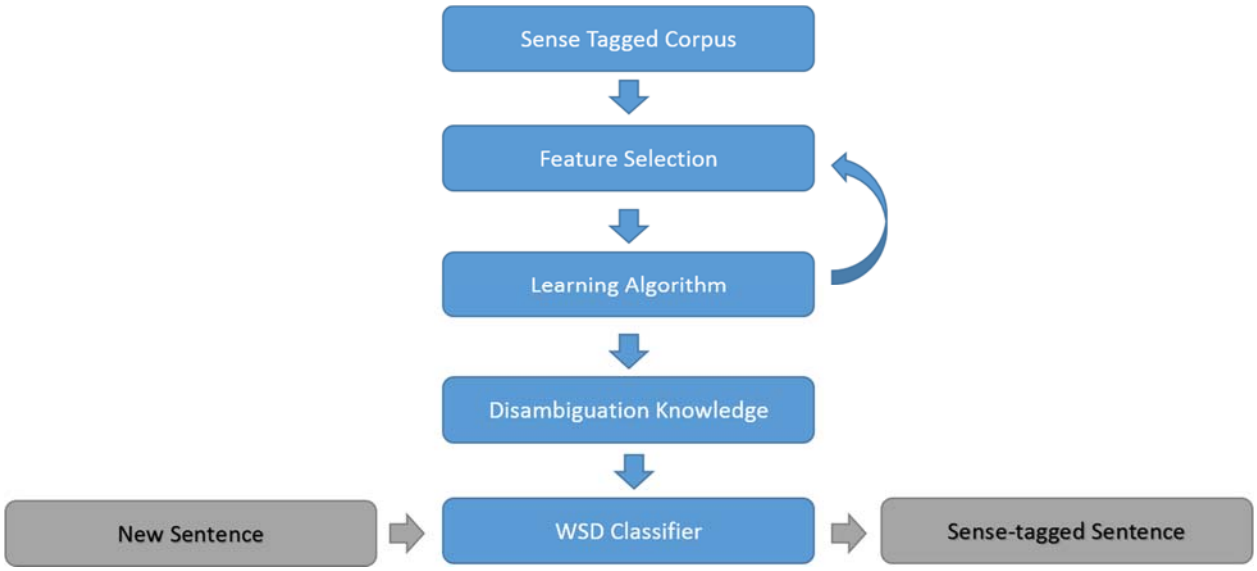

Figure 3. Machine Learning Algorithm Framework

Next, 75% of the 2,000 sense tagged documents is used to train the model. The remaining 25% of the tagged data is used to eventually test the model. The separation of datasets is stratified to ensure there is an equal representation of classifications in both the training and testing datasets. A 5-fold cross validation on the 75% training data set is also performed. We also compare the importance of including or excluding common English stop-wordsin building the model.

Once the important features for each model has been tested and validated, disambiguation knowledge has been achieved. The model's performance is evaluated using a testing data set and the F1 score calculated as in Equation (3) (Chang and Sung 2005).

Table 2. Description of multiclass confusion matrix

	Predicted Class 1	Predicted Class 2 …	Predicted Class n
True Class 1	$T_1 \cap P_1$	$T_1 \cap P_2$	$T_1 \cap P_n$
True Class 2	$T_2 \cap P_1$	$T_2 \cap P_2$	$T_2 \cap P_n$
:	:	:	:
True Class n	$T_n \cap P_1$	$T_n \cap P_2$	$T_n \cap P_n$

(1)

$$P_r = x_1 * \frac{T_1 \cap P_1}{\sum_{i=1}^{n} T_i \cap P_1} + x_2 * \frac{T_2 \cap P_2}{\sum_{i=1}^{n} T_i \cap P_2} + \cdots + x_n * \frac{T_n \cap P_n}{\sum_{i=1}^{n} T_i \cap P_n}$$

(2)

$$R_e = x_1 * \frac{T_1 \cap P_1}{\sum_{i=1}^{n} T_1 \cap P_i} + x_2 * \frac{T_2 \cap P_2}{\sum_{i=1}^{n} T_2 \cap P_i} + \cdots + x_n * \frac{T_n \cap P_n}{\sum_{i=1}^{n} T_n \cap P_i}$$

(3)

$$F1 = \frac{2}{\frac{1}{P_r} + \frac{1}{R_e}}$$

where *Pr* is the precision, *Re*, is the recall, *F1* is the F1 score for that model, T is the true or correct classification value and P is the predicted classification value. Based on the testing results, the best performing models are wrapped into a word sense disambiguation application (WSD). This application should accept a new sentence and replace the terms in these sentences with their most likely interpretation. We further discuss such an application in the discussion section.

3.5 Results

In this section, we describe the results of the machine learning algorithms for word sense disambiguation.

3.5.1 Model performances and evaluations: stop-words included

Figure 4 as well as Table 3 below demonstrates the performances of the 6 different feature extraction methods used for each model type. Stop-words are included in this evaluation. Each

feature extraction method pertains to the number words adjacent to the term of interest. We observe that in Figure 4, the Support Vector Machine (SVM) model performs the worst when $n = 2$. Its average F1 score is 0.824 with a standard deviation 0.153 across all 11 terms. The MTML model performs the best when the entire sentence is included during feature selection. The average F1 score is 0.879. Both the Random Forest (RF) and Support Vector Machine (SVM) algorithms have relatively similar performance measures across the different feature extraction methods used in this study. Both models perform the best when the five adjacent words to the term of interest are used for disambiguating the sense of that term. The average F1 score for both the RF and SVM models is 0.865 with a standard deviation of 0.108 and 0.871 with a standard deviation of 0.110 respectively. A breakdown of the precision, recall, and F1 scores for each of the terms is presented in the Appendix.

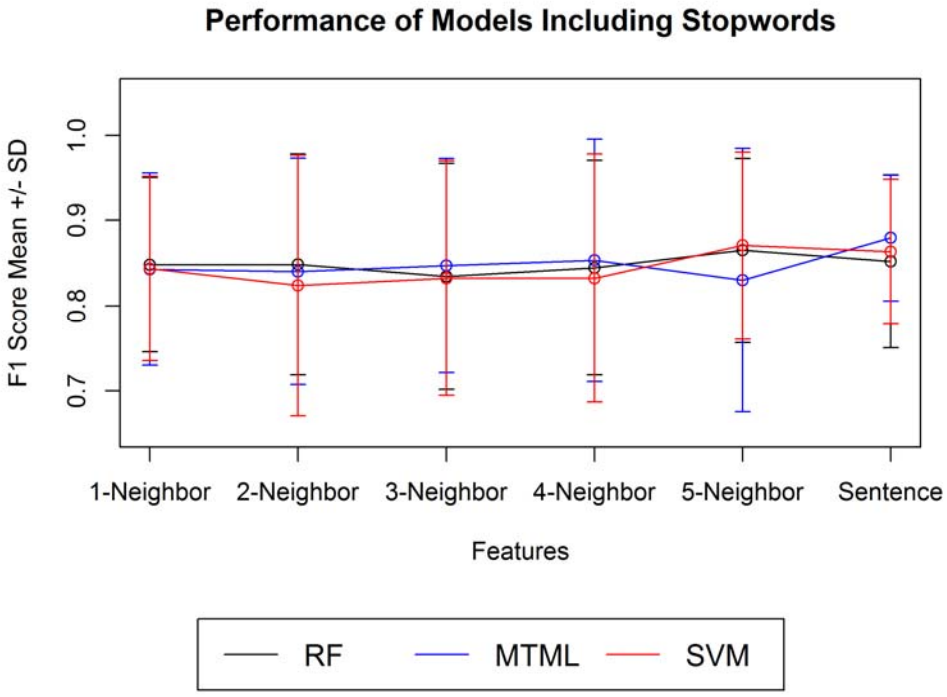

Figure 4. Different machine learning algorithm performances, stop-words included

Table 3. Average F1 score and standard deviation values for the different feature extraction methods and different machine learning algorithms. Stop-words included in text

Feature	MTML avg F1	MTML std F1	RF avg F1	RF std F1	SVM avg F1	SVM std F1

1-Neighbor	0.843	0.113	0.848	0.102	0.844	0.108
2-Neighbor	0.840	0.133	0.848	0.130	0.824	0.153
3-Neighbor	0.847	0.126	0.834	0.133	0.832	0.138
4-Neighbor	0.853	0.142	0.845	0.126	0.832	0.146
5-Neighbor	0.830	0.154	0.865	0.108	0.871	0.110
Sentence	0.879	0.074	0.852	0.102	0.863	0.085

3.5.2 Model performances and evaluations: stop-words excluded

Similarly, we evaluate the performance of the different algorithms when stop-words common in the English dictionary are excluded from our corpus. We use English stop-words from the R package tm (Feinerer and Hornik 2012). Words that are part of our list of terms are excluded from the stop-word list. The full list of stop-words is provided in the Appendix. The Multinomial Logistic Regression model performs the best as in the previous case with an F1 score of 0.879 and a standard deviation of 0.080. This is illustrated in Table 4 and Figure 5. Just as in the case when stop-words are included, the RF model attains its best performance when $n = 5$ with an F1 score of 0.860 and a standard deviation of 0.116. The SVM model achieves the same performance for $n = 5$. However, when the entire sentence is used, the standard deviation in performance between the terms improves, from 0.104 to 0.073. It should be noted that in both cases when the entire sentence is used, both the SVM and MTML models achieve the lowest variance in performance across all 11 terms.

Table 4. Average FI scores for all methods

Feature	MTML avg F1	MTML std F1	RF avg F1	RF std F1	SVM avg F1	SVM std F1
1-Neighbor	0.831	0.116	0.822	0.119	0.838	0.104
2-Neighbor	0.835	0.099	0.825	0.12	0.822	0.131
3-Neighbor	0.821	0.128	0.846	0.127	0.842	0.132
4-Neighbor	0.857	0.098	0.84	0.165	0.819	0.164
5-Neighbor	0.854	0.113	0.861	0.116	0.86	0.104
Sentence	0.879	0.08	0.839	0.152	0.86	0.073

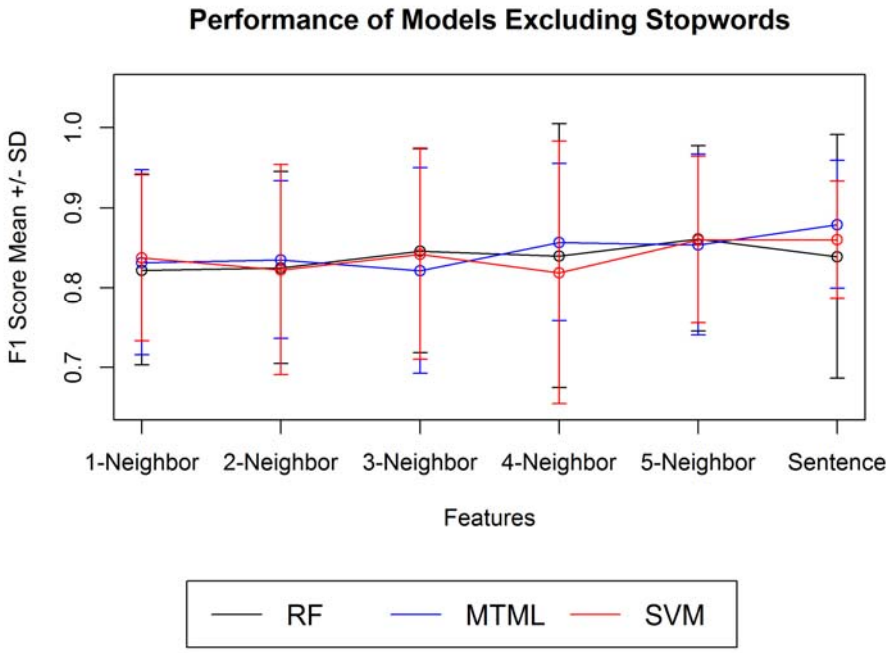

Figure 5. Different machine learning algorithm performances, stop-words excluded

3.6 Discussion and conclusion

Our aim in this study is to identify and replace ambiguous words, which we call 'terms', in these patient safety reports, with their most likely interpretation. We apply two simple, well studied techniques: bag-of-words and collocation, which has been shown to be quite effective in disambiguating clinical texts (Pedersen 2000, Moon, McInnes et al. 2015). We also compare the effectiveness of including or excluding stop-words while building our machine learning models. From the 69,000 available reports, we identified 11 ambiguous words and evaluated the performances of three different algorithms: multinomial logistic regression, support vector machine, and random forest. The multinomial logistic regression performed the best although we should note that the run time for this algorithm is on average about twice as long the random forest algorithm or support vector machine algorithm. Hence, the algorithm might not be suitable for larger datasets. We recommend using SVMs for larger data sets as this algorithm performed relatively well in comparison to the multinomial regression model and with faster run times.

This work has shown that the same approaches that were effective in disambiguating clinical texts can also be extended to patient safety reports. The use of stop-words or excluding stop-words surprisingly had little effect on the performances of these algorithms. This is possibly due to the extensive use of sentence fragments in PSEs. Because front-line staff described incidents in phrases rather than complete sentences, the effect stop-words could have played is significantly reduced. For example, the sentence: “The **pt** was waiting in the **er**" is commonly replaced with this sentence fragment:

pt waiting in **er**

where “pt" and “er" are the ambiguous terms to consider. Common stop-words that would traditionally have been added to the bag-of-words such as “The” or “was” are excluded.

Future research directions will involve exploring the use of synonyms or hypernyms in extracting features for our machine learning algorithm. We applied some hypernyms in this application whereby we replaced permutations of dates to the word "date", permutations of different doses in the report to the word "dose", and the same for time. We can look into other words that have broader meanings and use these broader meanings in our feature extraction stage. In addition, there were many typographical errors or inconsistencies in how words were represented in our corpus. An example is the word "approximate" which was also written as "approx", "appr", "approxmt", etc. Misspellings and stemming patient safety specific words (i.e., scripts and prescriptions or meds and medications) will also be helpful in the analysis of patient safety events. As syntactic parsing and word embedding are commonly applied to WSD, we can evaluate the performances of different models if these techniques are included.

Finally, future work will look into making these models publicly available through a public application programming interface (API) or web application that can help healthcare systems better analyze their data. In addition, we intend to include features into the web application that allows end users to provide feedback on the current models, particularly, if the models misinterpret the proper sense of a term. Furthermore, there is an opportunity to leverage the input of end users in identifying other ambiguous words or senses of a particular term. These developments have the potential of significantly improving the number of terms identified and the accuracy of the model. The end result is a word sense disambiguation classifier that can be applied to numerous healthcare related fields.

4 DESIGNED A COMMUNITY EXTRACTION METHODOLOGY FOR TOPIC MODELING OF PATIENT SAFETY EVENT REPORTS

4.1 Overview of chapter:

Patient safety event reports are free-text narratives written by the front-line staff. These narratives describe incidents whereby a healthcare service delivery did not go as expected. During these instances, the front-line staff witnessing the incident can document his/her perspective of the events that occurred. Therefore, aggregating similar PSEs has the potential to give insights into trends of the different types of medical errors healthcare organizations encounter. There is a significant amount of variation between documents because these narratives do not have to follow any specified format. For example, documents describing similar events can vary drastically in their word usage, vocabulary, document length, and prevalence of grammatical errors. Therefore, a methodology that is robust to these variations within documents, yet able to aggregate similar documents into communities is significant.

4.2 Significance of research

A variety of topic modeling approaches exist for aggregating similar documents into communities. However, many of these methods assume that a majority of documents belong to a community. The result of these topic modeling approaches when applied to a unique data source, such as patient safety event reports (PSEs), results in communities that are not intuitive in their meaning. This is because PSEs are written by various staff members to describe incidents they have noticed during healthcare service delivery. There are a variety of instances that a front-line staff observes, which do not naturally belong to a group or community of documents. In fact, there is a possibility that most of the reports are unique observations and should not be grouped together. Other topic modeling approaches, which rely on the presumption that most of the documents should be placed in a community are likely to group disparate documents together. Therefore, a methodology is needed, that will aggregate only similar documents into communities and leave out documents describing a unique instance or observation.

4.3 Method

In our paper, we first create a manually curated dictionary of commonly misspelled words in our corpus and replace them with their proper spelling. To do this, we extracted terms that

appeared in 2 or more documents and correct any misspelled terms. As PSEs typically contain information such as the date an event occurred, the time it occurred, or dosage of a particular medication, any permutation of a date, dosage, or time is replaced with the words "date", "dose", and "time" respectively. This is because, the exact time an event occurred or the exact dosage of a medication is irrelevant for our analysis.

For example, this sentence:

"On Dec. 13 at 5PM resident was prescribed 2mc/mg of oxycotine" is converted to

"On date at time resident was prescribed dose of oxycotine"

Special characters are removed, except for periods and all other numbers are removed from the text. To ensure that words with similar morphologies are presented as the same, a stemming function from the R package, "**tm**" is used to convert words to their root.

The entire corpus is converted to a term-document matrix where rows represent terms and columns represent documents. The weighting of the terms in our term document matrix is critical to any fuure analysis. We use the term frequency-inverse document frequency approach to weigh our terms. Once we have a weighted term document matrix, we apply Latent Semantic Analysis for dimension reduction. Previous studies have demonstrated the applicability of LSA for document retrieval (Turney 2001, Dumais 2004). LSA has the ability to handle obstacles prevalent in natural language processing and analysis such as presence of synonyms and polysemy. This paper only contains a brief description of LSA, for a more detailed description of the approach, see Foltz et. al., (1998) (Landauer, Foltz et al. 1998).

Gollub and van Loan (1989) demonstrated that any rectangular matrix can be converted to a product of three matrices using the concept of singular value decomposition, or SVD as it is commonly referred to. For a term-document matrix $\boldsymbol{X}$ of *m* terms and *n* documents with rank *r*, its SVD product can be written as

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{\Sigma}\boldsymbol{D}^T$$

Where $\boldsymbol{X}$ is the *m X n* term-document matrix, $\boldsymbol{T}$ is a *m X m* matrix whose columns are the orthogonal eigenvectors of $\boldsymbol{X}\boldsymbol{X}^T$ where we denote $\boldsymbol{X}^T$ as the transpose of the matrix $\boldsymbol{X}$**.** The matrix $\boldsymbol{D}$ is a n X n matrix whose columns are the orthogonal eigenvectors of $\boldsymbol{X}^T\boldsymbol{X}$ and $\boldsymbol{\Sigma}$ is a *m X n* diagonal matrix whose diagonals are $\sqrt{\lambda_i}$ where λ corresponds to the eigenvalues of $\boldsymbol{X}\boldsymbol{X}^T$ and $1 \leq i \leq r$ and 0 everywhere else. The eigenvalues of $\boldsymbol{X}\boldsymbol{X}^T$ are the same as the eigenvalues of $\boldsymbol{X}^T\boldsymbol{X}$**.** The values $\sqrt{\lambda_i}$ are called the singular values of $\boldsymbol{X}$.

The implementation of LSA used in this work is a low rank approximation of the SVD. For this, we find a positive integer, $k \leq r$ such that it closely approximates the term document matrix. The value *k* is selected such that it minimizes the error between the original matrix $\boldsymbol{X}$ and its low rank approximation $\boldsymbol{X_k}$. This is achieved through the following steps:

Since $\lambda_i \geq \lambda_{i+1}$, setting $\lambda_{i+1} = 0$ if it is close to zero will not significantly affect the original matrix $\boldsymbol{X}$**.** We therefore find a k where $1 \leq k \leq r$ such that it minimizes the difference in the Frobenius norm between $\boldsymbol{X}$ and $\boldsymbol{X_k}$**.** If k = r, then the difference in the Frobenius norm is 0 but if k << r, we have a low rank approximation of our matrix that is also easy to manipulate. By keeping only the k columns or entries for each of our matrices, we obtain $\boldsymbol{X_k}$ and furthermore a low rank approximation of both terms and documents. Therefore, we have

$$\boldsymbol{X_k = T_k \Sigma_k D_k^T}$$

Where we only keep the k columns of matrices $\boldsymbol{T}$ so $\boldsymbol{T_k}$ is a $m \; X \; k$ matrix**,** $\boldsymbol{D^T}$ so $\boldsymbol{D_k^T}$ is a $k \; X \; n$ matrix and $\boldsymbol{\Sigma_k}$ is a diagonal $k \; X \; k$ matrix.

Finally, we generate a network of documents by creating a similarity matrix from the matrix $\boldsymbol{D_k^T}$**.** In this paper, the similarity matrix is created by calculating the correlation between pairs of documents, resulting in a $n \; X \; n$ correlation matrix. The correlation matrix serves as our adjacency matrix for the community extraction phase.

4.4 Applying extraction algorithm to network of documents

Most community detection methods aim to partition a network into communities with the goal of maximizing the number of edges within communities and minimizing edges between communities (Zhao, Levina et al. 2011). This traditional framework assumes that all nodes belong to some community. However, there are many other scenarios whereby some nodes do not belong to any particular community and forcing these nodes into a community will distort the community detection results. For example, let's assume we have a network of high school students where links between students signifies that these students participate in similar extra-curricular activities. Applying some of the traditional community detection algorithms to this network will result in unsatisfactory results. This is because some students naturally do not participate in any extra-curricular activity and therefore do not belong to a community. However, these community detection algorithms will force these nodes to one of the formed communities.

Community extraction can handle these types of networks. Zhao et. al., applied community extraction to some well-studied networks such as the karate club network (Zachary 1977), school

friendship network (Hunter, Goodreau et al. 2008), and political books network (Newman 2006). Their results show that community extraction performs better when compared to other popular community detection methods such as modularity using the approximate eigenvector solution (Newman and Girvan 2004, Newman 2006), fitting a block model via Markov chain Monte Carlo (Nowicki and Snijders 2001), or using the latent position cluster model (Handcock, Raftery et al. 2007). In this paper, we use the community extraction method proposed by Zhao et. al (Zhao, Levina et al. 2011).

We describe a network graph *G* as composed of vertices *V* and edges *E, G = (V, E)*. The total number of vertices in a network graph *G* gives us the network size *N*. That is, *N* = |V|. Also the number of edges in a network graph is *M, M* = |E|. We consider only non-overlapping communities in this paper, therefore once community extraction is applied to a network *G* , the partition results in two distinct sets, V_1 and V_2 where $V_1 \cap V_2 = \emptyset$ and $V_1 \cup V_2 = V$.

A network can also be represented as an *N X N* adjacency matrix referred to as **A**, where its elements are $\mathbf{A}_{ij}$ and *i,j = 1,2,...,N*, $\mathbf{A}_{ij}$ = (-1,1) making it a weighted network. The adjacency matrix **A** is equal to the correlation matrix of $\boldsymbol{D_k^T}$**.** Communities are extracted one at a time with the criterion of extracting a set of nodes with the sum of its weights largest within that set and smallest between the set and its complement (Zhao, Levina et al. 2011). We will call this set of extracted nodes *S,* and its complement, S^c.

The objective function we are therefore maximizing in each iteration step is below (Zhao, Levina et al. 2011):

$$\widetilde{W}(S) = |S|\,|S^c|\left[\frac{O(S)}{|S|^2} - \frac{B(S)}{|S|\,|S^c|}\right]$$

$$O(S) = \sum_{i,j \in S} A_{i,j}, \qquad B(S) = \sum_{i \in S, j \in S^c} A_{i,j}$$

The term O(S) is twice the weight of the edges within *S* and *B(S)* represents the weights from the set S to the rest of the remaining network. In large sparse networks, particularly as in our application, a small community *S* could result in a large $\widetilde{W}(S)$ value, the term $|S|\,|S^c|$ serves to ensure that sufficiently sized communities are extracted at each step as very large communities or very small communities will be penalized. This is because the term, $|S|\,|S^c|$ is maximized at $|S| = \frac{N}{2}$

In our paper, the community extraction algorithm is repeated till only a small subset of nodes, 30 nodes or less, are left in the network and this was sufficient for our application. Zhao et. al., propose a stopping criteria only for a network that can be represented by the block model (Zhao, Levina et al. 2011). Future works will investigate a more appropriate stopping criteria.

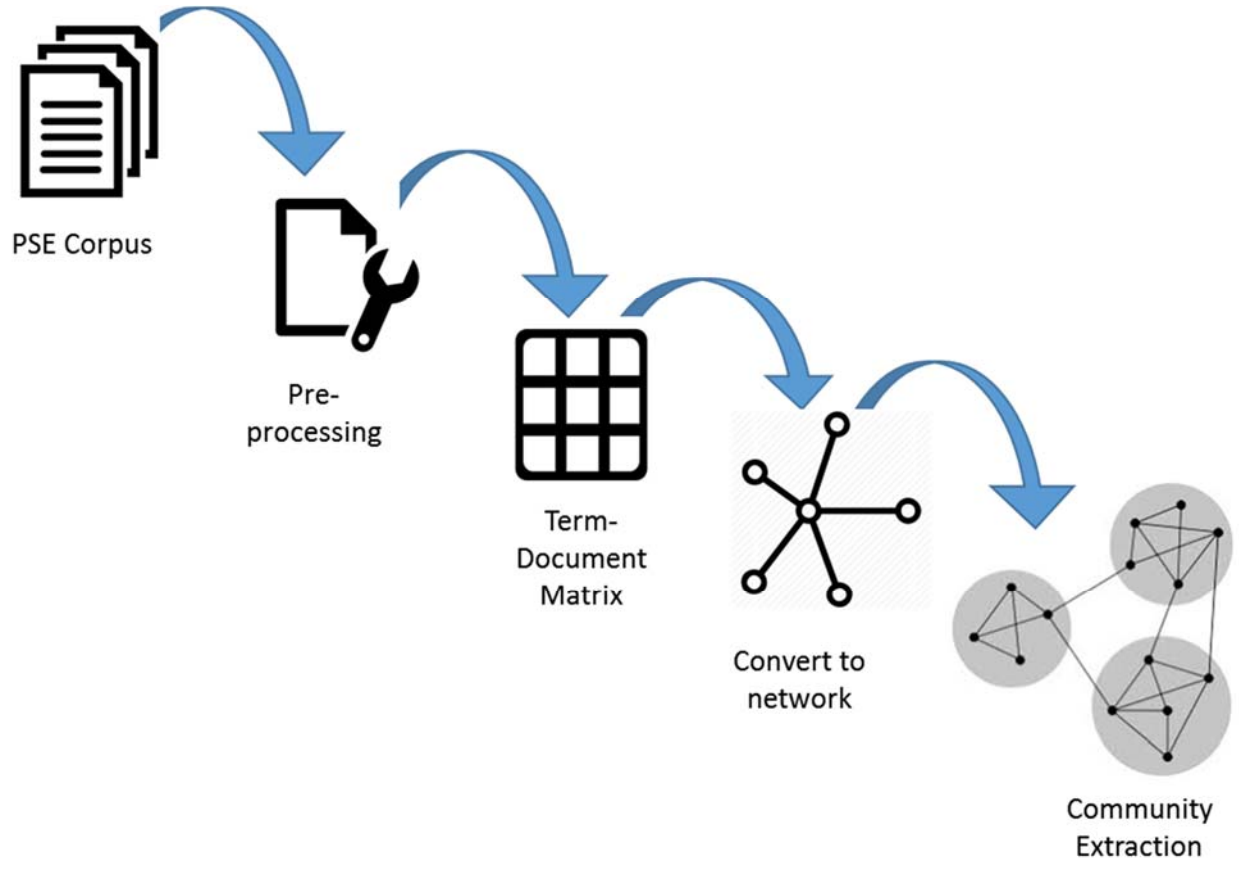

Figure 6. Framework for community extraction of PSE corpus

To maximize the objective function, we implement the *tabu search* maximization technique which is a local optimization technique based on label switching (Beasley 1998, Glover and Laguna 2013). In this optimization technique, a string of binary values representing nodes in either community S or S^c is passed to the tabu search function (Beasley 1998, Zhao, Levina et al. 2011, Glover and Laguna 2013). The function tracks which nodes have been switched, ensuring that they are not switched again until a certain number of iterations have passed, making these nodes, "tabu". To guard against being trapped at a local maxima, the algorithm is run with random label assignments each time. In practice, we observed run times of the order $O(n^2)$ where n is the size of the corpus. Our original PSE corpus is 2,072 documents, and running one iteration of the tabu

search algorithm on the entire corpus takes over 120 hours. One alternative is splitting the entire corpus of 2,072 documents into chunks of 200 documents or chunks of 400 documents. We observed run times of about 22 hours and 44 hours when partitioned into sizes of 200 and 400 respectively. However, it is crucial to knit similar communities in each partition of 200 or 400 back together.

Partitioning the entire document will also result in some communities being arbitrarily split up. We also developed a methodology for combining similar communities from different partitions. Our methodology relies on the correlation matrix of the entire 2,072 corpus. We compare pairs of communities across the different partitions and combine communities that have a combined density greater than some threshold.

We denote $S_{a,p}$ as the identity of a community extracted during the implementation of our algorithm. The integer, a, refers to the iteration number at which the community is extracted in that partition. The integer, p, refers to the partition the community belongs to. Where $1 \leq a \leq x$ with x representing the number of communities extracted for that partition and $1 \leq p \leq y$ where y is the total number of partitions for that particular implementation. Therefore, to establish if two extracted communities, $S_{1,1}$and $S_{4,2}$ originally belong to the same community, we compare each of their densities, $D_{a,p}$ to their combined density, $D_{(1,1),(4,2)}$.

$$D_{1,1} = \frac{1}{|S_{1,1}|^2} \sum_{i,j \in S_{1,1}} A_{i,j}$$

$$D_{4,2} = \frac{1}{|S_{4,2}|^2} \sum_{i,j \in S_{4,2}} A_{i,j}$$

$$D_{(1,1),(4,2)} = \frac{1}{|S_{1,1}| * |S_{4,2}|} \sum_{i \in S_{1,1}, j \in S_{4,2}} A_{i,j}$$

In this paper, two communities are combined together if $D_{(1,1),(4,2)} > 0.85 * D_{1,1}$ and $D_{(1,1),(4,2)} > 0.85 * D_{4,2}$.

4.5 Results

Our PSE reports are tagged by the user with options available from a drop-down menu for a front-line staff to tag a PSE report. Reports are tagged with both a general event description, and there are 20 options to select from in our report, and 187 specific event descriptions which are sub-categories of any one of the general event descriptions. From our analysis, these event tags do not

properly describe the report. To illustrate this observation, we create a correlation matrix for the different event tags. Specifically, we use the general tag “Medicaton Fluid” and investigate the congruency in Medication Fluid reports tagged by frontline staff with our methodology. As the correlation matrix in some instances has over 150 cells, we present our results using heatmaps. If the tags are descriptive enough, then we would expect the diagonals of the correlation matrices to be of a similar hue and the off diagonals should be a different hue. This would suggest that front-line staff are tagging similar documents with similar tags. However, if the heatmaps do not follow this pattern, then it suggests that the tags available to the front-line staff are not descriptive enough for each report type. We also compare these heatmaps to our community extraction results.

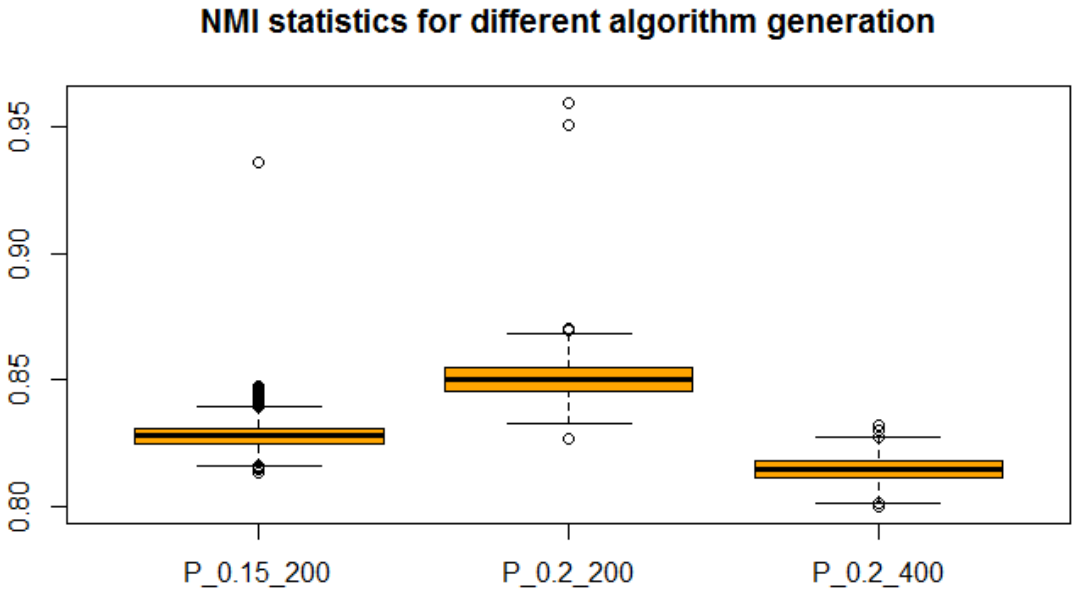

Figure 7. Boxplot showing NMI distribution between 100 randomized simulations for each communication methodology investigated (threshold at 0.15 with partition at 200, threshold at 0.2 with partition at 200 and threshold at 0.2 with partition at 400)

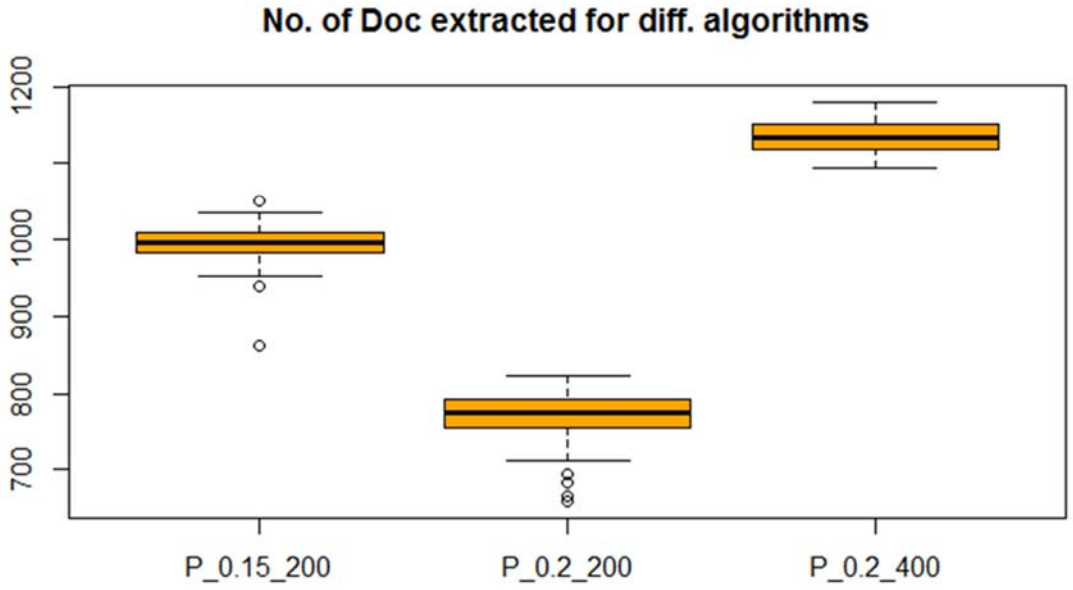

Figure 8. Boxplot showing number of documents extracted between 100 randomized simulations for each communication methodology investigated (threshold at 0.15 with partition at 200, threshold at 0.2 with partition at 200 and threshold at 0.2 with partition at 400)

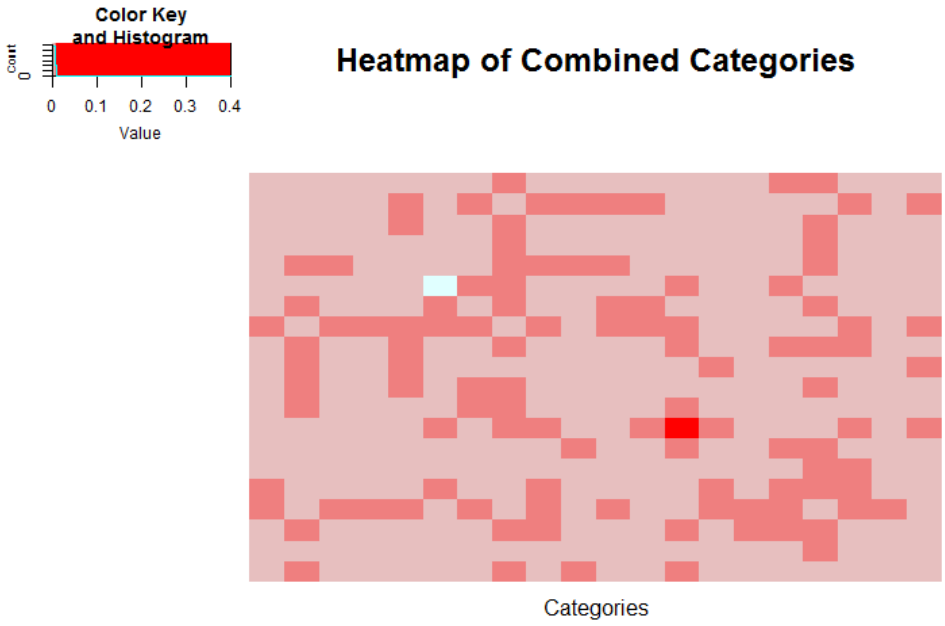

Figure 9. Heatmap of Combined event types generated from 20 X 20 correlation matrix of documents that fall into this tag

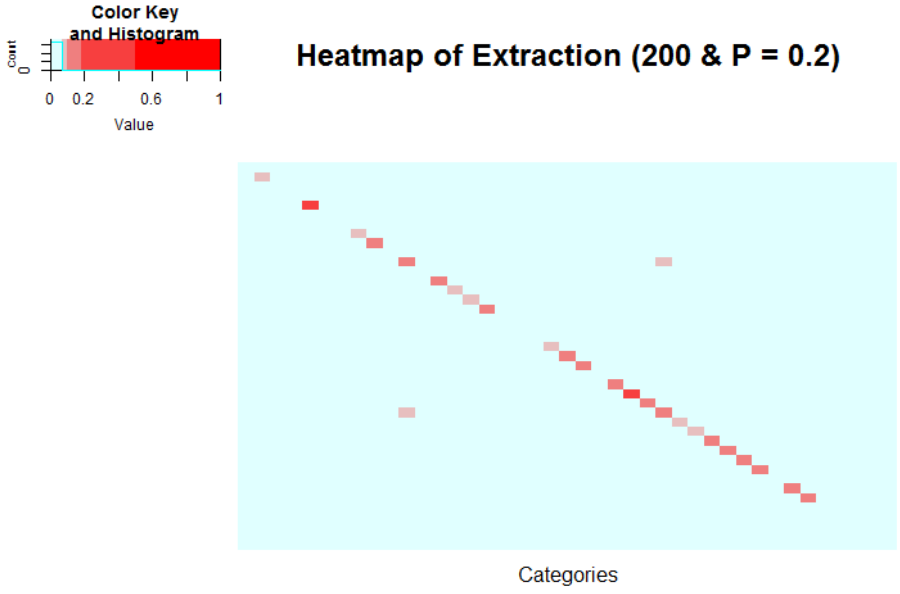

Figure 10. Heatmap of communities generated from 41 X 41 correlation matrix of documents that fall into the respective communities after community extraction is applied. Partition (200) with correlation threshold set at 0.2

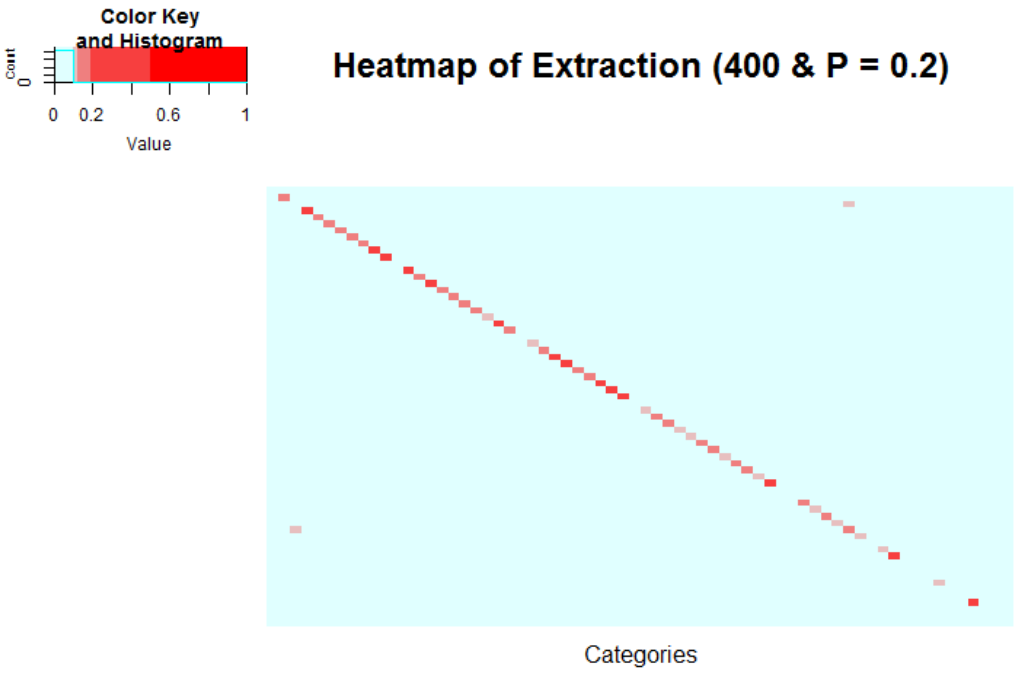

Figure 11. Heatmap of communities generated from 65 X 65 correlation matrix of documents that fall into the respective communities after community extraction is applied. Partition (400) with correlation threshold set at 0.2

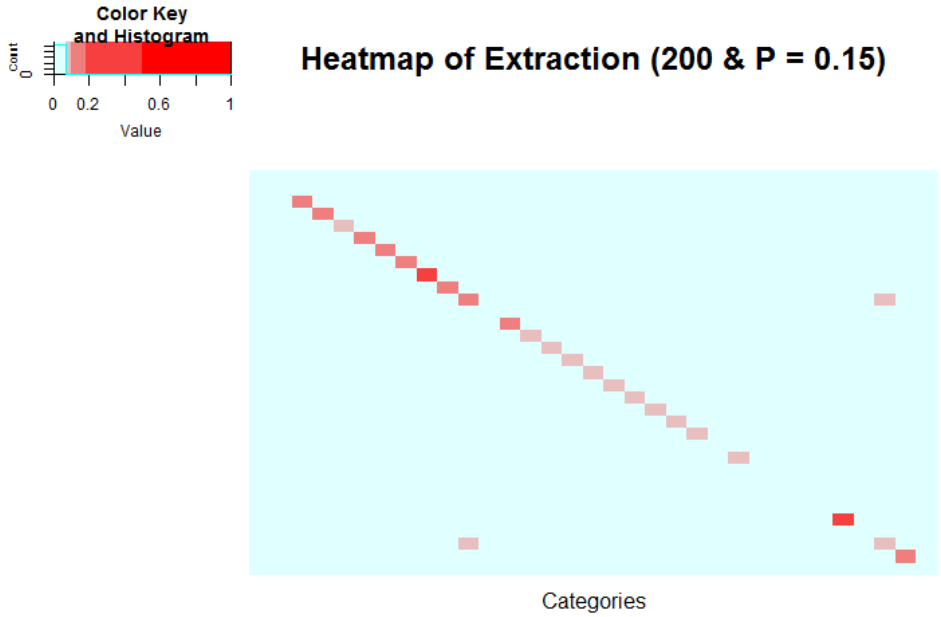

Figure 12. Heatmap of communities generated from 33 X 33 correlation matrix of documents that fall into the respective communities after community extraction is applied. Partition (200) with correlation threshold set at 0.15

4.6 Discussion

Overall, the results show that community extraction is a viable methodology for topic modeling of text data. In this work, we apply community extraction to 1773 patient safety even reports and compare our unsupervised community extraction method to tagged documents by frontline staff. Although there are many approaches to applying community extraction, we explore three methods. There is general agreement in the methodologies. However, we noted that increasing the threshold from 0.15 to 0.2 reduces the number of overall documents extracted but the documents are more similar overall. Secondly, to combat the concerns with run time, we proposed different partitioning values, (200, 400) and we observed that a higher partition results in a higher number of overall documents extracted but the run time is almost doubled.

5 MODELED OVERALL HEALTH OUTCOMES USING NEWLY DEFINED METHODS FOR SOCIAL DETERMINANTS OF HEALTH

5.1 Overview of chapter

Social determinants of health (SDOH), are the social conditions in the environments in which people are born, live, learn, work, play, worship, and age that affect a wide range of health, quality-of-life outcomes and risks (Koh, Piotrowski et al. 2011). There is general consensus that these factors impact individual health. This chapter discusses our methodology for identifying critical SDOH characteristics. I also introduce a novel dasymetric mapping method to define our local neighborhoods. Furthermore, I describe how I created prospective statistical models that can be used to predict individual health outcomes for the population at large. It should be noted that a paper on this topic has been submitted to The Lancet Journal (Komolafe *et. al 2018*).

5.2 Significance of research into statistical modeling of overall health outcomes using social determinants of health

SDOH have attracted a lot of attention from the healthcare community in recent years and they span a wide range of areas, from economic well-being to food insecurity, housing, transportation among others. In many SDOH studies, researchers pair a particular SDOH with a specific set of health outcomes (Kuruvilla, Schweitzer et al. 2014, Klinenberg 2016). However, we know that the effect of income is not related only to heart related outcomes. Transportation barriers

limit the ability to have regular preventative care, for all patients, regardless of whether they have a heart or lung disease. Similarly, food insecurity or poor housing affects a range of health concerns. Therefore, a study that focuses on pairing SDOH with general health outcomes, rather than a specific set of diseases, is beneficial. In this work, I investigate the effect of SDOH on medical risk factors and behavioral patterns such as emergency department (ED) utilization and body mass index (BMI). These risk factors aim at capturing both the at-risk population and the possible range of health complications that can arise from SDOH. Looking at health outcomes through general features brings together different aspects of the health care system.

Secondly, many data sources in existing SDOH research are wholly comprised of participant responses to survey questions. The data and results are therefore reliant on the recall and/or attentiveness of a participant. Furthermore, relying on individual measurements is costly and difficult to obtain. Rather, it would be beneficial to use data sources which are publicly validated, credible, and statistically sound. Lastly, current research areas like (Pickett and Pearl 2001, Sampson, Martins et al. 2018) use census blocks or administrative districts to define these neighborhood boundaries, however, this methodology of sampling is not designed for and does not account for human interaction with the neighborhood. This approach crosscuts neighborhood boundaries and infrastructure systems and do not account for the local interactions of neighborhoods and their community. Therefore defining neighborhoods that are more representative of social conditions people live in is critical for SDOH research.

In this work, we study SDOH effects at the neighborhood level. Previous studies like (Pickett and Pearl 2001, Sampson, Martins et al. 2018) use census blocks or administrative districts to define these neighborhood boundaries, however, these methodologies have some limitations. For example, census sampling areas (census tracts) in the U.S. are drawn to encapsulate about 4000 households (Schlossberg 2003). However, this methodology of sampling is not designed for and does not account for human interaction with the neighborhood. This approach crosscuts neighborhood boundaries and infrastructure systems such as roads, bridges, rivers, which impact access and utilization of health care resources. In essence, these administrative lines do not account for the local interactions of neighborhoods and their community. Furthermore, these methods for neighborhood catchment area are subject to being redrawn in the future, making neighborhood specific SDOH results irrelevant in future studies.

The main contributions in this paper are summarized as follows:

- As with previous studies, we use similar features for defining our SDOH, however, we are getting the data from the federal government, large healthcare organizations, and other credible organizations which are both reliable and statistically sound
- In our study, we use a sophisticated hexagonal binning to create our local neighborhoods. This methodology ensures that our defined neighborhoods are more intuitive as they have similarly shared infrastructure and SDOH limitations. This is more pertinent to studying SDOH as it reflects the inequities present in how social privileges are distributed among populations while also providing the added benefit of masking patient clinical data (Koh, Piotrowski et al. 2011). We discuss further our hexagonal grid system methodology in section 2.
- On the medical side, what we are looking at are general risk factors and behavioral patterns which are not restricted to a set of diseases. This is also more informative because SDOH will affect health care outcomes in general
- We also have very different approaches in connecting SDOH factor with general health outcomes. We apply machine learning algorithms, Spearman correlation and other statistical tools to identify SDOH variables and the general health outcomes with which they are paired. This allows us to build models that are not just retrospective, but prospective.

We demonstrate that given SDOH factors such as economic well-being, transportation, housing, and food insecurity for a particular demographic, we can transcribe a particular health outcome risk score. This is crucial as we show that given only the location of an individual, we can transcribe certain health risks based on the economic well-being, transportation, housing, and food insecurity of his or her nearest neighbors.

The rest of the paper is as follows. In Section 2 we go over the data sources used in this paper as well as the data wrangling methods. We also describe our hexagonal binning methodology in more detail. In section 3, we briefly highlight our findings for the four thematic SDOH areas we explore in this study. In section 4, we go into details of these findings. In section 5, we conclude by discussing the impact of these findings.

5.3 Data sources

The data sources for building our models are shown below.

- Patient health data: Patient electronic health records from two major health care organizations for two different regions in Ohio are used in this study. Healthcare provider in the Cleveland region (Cuyahoga County) supplied us with 161,000 Medicaid patient electronic health records. Electronic Health Records (EHR) contain medical information of a patient such as lab results, Body Mass Index (BMI), presence of chronic diseases, among others. In the Toledo region (Lucas County), we obtained 273,000 patient EHRs. However, these are not Medicaid patients.
- Patient claims data: We obtained over 730,000 claims records for the 161,000 Medicaid patients in Cuyahoga County. Claims data is recorded when a healthcare provider or institution requests some form of reimbursement from a payer for services rendered. Claims data contains at a minimum information such as: general patient demographic information, diagnosis and procedure codes, as well as where the service was rendered. We use this information to ascertain if a visit is to the emergency department or not.
- GIS data: National Land Cover Database data which is aggregated by the United States Geological Survey among a consortium of other federal agencies (Multi-Resolution Land Characteristics (MRLC) Consortium which is comprised of the National Oceanic and Atmospheric Administration, the U.S. Environmental Protection Agency, the U.S. Department of Agriculture (USDA) National Agricultural Statistics Service, the U.S. Forest Service, the National Park Service, the U.S. Fish and Wildlife Service, the Bureau of Land Management, NASA, and the U.S. Army Corps of Engineers) for the geographic area of responsibility.
- American Community Survey (ACS) data: contains survey information for multiple SDOH and aggregated by the United States Census Bureau.
- Business information: Latitude and longitude information on businesses in a region as well as type of service offered. Business information is retrieved from Google's Application Program Interface (API).

We describe the results of our study as well as briefly highlight the distinguishing aspects of our model building and evaluation methodologies. To allow for comparison between different SDOH and various health outcomes, we convert our results to risk scores. These risk scores can vary from 1 - 5, with 1 being a hex that is least at-risk and 5 corresponding to a hex that is most at-risk. The score of 0 is assigned to non-residential hexes.

5.3.1 Hexagonal methodology

In this study, hexagons are used as the unit of measurement for a particular set of SDOH and a given health outcome. With hexes, we can assign a risk score to a uniform area. Furthermore, this allows us to mask patient information in our analysis. Hexes measure 0.1 km^2 in size.

Dasymetric mapping, specifically hexagon grid systems, have been applied to other geo-spatial projects with the most recognizable application being Uber's ride sharing application. Hexagons allow us to apply comparable risk scores to areas of equal sizes. They are also the closest polygon shapes to a circle, i.e., uniform distance from the center to the edges, yet they still allow for complete tessellation of a land area. Furthermore, we account for residential and non-residential land areas in our dasymetric mapping, such as bodies of water, highways, and bridges, among others. We use the National Land Cover Database provided by the United States Geological Survey (USGS) to define residential and non-residential areas. These attributes allows us to define specific neighborhoods based on homogeneous spatial location.

To illustrate the scale at which we analyzing SDOH factors, we show a census block in the north eastern corner of Toledo, Ohio and a zoomed in portion of that census block in Figure 14 with the hexes that comprise it.

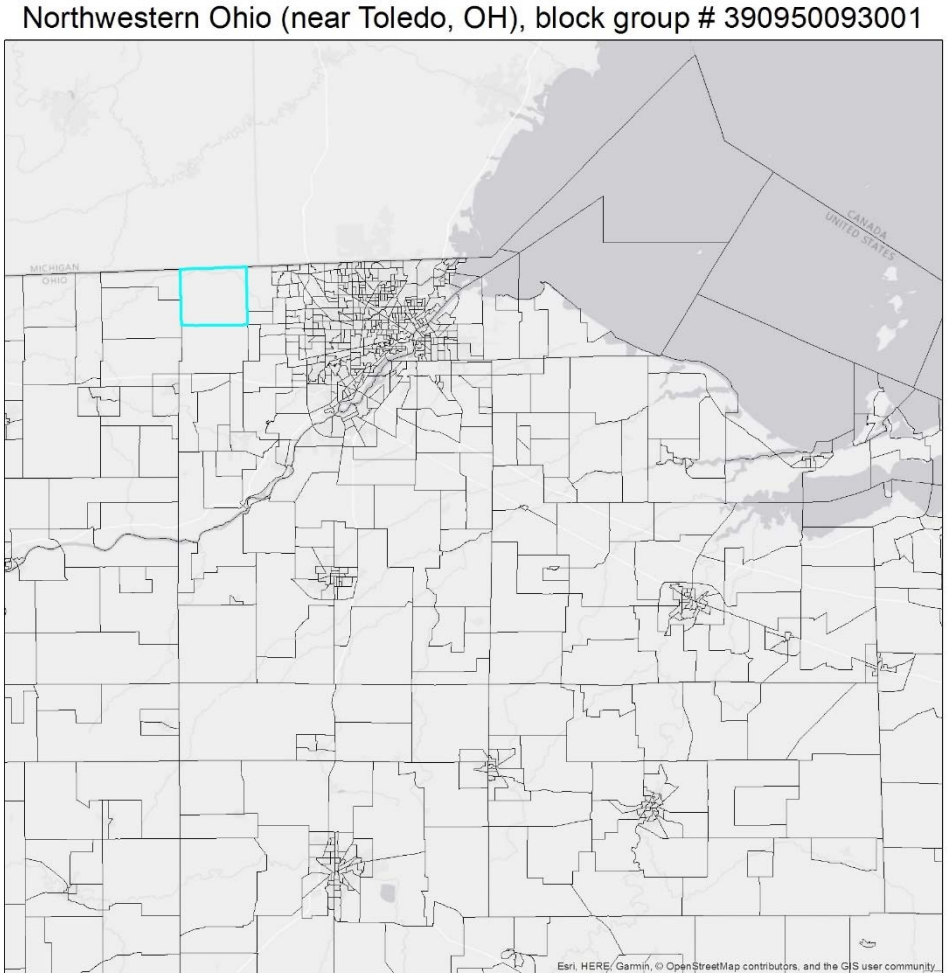

Figure 13. A census block in the Northern Toledo, Ohio region highlighted in blue

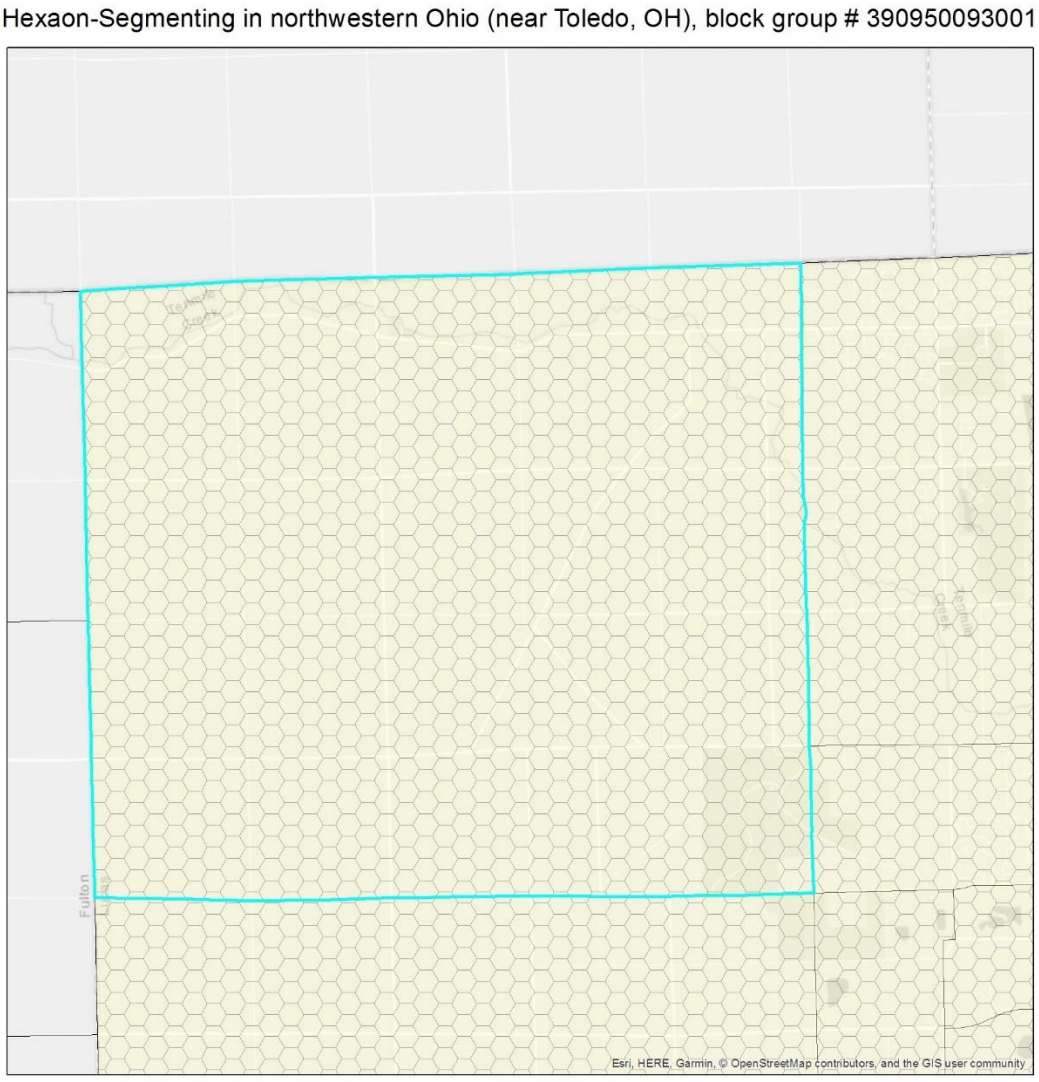

Figure 14. Zoomed in portion showing the hex map that comprises the census block

We analyze two populations for our study: Lucas County near Toledo, Ohio and Cuyahoga County near Cleveland, Ohio. We note that patient clinical information for our analysis in Cuyahoga is derived from Medicaid patients in the County, and this could lead to bias in our analysis. The hex representation for the two counties as it relates to the housing SDOH is shown below in Figure 14 for purposes of illustration. For the housing SDOH, we pair hexes housing related SDOH covariates with their predicted annual ED visit utilization. In our study, we calculate a risk score, on the scale of 1 - 5, from least at-risk population to most at-risk population. The cut-offs for each score are based off of input from physicians and health care personnel. For example, for the behavioral health outcome of average annual ED visits, hexes that have over 2.5 average ED visits a year are given a score of 5. The figure below shows a hex map with associated housing

risk scores. We see that there is a spatial continuity in risk scores. As neighborhoods that are next to each other, have similar risk scores.

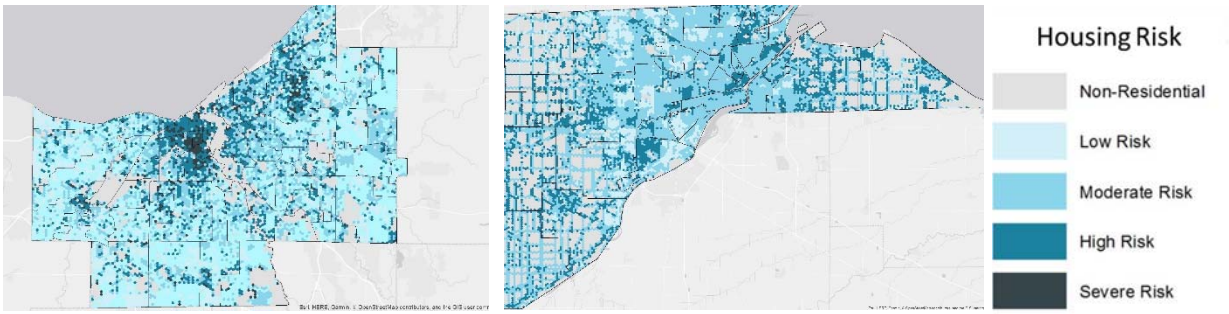

Figure 15. Left: Cuyahoga County hex map of housing, Middle: Lucas County hex map of housing, Right: Legend

5.3.2 Data wrangling methods for SDOH factors

We talk about the different data wrangling methods as well as statistical tools used for each of the thematic SDOH areas explored in this paper.

Moving Window: We use moving window plots to smooth the effects of discrete data in our analysis.

Distance Matrix and binning: We calculate distances from the center of hexes to various business locations for future binning.

Random Forest: We use a random forest regression algorithm to create predictive models. The hyper-parameters tuned in our regression models are: number of trees, minimum number of samples required to be at a leaf node, the maximum depth of the tree, and minimum number of samples required to split an internal node.

5.3.2.1 Economic well-being

The claims data for Cuyahoga County consisted of 730,923 Medicaid claims records and 11 variables. We are interested in CMS place of service (POS) codes. The major POS codes are 11, 12, 19, 20, 21, 22, 23, and 81. These POS codes allow us to identify two main categories for visits, preventative care and emergency. Preventative care, or "good", visits occur at a physician's office or an outpatient facility, such as an urgent care office. Emergency, or "bad", visits are emergency department visits or inpatient hospital stays. We define good visits as POS codes 11, 19, 20, and 22, and bad visits as POS codes 21 and 23.

First, we removed 313,268 claims that had a zero POS code. Overall, this left 417,655 visits of which 270,418 (about 65%) are good visits, 98,859 (about 24%) are bad visits, and the final 48,378 (about 11%) are considered neither. The ratio of bad visits to good visits is 0.37.
For a rough comparison, from the Center for Disease and Control (CDC) data we know that the annual number of visits to the emergency department is 141.4 million, whereas the annual number of physician office visits is 990.8 million. Therefore at the national level, the ratio of ED visits to Physician office visits is around 0.143. The ratio in our claims data is much worse than this national benchmark, which is unsurprising because these claims are for Medicaid patients. Also, the national benchmark is a rough proxy and might refer to a different set of CMS codes. Then we map the claims data to the hex data, and calculate the number of good and bad visits from each hex.

5.3.2.2 Transportation

We use similar claims data as that used for economic well-being. However, our SDOH of interest is the distance between a particular hex center to a nearby pharmacy or clinic. We obtain clinic and pharmacy business information (longitudes and latitudes) for the two Ohio counties, Cuyahoga (12,074 hexes with 1886 hexes being non-residential) and Lucas counties (16,524 hexes with 4767 hexes being non-residential). We removed non-residential hexes during model generation. We use Cuyahoga County demographic information to build our model, for transportation and other thematic areas. The number of pharmacies within a certain distance from the center of a particular hex is used as the features for input into our model. This is created using a binning methodology. In our final feature matrix, we created four bins, number of pharmacies less than 1 mile from the center of that hex, number of pharmacies within 1 and 2 miles of the hex, between 2 and 5 miles of the hex and greater than 5 miles from the hex center. All observations of the feature matrix are standardized using the formula

$$\frac{x}{1+x}$$

before we input into our machine learning model. This way, all hexes are transformed to a normalized range between 0 - 1. This also reduces the effect of outliers on our standardization methodology. Similar hex information and standardization procedures are used for other thematic area analysis.

Emergency department (ED) visits is our transportation response variable. However, since we are aggregating patient information up to a hex, we have to also define some statistic metric for

representing a hex's ED visit. We compare aggregating by the mean, median, and 3rd quantile. We observed that aggregating by the mean performed the best in our statistical models. To reduce the effect of sparsity, we only include hexes with more than 4 patients in a given hex.

5.3.2.3 Housing

We use similar claims data as well as hexes for exploring housing SDOH. However, we use American Community Survey (ACS) housing information data to capture SDOH factors related to housing. This ACS housing matrix contains 870 variables. However, we only include variables that pertain to the total of a specific factor rather than specific attribute types. For example, we ignore factors that are listed by race, by age, or by gender. We also remove highly correlated features (0.85 correlation). Our final feature matrix has only 24 variables, and it is standardized similar to the above normalization technique.

We also use (ED) visits as our housing response variable. The same aggregation methodology performed for the transportation SDOH is applied here.

5.3.2.4 Food insecurity

We consider the following factors for food insecurity:

- SNAP utilization data from ACS
- Number of food options within a certain distance from the hexagon centroid

The SDOH factors used are distance to food options and (Supplemental Nutrition Assistance Program) SNAP utilization. Food options are businesses entities such as groceries or food markets as well as fast food restaurants. SNAP utilization is obtained from ACS data. Food options are binned in a similar fashion as above, however, the binning distance varies. The binning distance that resulted in the best performing model (average least mean squared error (MSE) in 10 randomized runs) is the following: less than 0.5 miles, between 0.5 and 2 miles, between 2 miles and 5 miles, and greater than 5 miles. The ensuing feature matrix is standardized as in prior cases and fed as input to our machine learning algorithm.

We selected BMI as our food and insecurity response variable. As we are aggregating patient information up to a hex, we also have define some statistic metric for representing a hex's BMI. We compare aggregating by the mean, median, and 3rd quantile. We observed that aggregating by the 3rd quantile performed the best in our statistical models. We only include hexes with more than 15 patients in a given hex for this thematic area.

5.3.3 Summary of findings

The SDOH thematic areas explored in this paper as well as information on the data sources, medical risk response variables and data transformations used are presented in the Table below.

Table 5. SDOH factors and description of how they are obtained and used in this study

SDOH	Socio-economic factors	Data Source	Transformation	Response Variable
Economic well-being	Median income or median earning	ACS	Moving window	Index of Visit Quality
Transportation	Number of pharmacies within certain distances from the center of a hex	Google API	x/ (1+ x)	Annual ED Visits
Housing	Housing units by occupancy status	ACS	x/ (1+x)	Annual ED Visits
Food insecurity	Receipt of food stamps/SNAPS in the past 12 months by disability status for households, Number of food options within certain distances from the center of a hex	ACS	x/ (1+x)	BMI

We analyzed GIS data, patient health outcome data, and ACS data using a variety of statistical techniques to isolate high-priority variables and develop meaningful metrics to quantify important health care outcomes as functions of social determinants. Some examples of methodologies we used in our analysis are below.

We developed data-driven patient scoring metrics for four thematic areas, namely, Economic well-being, Transportation barriers, Housing insecurity, and Food insecurity. For economic well-being, we were able to use a moving window to encapsulate the relationship between this SDOH and its associated medical risk factor. However, for three of the thematic areas: transportation barriers, housing insecurity, and food insecurity, we relied on a machine learning algorithm (random forest regression model), to identify inherent relationships.

Economic Well Being: We discovered that the quality of physician/hospital visits are strongly related to income. We developed a novel metric, called the Index of Visit Quality (IVQ) to quantify

the quality of visits. The following rolling window plots in Figure 16 show how the IVQ improves with income and earning level, with the red line indicating the national benchmark.

- **Transportation Barriers**: We discovered that the general medical risk factor, annual

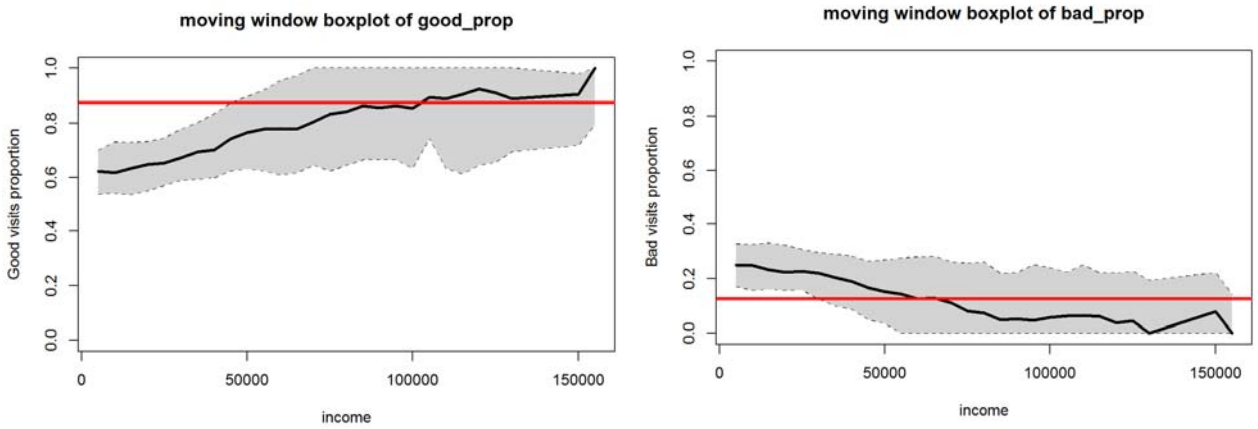

Figure 16. IVQ moving window plots. Left figure shows the relationship between the IVQ and income. Right figure shows the relationship between IVQ and earnings

emergency department visits (ED), are strongly related to transportation barriers, for which we developed a scoring mechanism. We developed a random forest model to encapsulate this relation. The following box plot in Figure 17 shows the strong relation between ED visits and transportation barriers.

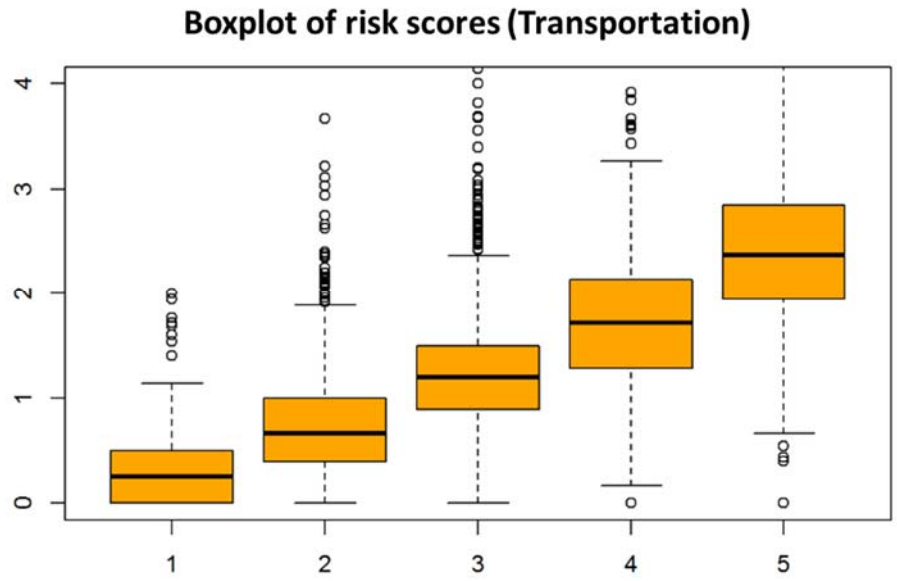

Figure 17. Box plot showing the performance of the random forest model as a predictor of ED visits using transportation barriers as the input

- **Housing Insecurity:** We discovered that the annual emergency department visits (ED) are strongly related to housing insecurity, and Figure 19 shows the predictive power of our random forest model.

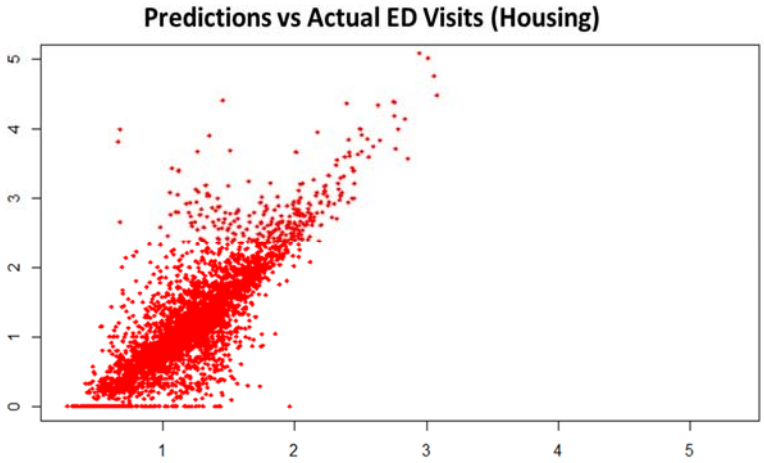

Figure 19. Scatter plot showing actual response health outcomes of patients (ED visits) against the predictions of the random forest model

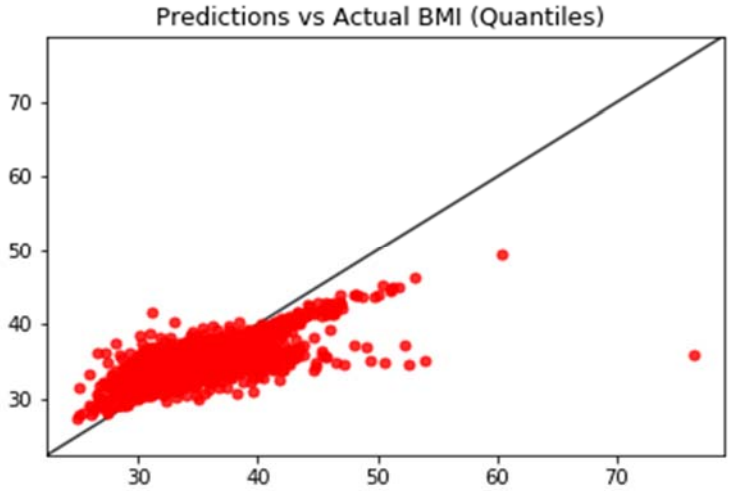

Figure 18. Scatter plot showing actual response health outcomes of patients BMI against the predictions from the random forest model

- **Food Insecurity**: We discovered that Body Mass Index (BMI) are strongly related to food insecurity. Figure 18 shows the plot of actual BMI of patients against the predicted BMI

using our model. The prediction accuracy of the model demonstrates the strength of the relation between BMI and food insecurity.

Detailed findings, as well as comprehensive empirical results for all four areas are described in the rest of the paper.

5.4 Detailed findings

We go over the details of our analysis in this section. We include County maps with the neighborhood risk scores overlaid. We also include some chronic disease heatmaps for comparison purposes.

5.4.1 Findings for economic well-being

We show the effects of economic well-being, namely median income and total earnings of that Hex as obtained from Census data. Histograms (frequency distribution) of these two variables, and their scatter plots, are shown in Figure 20. The scatter plot shows a strong positive relation between earnings and income, which is expected. The red dots in Figure 20 correspond to hexes with recorded 0 total earnings which is roughly 5% of the hexes used in our model.

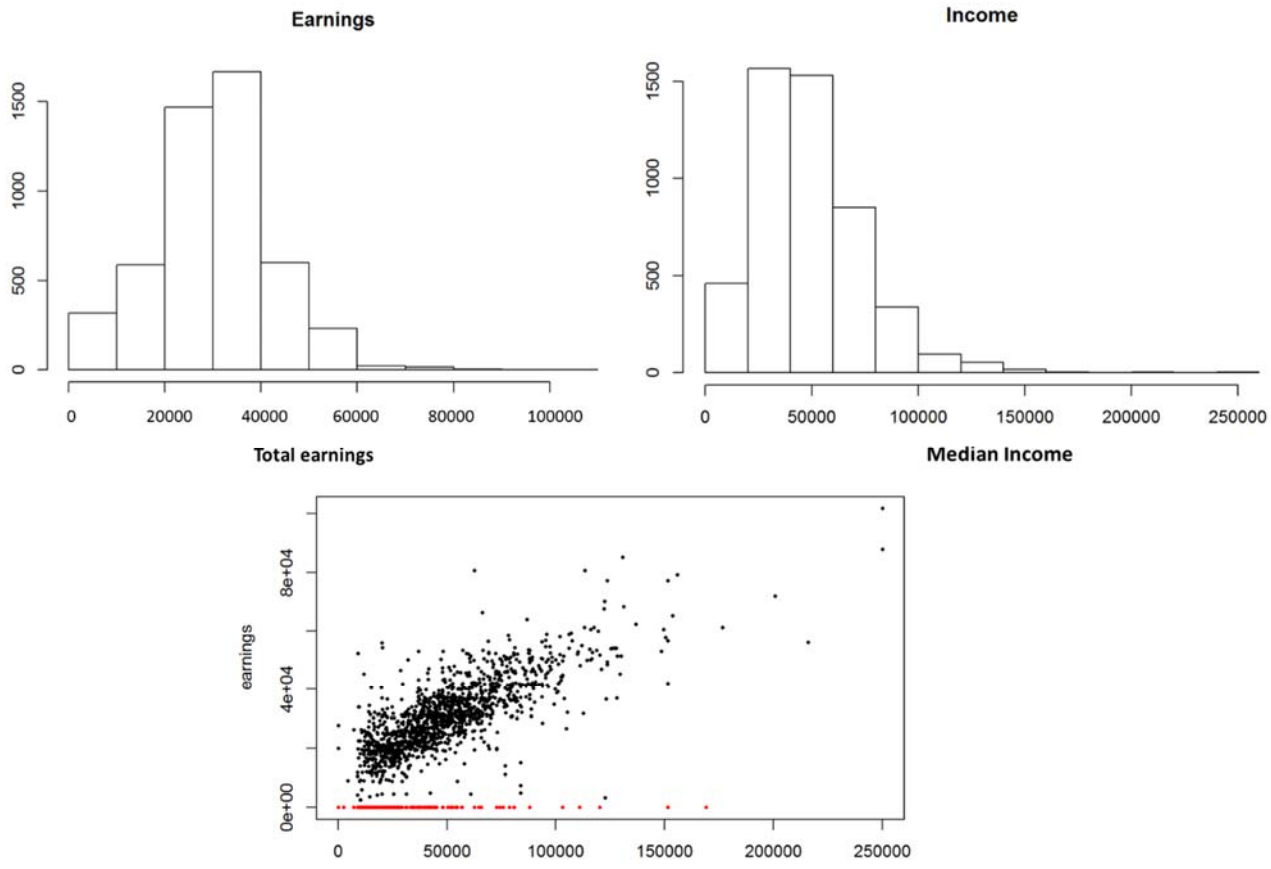

Figure 20. Histograms and scatter plot of income and earning

5.4.1.1 IVQ: A composite metric for patient behavior

We develop a composite metric called Index of Visit Quality (IVQ) that incorporates good visits (i.e, y = number of good visits / total number of visits) as well as bad visits.

From CDC data, we know that the annual number of visits to the emergency department is 141.4 million, whereas the annual number of physician office visits is 990.8 million.

For each hex, we define *x* to be the ratio of bad visits to good visits, i.e., *x* = number of bad visits by patients in that hex/number of good visits by patients in that hex.

Therefore, the ratio

$$x_0 = \frac{141.4}{990.8} \approx 0.143$$

can be considered the national benchmark for good:bad visits. We define the risk score

$$r = \frac{x}{x_0}$$

that expresses the hex-specific number relative to the national benchmark. If $r > 1$, that hex is more risky than the national average, and if $r < 1$, the hex is less risky than the national average. Finally, we define the following hex-specific index of visit quality (IVQ)

$$IVQ\ (hex) = 1 + \frac{4r}{1+r}$$

The IVQ ranges from 1 to 5. A high IVQ implies higher odds of bad visits than good visits, a low IVQ implies lower odds of bad visits than good visits. The advantage of using the IVQ, rather than simply using the ratio of bad visits to good visits (which is *x*), is that

- the IVQ ranges from 1 to 5, and is therefore a natural scoring metric
- the national average IVQ is 3 (corresponding to the case $r = 1$), IVQ greater than 3 implies riskier than national average, and IVQ less than 3 implies less risky than national average.
- the IVQ can easily incorporate $x = \infty$ cases where the number of good visits is zero the IVQ is 1 in such cases.

We now analyze the IVQ with respect to median income and total earnings. The results are presented as moving window box plots and categorical box plots in Figure 21.

5.4.1.2 Statistical methods employed

- **Moving window box plot**: To fully understand the trend of the relation of IVQ to income, we created a moving window box plot. For a given level of income (say $20K) we consider a window of width $10k centering that value (i.e., $15K-$25K for income = $10K), and look at all hexes belonging to that window. We then compute the median good-proportions, the first quartile (25 percentile) good proportions, and the third quartile (75 percentile) good-proportions for hexes in that window.
- **Categorical box plot**: To see the impact of economic well-being on IVQ, we classify the demographic hexes into three categories:
  - Lowest income neighborhood: median income less than $30K/ earnings less than $22K
  - Lower income neighborhood: median income greater than (or equal to) $30K but less than $60K/ earnings greater than (or equal to) $22K but less than $35K
  - Middle income neighborhood: median income greater than (or equal to) $60K/ earnings greater than $35K

First we look at moving window box plots. For IVQ, we restrict to windows with at least 20 hexes in them. The moving window box plots show that the IVQ improves (i.e., decreases) as income/earning gets higher. At around $65K income/ $40K earning, the IVQ gets better than the national benchmark.

Economic well-being hex risk score geo-spatial maps are shown in Figure 22 below. Heat maps, for the same hex maps, which are created from weighted smoothing of the hex boundaries is show also. In terms of correlation, the correlation coefficient between IVQ and income is 0.30 and that between IVQ and total earnings is -0.24. The Spearman correlation between IVQ and income is 0.28 that between IVQ and total earnings is -0.23. Thus, for the IVQ, rank correlations are similar to linear correlation. However, overall there is strong correlation between IVQ and economic well-being.

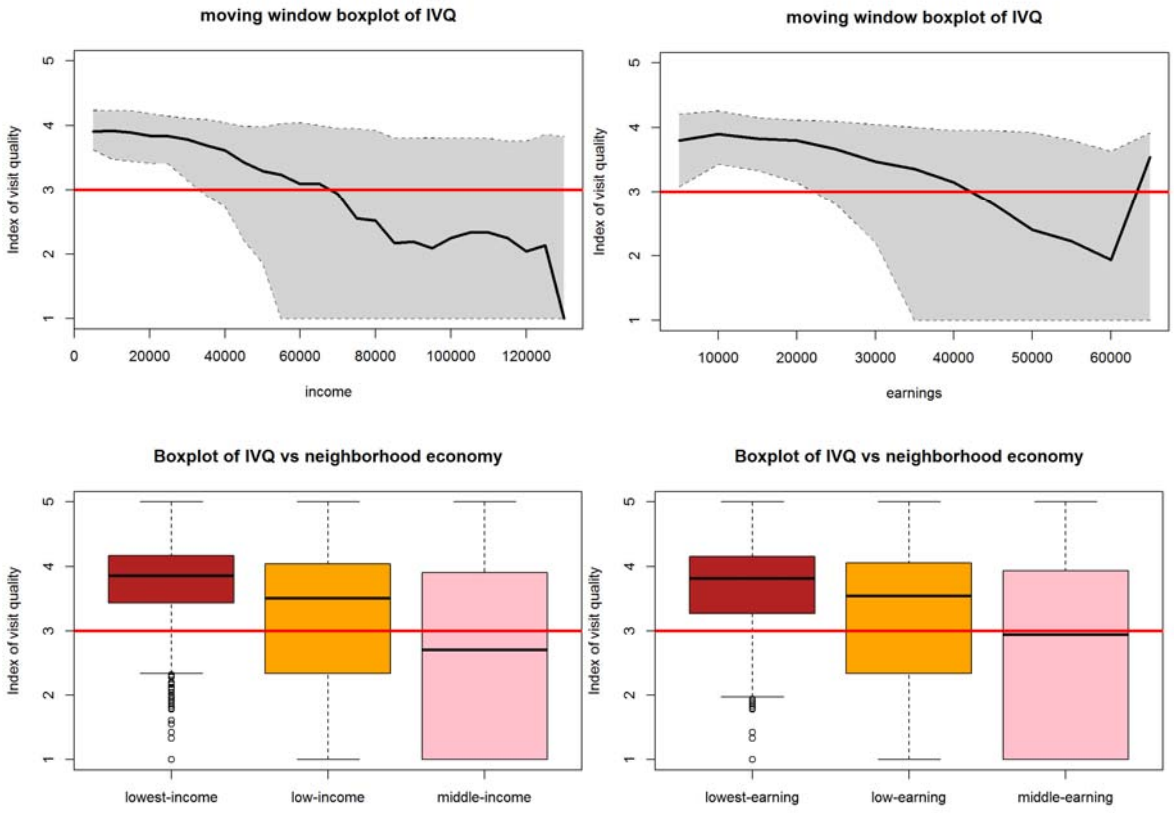

Figure 21. Top: Moving window IVQ plots for median income (left) and median earnings (right). Bottom: Box plots for different strata of earnings for median income (left) and median earnings (right)

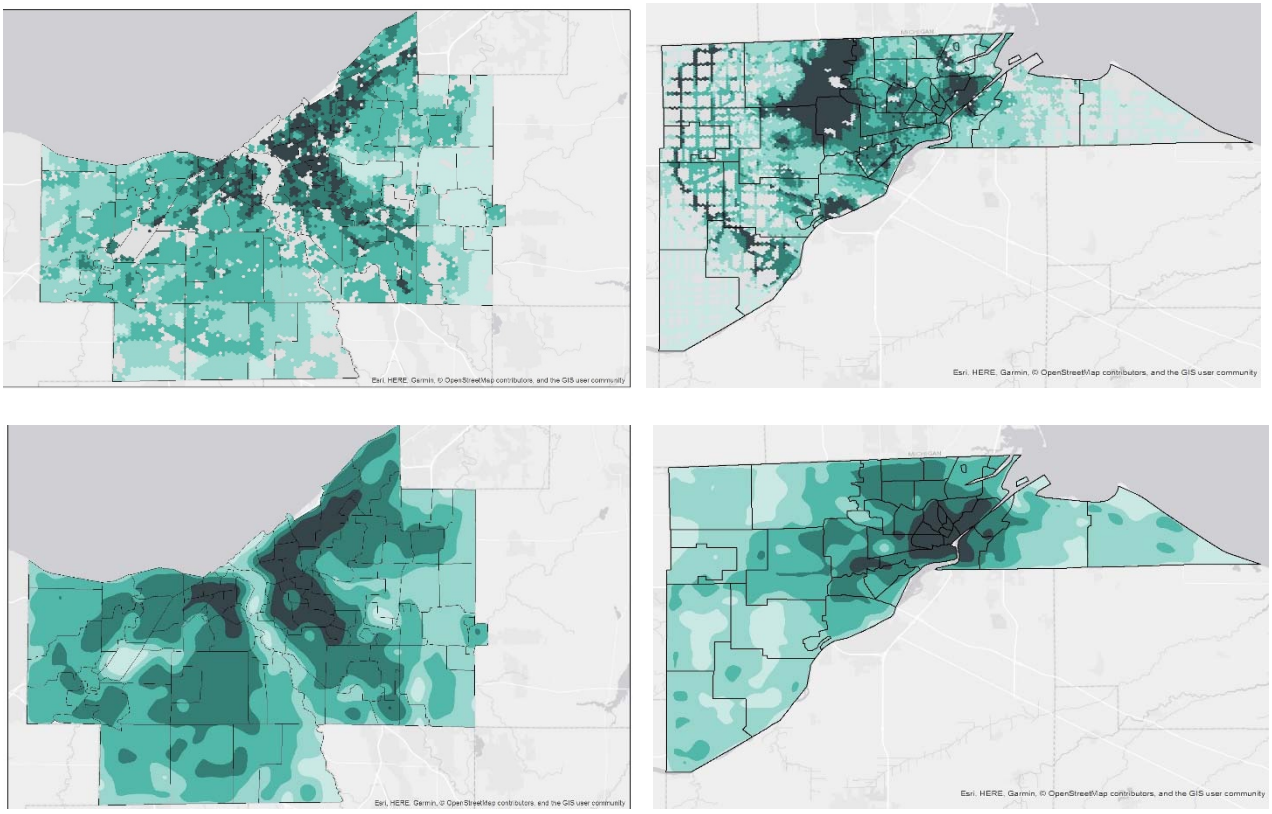

Figure 22. Hex maps and heat maps of economic well-being risk scores. Left figures are hex maps and heat maps for Cuyahoga County. Right Figures are hex and heat maps for Lucas County

5.4.2 Findings for transportation barriers

For transportation barriers, the data consists of the latitudes and longitudes of 12082 hex centroids in the Cuyahoga region, as well as 216 pharmacies. We compute the geo-spatial distance between each hex-pharmacy pair using the Vincenty algorithm, which is more advanced than other algorithms since it takes into account the fact that earth is an ellipsoid while other algorithms assume that the earth is a sphere.

We iterate through a number of binning parameters select the combinations which produce the smallest MSE error in our cross validation set. The 4 features are, hexes with number of pharmacies within 1 mile, pharmacies within 1 to 2 miles, number of pharmacies within 5 miles but greater than 2 miles away, and hexes with number of pharmacies greater than five miles away and less than 10 miles away.

Random forest model for transportation

The medical risk factor selected for pairing with transportation barriers is the annual number of visits to the Emergency Department. Transportation barriers are likely to impede individuals from getting regular check-ups and follow-ups, leading to severe medical conditions that require emergency care. Had there been no transportation barriers, such emergency medical conditions could have been prevented by regular preventative care. We split the data into 70% training and 30% testing. We tune the following hyper-parameters for the Random Forest Regressor module present in the python sklearn module.

- Number of trees in forest is varied from 100 to 1000 in increments of 100.
- Bootstrap sampling/no bootstrap sampling
- Maximum number of levels in tree is varied from 10 to 110 in increments of 10
- Maximum number of features is varied between the square-root of the total number of features or total number of features.

Of the 70% used for training, we run 10 randomized Random Forest algorithms for each hyper-parameter combination, keeping 10% of this data set for cross validation in each run. We select the model whose parameter combinations results in the lowest mean MSE error. The selected hyperparameters are shown in Table 6

Table 6. (Transportation) Hyperparameters selected after cross-validation for Random Forest algorithm.

Parameter	Best Values
Number of Trees	1000
Bootstrap	TRUE
Maximum number of levels	50
Maximum number of features	square-root

After building our random forest model, and relying on only these four features, we generate risk scores for the different hexes and show our results below. Figure 23 shows the performance of our model as well as the stratification of hexes into risk scores. We define risk scores as follows, lowest risk score of 1 pertains to hexes with annual ED visits equal to 0, risk scores of 2 is associated with hexes who have an average annual ED visit between 0.5 and 1, risk score of 3 is associated with hexes between 1 and 1.5, hexes with risk scores of 4 have annual ED visits between 1.5 and 2, while hexes with a risk score of 5 have annual ED visits greater than 2.

Finally, we apply the same model to a different County, Lucas County, and show that our model is able to suitably partition the hexes of Lucas County. The histogram in Figure 24 shows

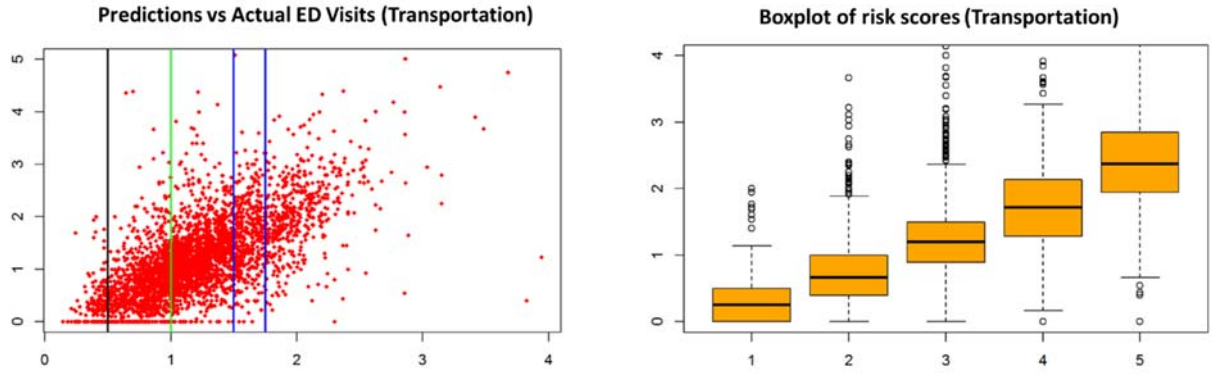

Figure 23. The left figure shows a scatter plot of actual ED visits against predicted ED visits using our model. Vertical lines demonstrate the discretized transportation barriers scores (1-5) using our model. The right figure is the categorical box plot showing a strong trend of the risk of ED visits increasing with higher transportation barriers

the predicted risk score distribution of our original County, Cuyahoga, with a new County, Lucas County. The accompanying geo-spatial maps in Figure 26 show spatial distribution of risk scores for the two counties. Furthermore, we include a heat map of patient ED utilization, generated from claims data in Figure 26 to reflect the similarities in our spatially predicted risk scores and actual ED utilization data.

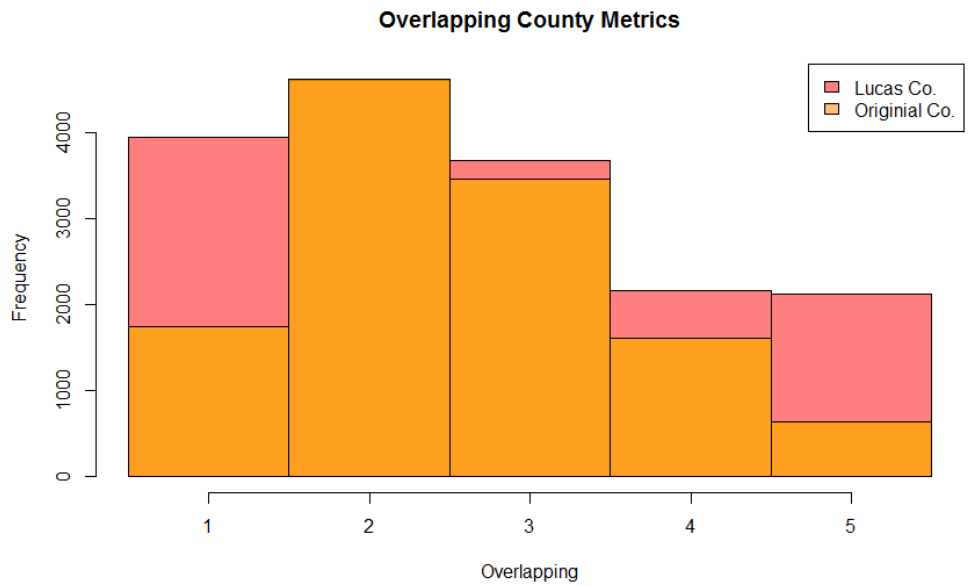

Figure 24. Overlay of risk scores for two different counties predicted using the random forest model. The orange plot reflects the distribution when the model is used to predict risk scores for the County it is trained in. The pink histogram is the risk score distribution for a different County using the same random forest model

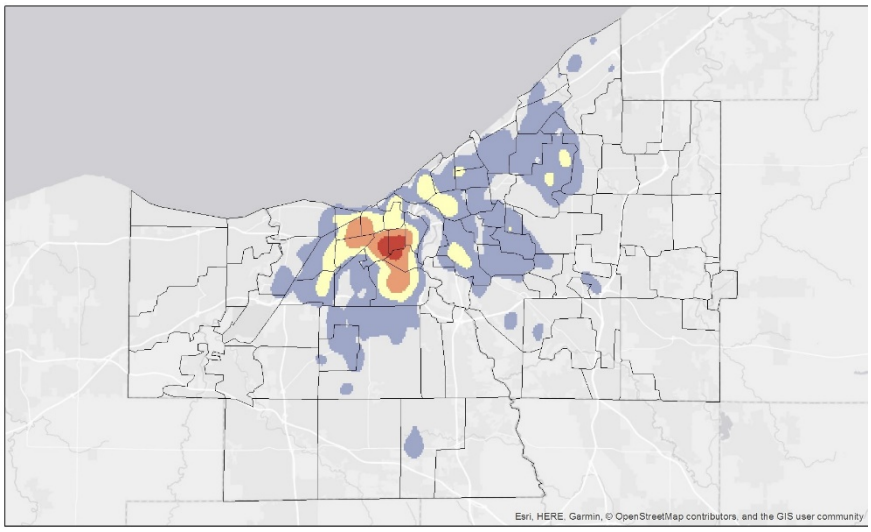

Figure 25. Patient clinical data (ED utilization) mapped for Cuyahoga County

5.4.3 Findings for housing insecurity

In this section, our goal is to investigate how patient health care outcomes are related to housing insecurity. The first step is to develop a metric to quantify housing insecurity.

5.4.3.1 Quantifying housing insecurity

Our primary data source for housing related variables is the ACS survey data published by the US Census Bureau. We use the ACS data related to housing as our covariates and remove variables that are highly correlated (> 0.85). Out of 76 possible variables, 52 are highly correlated to the other 24 variables so these 52 variables are excluded from future analysis.

5.4.3.2 Random forest model for housing

The health care outcome variable we used for this analysis is also the number of visits to the Emergency Department, similar to transportation. Also, we experimented with multiple aggregation strategies for combining patient level ED visits for a given hex. As mentioned above, we aggregated the data using the mean, after comparing its performance to other statistical aggregation methods. To reduce instances of noise, we only include in our model hexes with at least 5 patient claims data. Figure 19 shows the results of our model building and how well it fits with the data. We follow a similar procedure to the transportation SDOH for creating our random forest model and also for tuning the hyper-parameters.

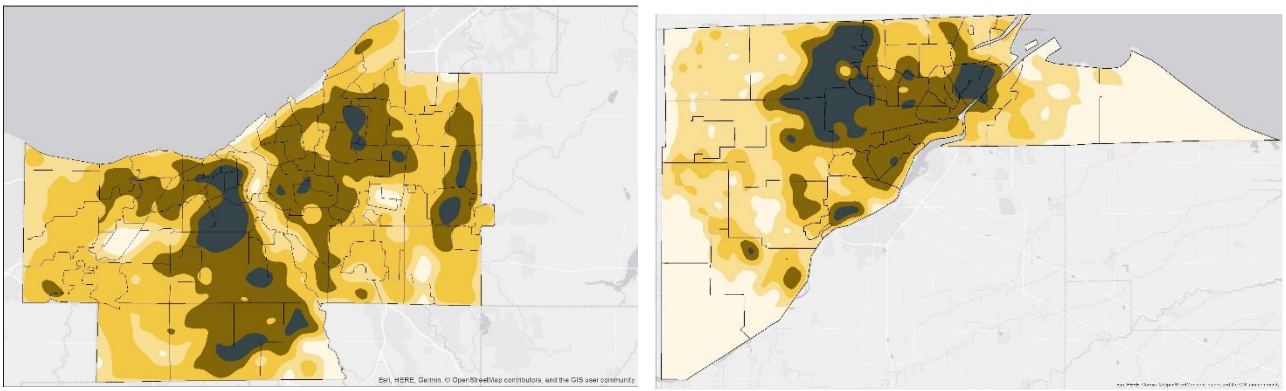

Figure 26. Heat maps of transportation risk scores. Left figures are heat maps for Cuyahoga County. Right Figures are hex and heat maps for Lucas County. The heat maps are derived from smoothing the edges of the hexes

Table 7. (Housing) Hyperparameters selected after cross-validation for Random Forest algorithm.

Parameter	Best Values
Number of Trees	1000
Bootstrap	TRUE
Maximum number of levels	40
Maximum number of features	square-root

We also applied the same analysis to a different County, Lucas County, and those results are below. We show how the risk scores for the different counties compare by overlaying the Cuyahoga County results on the Lucas County results. The results show that our random forest model is capable of predicting the risk scores of a different County reasonably.

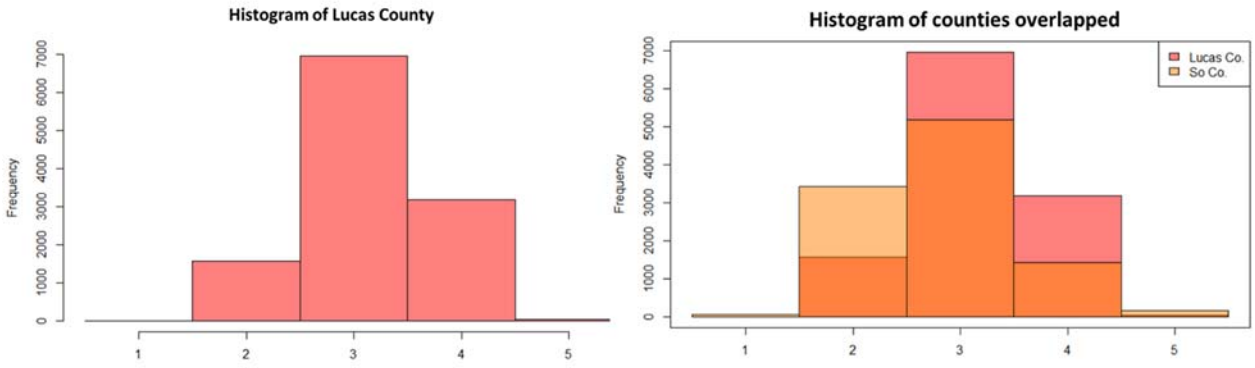

Figure 27. Left: Housing distribution of risk scores when using generated model on a new County, Lucas County. Right: An overlay of the histograms of two counties, where the light orange plot is Cuyahoga County (County used to train the model) and the pink plot is Lucas County. The orange portion of the overlaid histograms is where the distributions of the two counties overlap.

The accompanying geo-spatial maps in Figure 28 show spatial distribution of risk scores for the two counties. We also include a heat map of asthma, a common chronic disease associated with housing insecurity in Figure 29.

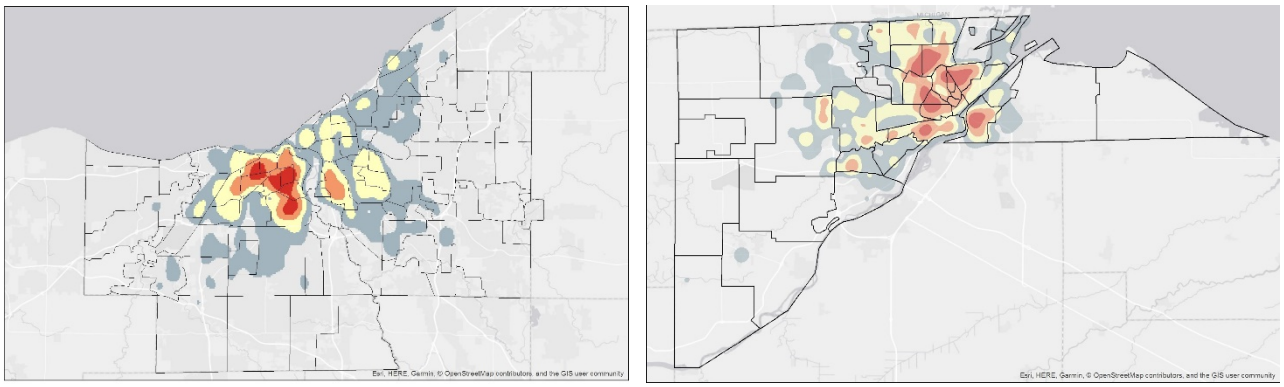

Figure 29. Heat maps of asthma prevalence in both counties. Left: heat map of Cuyahoga County, Right: heat map of Lucas County

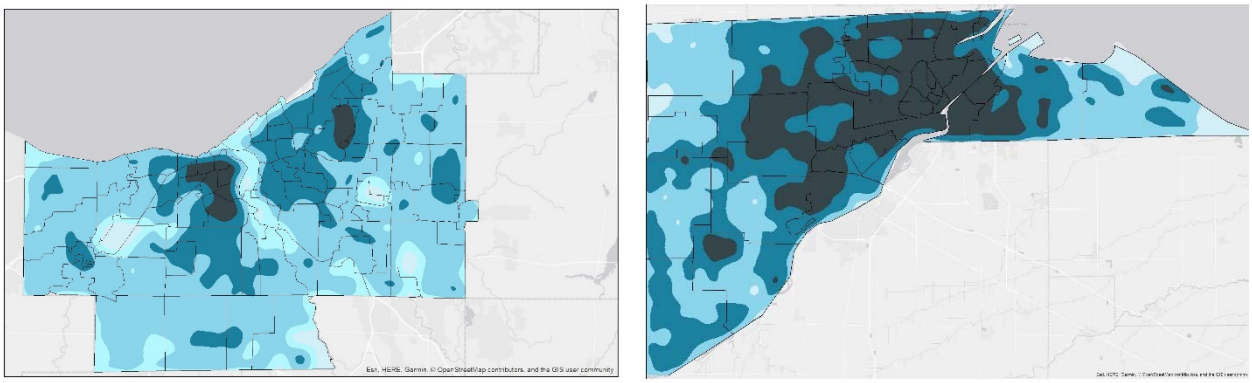

Figure 28. Heat maps of housing insecurity risk scores. Left figures are heat maps for Cuyahoga County. Right Figures are heat maps for Lucas County

5.4.4 Findings for food insecurity

Our next area of interest is Food Insecurity. The distances are separated into bins and a count of number of food options in each bin is used to generate a feature matrix. We iterate over different combinations of binning parameters and settle on these, 0.5 mi, (0.5mi to 1mi), (1mi to 5mi), (5mi to 10mi), and greater than 10mi. We combine this feature matrix with ACS SNAP utilization data. We found that food insecurity SDOH are better predictors of BMI, specifically we include in our model only hexes with at least 15 patient claims data and use the 75th quantile for that hex as the aggregate health outcome for that hex. This makes sense as BMI is closely associated with dietary choices. We built a random forest model using the health variables as outcome variables, and the food insecurity factors as input variables. Figure 18 shows the actual BMI vs predicted BMI (from this model) for all the hexes in Cuyahoga County.

Figure 31 shows the food insecurity predicted hex risk score geo-spatial maps. We also include a heat map of diabetes in Figure 30. Diabetes is a common chronic disease associated with food insecurity.

5.5 Overall conclusions

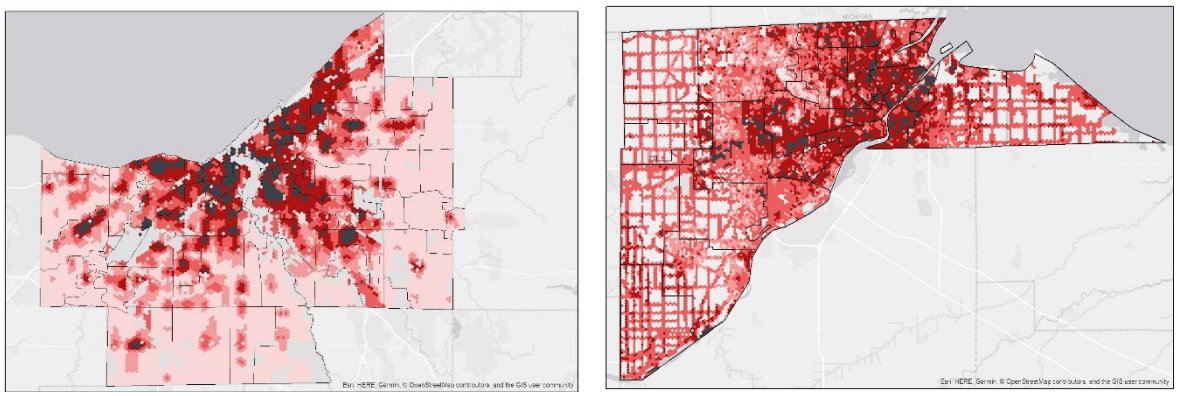

Figure 31. Hex maps of food insecurity risk scores. Left figures are hex maps and heat maps for Cuyahoga County. Right Figures are hex and heat maps for Lucas County

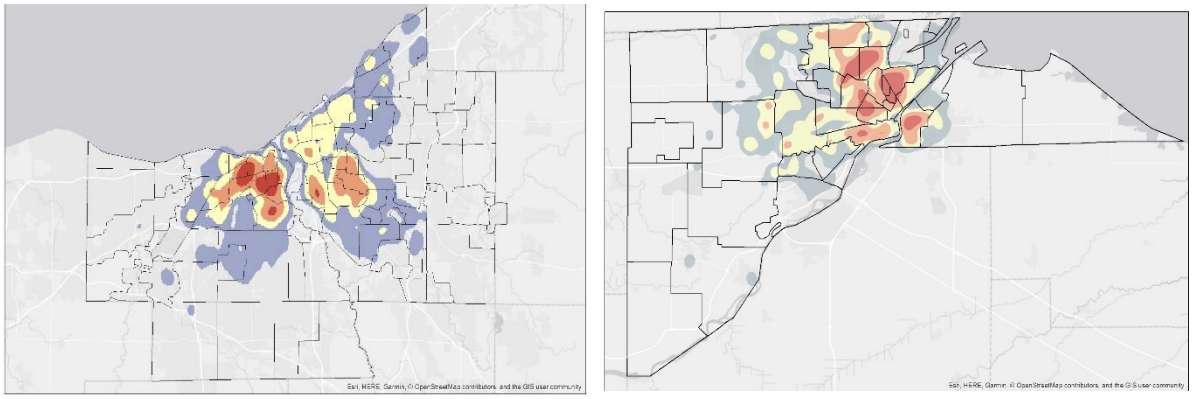

Figure 30. Heat maps of diabetes prevalence in both counties. Left: heat map of Cuyahoga County, Right: heat map of Lucas County

In conclusion, our investigation of the clinical records and demographic factors yielded important and interesting connections between general health outcomes and social determinants of health. We quantified these connections by developing a novel framework of risk scores. The results show that our models tend to conform conventional wisdom. For example, we see that patient health outcomes, as measured by good visits and bad visits, improves with economic well-being. Similarly, our analysis on other SDOH: transportation, housing, and food insecurity, shows that these SDOH are reliable metrics for predicting overall patient health outcomes. The accompanying patient diagnoses heat maps reinforce this point.

Our analysis also shows interesting patterns documenting the evolution of behavior at various risk levels and health outcomes. We note the interesting patterns at different economic levels for example. This work shows that we can use SDOH to infer health risks for a specific population, using only spatial and demographic information. We observed connections between multiple SDOH, medical risk factors, and behavioral patterns. This was achieved through utilizing multiple statistical and data analytics tools such as the moving window, Spearman's correlation, proper aggregation of patient claim information to create a representative value for a given hex, and employing random forest algorithms using proper standardization procedures.

There are opportunities to further this research by extending the hex-level analysis to the individual patient, either by obtaining specific individual social data or by devising methods to better map publicly available social data to an individual. Further, by investigating more populations that provide a cross-section of socioeconomic levels, such as the Lucas County patient population, we expect to obtain more representative results.

By leveraging the connections generated from this study, high return-on-investment interventions can be made to improve the health care outcomes of at-risk patients. In 2008, the WHO commission reported that SDOH were not adequately accounted for as significant factors that impact the the overall health of a population. Other commissions such as the RWJF Commission to Build a Healthier America, MacAurthur Foundation Network on Socioeconomic Status and Health, and the ACOs formed through the Affordable Care Act also echoed the same understanding. This study reaffirms the goals of these studies - i.e., SDOH do have an impact on overall health outcomes of a population. In addition, this study provides some details to what actions can be taken.

6 INVESTIGATED A SENSITIVE SPECTRAL METHOD FOR ANOMALY DETECTION, IDENTIFIED CRITICAL SHORTCOMINGS, AND MADE IMPROVEMENTS

6.1 Overview of chapter

The problem of anomaly detection in networks has attracted a lot of attention in recent years, especially with the rise of connected devices and social networks. Anomaly detection spans a wide range of applications, from detecting terrorist cells in counter-terrorism efforts to

identifying unexpected mutations during RNA transcription. Fittingly, numerous algorithmic techniques for anomaly detection have been introduced. However, to date, little work has been done to evaluate these algorithms from a statistical perspective. This work is aimed at addressing this gap in the literature, by carrying out statistical evaluation of a suite of popular spectral methods for anomaly detection in networks. Our investigation on statistical properties of these algorithms reveals several important and critical shortcomings, and we make methodological improvements to address such shortcomings. Further, this work carries out a performance evaluation of these algorithms using simulated networks, and also extends the methods to count networks. A paper on this topic has been submitted to the Journal of Network Science (Komolafe *et. al 2018*).

6.2 Significance of research

In this study we evaluate the statistical properties of the chi-square algorithm and L_1 norm algorithm. It is important to understand these properties when implementing the algorithms as a practitioner will want to know how the algorithms behave for different network sizes and network types. In (Miller, Beard et al. 2015), the only case that is explored for both algorithms is a network size of 4096 vertices and average degree of twelve. However, it is possible that the algorithms perform differently when the network size or sparseness of the network changes. Hence, evaluating the performances of these algorithms for other network size and average degree combinations would provide practitioners with insights on how the algorithms perform under various conditions. Also, in Miller et al., there is little discussion on establishing a signaling threshold. This is also very important if these algorithms are to be implemented. We will also demonstrate the effectiveness of these algorithms when applied to count networks, an area not explored in (Miller, Beard et al. 2015) or by any other investigators. Our main contributions in this work can be summarized in two main points:

- Evaluate the statistical properties of the chi-square algorithm and L_1 norm algorithm and identify critical shortcomings pertaining to their statistical properties as well as implementability
- Introduce methodological improvements to both algorithms. Specifically providing more practical and appropriate signaling and detection schemes in both algorithms

The networks we monitored are unlabeled static networks and we applied the above mentioned algorithms to the three network models generated in (Miller, Beard et al. 2015). Additionally, simulations rather than case studies, will be the primary method used in this chapter

to evaluate the methods as in (Miller, Beard et al. 2015). With simulations, anomalies can be introduced in a controlled manner and the ability to detect particular types of anomalies tested (Savage, Zhang et al. 2014, Azarnoush, Paynabar et al. 2016, Woodall, Zhao et al. 2017). Also, other investigators such as Woodall et al., and Savage et al., both concur that anomaly detection methods should be compared using simulated networks because case studies could contain conflating effects of some of the critical factors mentioned above (Savage, Zhang et al. 2014, Woodall, Zhao et al. 2017).

The rest of the chapter is as follows. In chapter 3.3 - 3.5, we describe the mathematical formulations used in defining the spectral properties of the networks and also describe the network models. In chapter 3.6 – 3.8, we observe the behavior of the algorithms for the non-anomalous cases in binary networks and in chapter 3.10, we observe the behavior for the anomalous subgraph present. In chapter 3.11- 3.12, we provide some methodological improvements to both algorithms investigated. Chapter 3.13 includes the discussions and our proposed future direction for this investigation.

6.3 Model setup and methodology

In this section, we discuss the formulation of the residual matrix that is used in the ensuing algorithms. We also describe the formulation of the three network models along with the spectral properties of their residual matrix (Miller, Beard et al. 2015).

6.4 Mathematical definitions

We describe a network graph G as composed of vertices V and edges E, $G = (V, E)$. A subgraph of such a network G is G_s such that all vertices of the subgraph, V_s, belong to the network graph G, $V_s \subseteq V$. Similarly, all edges in G_s are a subset of the edges in G, giving $E_s \subseteq E$. The total number of vertices in a network graph G gives us the network size n. That is, $n = |V|$. Also the number of edges in a network graph is M, $M = |E|$.

A network can also be represented as an $n \, X \, n$ adjacency matrix referred to as $\mathbf{A}_{ij}$. As there are two types of networks that are explored in this study: binary networks and count networks, the elements, a_{ij} of the adjacency matrix which can either be 0 or 1 for a binary network or a non-negative integer for count networks. For binary networks, we assume $A_{ij} \sim Bernoulli(p_{ij})$ and for count networks, $A_{ij} \sim Poisson(\lambda_{ij})$ where $i,j = 1,2,...,n$.

Our work investigates the use of some spectral properties of graphs for anomaly detection, specifically, the spectral properties of the residual matrix. Using the same terminology as (Miller, Beard et al. 2015), the residual of the observed adjacency matrix **A** is described in Equation (4)

(4)

$$B = A - E[A]$$

where **A** is the observed adjacency matrix which is a matrix of 1s and 0s in a binary network, or non-negative integers in a count network and E(Honeycutt) is the expectation of the adjacency matrix which is either a matrix with values p_{ij} for a binary network or λ_{ij} for a count network, $i,j = 1,2,...,n$.

6.5 Models

In (Miller, Beard et al. 2015), three types of network models with varying complexities are introduced. The models are the Erdös-Rényi model, the R-MAT model, and the Chung-Lu model. Their formulations are described below.

6.5.1 Erdös-Rényi

Erdös-Rényi (ER) networks are simple networks that are generated given only a single parameter, the background probability p_0 or λ_0 (Erdos and Rényi 1960, Chung, Lu et al. 2003, Miller, Beard et al. 2015). Figure 32 is a visualization of the adjacency matrix of a Bernoulli generated Erdös-Rényi Model with $p_{ij} = 0.1$ and $n = 1024$, and $i,j = 1,2,...,n$. The dots represent 1's in the adjacency matrix. The areas shaded white correspond to areas where the a_{ij} values are 0.

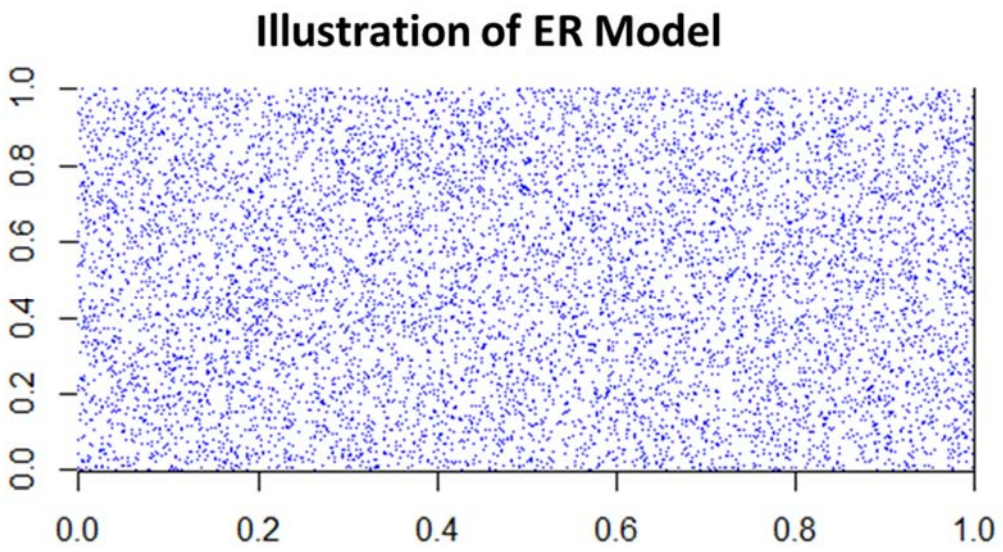

Figure 32. ER Model adjacency matrix illustration. n = 1024, E = 100000 and p_0 = 0.1

The residual matrix **B** for an Erdös-Rényi generated model is described in Equation (4)

$$B = A - E[A]$$

where the expectation of the adjacency matrix , E[**A**], is p_0***1** · **1**' for the binary network and λ_0***1** · **1'** for the count network.

6.5.2 R-MAT Model

Networks such as the world wide web, virus propagation networks, peer-to-peer networks, and so many others typically follow certain regularities or laws (Akoglu, McGlohon et al. 2010). The **R**ecursive **MAT**rix (R-MAT) model is a network model that is able to simulate these phenomena while requiring only a few parameters and has very fast generation speed (Chakrabarti, Zhan et al. 2004, Leskovec, Chakrabarti et al. 2005, Chakrabarti, Faloutsos et al. 2007). It was introduced by Chakrabarti et al, in 2004 (Chakrabarti, Zhan et al. 2004).

The R-MAT model is different from other network generation models in one important aspect - we specify the number of edges, *M*, to assign to the network and then generate the network (Chakrabarti, Zhan et al. 2004, Miller, Beard et al. 2015). The model is only used to generate binary networks.

To assign these pre-specified number of edges, *M*, in the R-MAT model, we start with a base edge assignment probability matrix shown below.

The edge assignment probabilities *a, b, c, d* have these relationships (Chakrabarti, Zhan et al. 2004, Miller, Beard et al. 2015):

$$a > d > c = b \; and \; a + b + c + d \; = 1$$

Assuming the 2 X 2 matrix below is an empty adjacency matrix, the values *a, b, c, d* correspond to the probability that a given edge will end up in one of the four cells. In our study, the edges are assigned using a multinomial distribution.

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

Then for a given empty adjacency matrix **A** of size n X n, we subdivide the matrix into four partitions and randomly choose to assign a particular edge to one of the partitions based on the probabilities of the base edge assignment probability matrix above. Once an edge is designated to a partition, we subdivide that particular partition again into four partitions and choose to assign that particular edge to one of the subdivisions based on the base edge assignment probability matrix. This process is repeated iteratively until we end up in cell a_{ij}, that is a *1 X 1* cell and this cell receives the edge (Chakrabarti, Zhan et al. 2004, Miller, Beard et al. 2015). Since our network is an un-directed network, the process is repeated for *M*/2 iterations. Once cell a_{ij} receives an edge, cell a_{ij} also receives the edge. As we allow for self-loops as in the Miller case, an edge can end in a cell a_{ij} where $i = j$.

Figure 33 is an example of the adjacency matrix formed using the R-MAT model generation technique. In the example, a = 0.5, d = 0.25, b = 0.125 and c = 0.125. Also n = 1024 and the number of edges is E ~ 100000

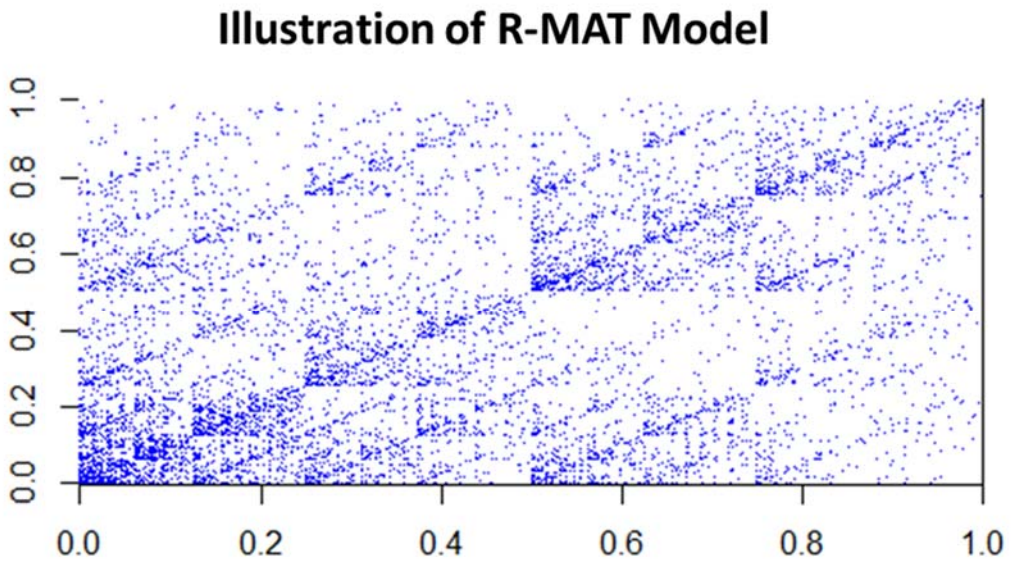

Figure 33. R-MAT Model adjacency matrix illustration. n = 1024, E ~ 100000

The resulting network could have a community structure as seen in Figure 33. Presence of community like patterns make the problem of detecting an anomalous subgraph even more challenging.

The associated n X n probability matrix **P**, with elements, p_{ij}, for the R-MAT model can be calculated as shown in Equation (5).

(5)

$$p_{ij} = 1 - (1 - \hat{p}_{ij})^t$$

where $\ddot{p}_{ij}$ is the i^{th} and j^{th} element of the matrix after performing the k-fold Kronecker product, where *i,j = 1,2,...,n* (Miller, Beard et al. 2015).

Because of the difficulty of obtaining the expected adjacency matrix, E[**A**]for the R-MAT model, a rank-1 approximation is used instead just as in (Miller, Beard et al. 2015). The rank-1 approximation is considered a very close estimate of the expected residual matrix (Miller, Beard et al. 2015). Under this model, the residual matrix of the observed network is calculated in Equation (6)

(6)

$$B = A - \frac{kk^T}{2M}$$

where **B** is the residual of the network, **A** is the observed adjacency matrix which is a matrix of 1s and 0s in a binary network, the vector ***k*** is the sum of the j^{th} columns across each i^{th} row, giving us a vector that represents the observed degrees of each node in the matrix and *2M* is the total number of edges in the network.

6.5.3 Chung Lu model

Under the Chung Lu model, the popular nodes get ever more popular (Chung, Lu et al. 2003). That is, popular nodes have a higher propensity to develop an edge between each other. This tendency leads to a community like structure. We observe the community like structure in Figure 34. In Figure 34, *n = 1024* and the number of edges is *E = 100000*.

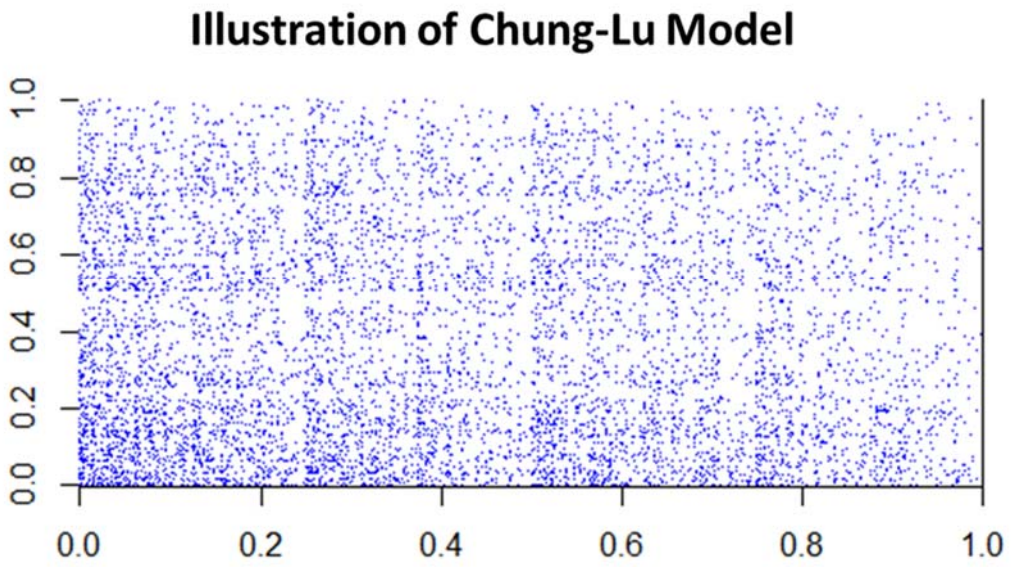

Figure 34. Chung Lu Model adjacency matrix illustration. n = 1024, E ~ 100000

In our study, as in (Miller, Beard et al. 2015), the probability matrix **P** in the Chung-Lu model for the binary case is calculated from the R-MAT randomly generated graph. Specifically:

(7)

$$P = \frac{kk^T}{2M}$$

where the vector **k** is the sum of the j^{th} columns across each i^{th} row of the observed R-MAT model and $2M$ is the total number of edges in the R-MAT network (Miller, Beard et al. 2015). The Chung Lu generated graph is rank 1 as the network is derived entirely from the vector **k**.

Under the Chung-Lu model, the residual matrix of the observed network is calculated as in Equation (4) which is as follows

$$B = A - EA$$

where **B** is the residual of the network, **A** is the observed adjacency matrix which is a matrix of 1s and 0s for the binary network and E[**A**].

6.6 Chi-square and L1 norm algorithms

In this section, we report the statistical properties of the L_1 norm and chi-square algorithm in (Miller, Beard et al. 2015).

6.6.1 Eigenvector L_1 norm algorithm

Given a vector $\mathbf{X} = (x_1, x_2, ..., x_n)$, the L_1 norm of **X** is

(8)

$$|X|_1 = \sum_{i=1}^{n} |x_i|$$

For the residual matrix **B**, the L1 norm of one of its eigenvectors, $\mathbf{X}_k$, in its eigenspace will be significantly lower than L_1 norms of other eigenvectors if a small clique or bipartite graph is present for k = 1,2,...,n. This is due to the orthonormal property of eigenvectors. When an anomaly is present, the elements of the eigenvector, $\mathbf{X}_k$, that correspond to the nodes of the anomalous subgraph will have absolute values that are significantly larger than other elements in the eigenvector $\mathbf{X}_k$ (Miller, Bliss et al. 2010, Miller, Beard et al. 2015). Since the eigenvector is orthonormal, $\sum_{i=1}^{n} {x_i}^2 = 1$ needs to be satisfied. But for eigenvectors affected by the presence of an anomalous subgraph, only a small portion of the elements, x_i, in the eigenvector have values whereby $\sum_{i=1}^{s} {x_i}^2 = 1$. Here *s* corresponds to the nodes in the anomalous subgraph (Miller, Bliss et al. 2010, Miller, Beard et al. 2015). For this reason, the L_1 norm for that eigenvector will be significantly smaller than L_1 norms of other eigenvectors. Therefore an anomalous subgraph can be detected by calculating the minimum L_1 norm in an eigenspace. The L_1 norm statistic, L is calculated as

(9)

$$L = -\min_{1 \le k \le m} \frac{|X_k|_1 - \mu_k}{\sigma_k}$$

where $|\mathbf{X}_k|$ is the L_1 norm of the eigenvector in question, μ_k and σ_k are the mean and standard deviations of L_1 norms estimated from historical observations. Therefore, to apply this statistic, one needs to first obtain the L_1 norms for graphs where it is known that no anomalies are present.

The standardization is as follows. For each historical network observation of size *n*, where no anomaly is present, its residual matrix as in Equation (4) is first calculated. Then for each residual matrix, an arbitrary set number of *m* largest eigenvalues, where $m \le n$, are sorted in decreasing order and the L_1 norms of the corresponding eigenvectors calculated. That is, an L_1 norm value is calculated for each eigenvector $\mathbf{X}_k$ where $k = 1,2,...,m$ and the corresponding eigenvalues, $\lambda_1 \ge \lambda_2,... \ge \lambda_m$. Then the mean of the historically observed L_1 norms for each of the eigenvectors $\mathbf{X}_k$'s is estimated, yielding μ_k where $k = 1,2,...,m$ along with their standard deviations σ_k. When a new graph is observed, its *m* largest eigenvalues are extracted in decreasing order and their corresponding eigenvector L_1 norms calculated.

The smallest, negative value is used as the detection statistic and if it crosses a specified threshold *K*, the presence of an anomaly is suspected. According to Miller et al., the detection statistic follows a Gumbel distribution so the specified threshold K should be set according to this distribution (Miller, Beard et al. 2015). The Gumbel distribution is defined by two parameters, the location parameter a_m and the scaling parameter b_m (Nadarajah and Kotz 2004, Wolpert 2014). Given that the random variable follows a standardized normal distribution, as we assume in our case, the parameters a_m and b_m can be calculated from the Extreme Value Theorem:

(10)

$$a_m = -\phi^{-1}(1/m)$$

(11)

$$b_m = \frac{1}{a_m}$$

Where φ is the cumulative density function of the standard normal distribution and *m* is the number of random variables the extrema is derived from. In our case, *m* is the number of eigenvectors used to derive the L_1 norm statistic as in Equation (9).

The parameters a_m and b_m can also be calculated using the Method of Moments (MOM) estimators which requires using historical data as shown in Equations (12) and (13) (Hoisungwan 2015).

(12)

$$a_m = \frac{1}{h}\sum_{i=1}^{h} L_i - b_m\gamma$$

(13)

$$b_m = \frac{\sqrt{6}S}{\pi}$$

where h is the number of historical observations, L_i is the L_1 norm detection statistic for each historical observation, $\gamma \sim 0.57722$, S is the standard deviation of the L_1 norm detection statistic from the h historical observations (Hoisungwan 2015). The steps for the L_1 norm algorithm are described below.

Algorithm 1:	L_1 norm algorithm
Input:	Observed network
Output:	Alert for anonymous subgraph detected
	Obtain μ_k and σ_k from historical observations:
	Standardized observed network L_1 norms for each eigenvector with corresponding μ_k and σ_k :
	Calculate location parameter a_m and scaling parameter b_m ** :
	Transform standardized observed L_1 norms to standard Gumbel distribution using parameters a_m and b_m** :
	Signal if observed network detection statistic for a given eigenvector crosses a specified threshold, $L_i > K$** :
** :	Identifies serious concerns in critical steps during implementation of algorithm

For these reasons, three main concerns have to be addressed when implementing the L_1 norm in practice:

- Which of the two approaches, Extreme Value Theorem or MOM estimators, should be used to estimate the location parameter a_m and scaling parameter b_m?
- Although Miller et al., claim that the algorithm is applicable to static networks, the statistic in Equation (9) requires historical observations to calculate the mean of the L_1 norms, μ_k

and their corresponding σ_k values. This makes the algorithm impractical for many static networks.

- The number of eigenvectors m to select from the eigenspace needs to be specified

These concerns could significantly impact a practitioners' ability to implement the algorithm as will be demonstrated in sections 3, 4 and 5. These are also limitations not stated explicitly in (Miller, Beard et al. 2015) as the paper claims the algorithm is applicable to a static observed network with no prior information. Also, a criteria for signaling is not explicitly presented in (Miller, Beard et al. 2015). Future sections further elaborate on possible detection values and their resulting relative performance.

6.7 Chi-square algorithm

For networks with no anomalies, empirical observations of the first two principal components of the residual matrix corresponding to the two largest eigenvalues shows that they are radially symmetric and scatter outwards when plotted (Miller, Beard et al. 2015). This is due to the orthonormal property of these eigenvectors (Miller, Beard et al. 2015).

The chi-square algorithm relies on this radial symmetry of the first two principal components of the residual matrix, **B**, to detect anomalies. We use the number of points in each quadrant when we plot the first two principal components to calculate our detection statistic. For illustration, in Figure 35, Figure 35a has no anomalous subgraph embedded while Figure 35b has a 15 node clique, where all 15 nodes are connected to every other node in the subgraph and randomly embedded into the 1024 node network.

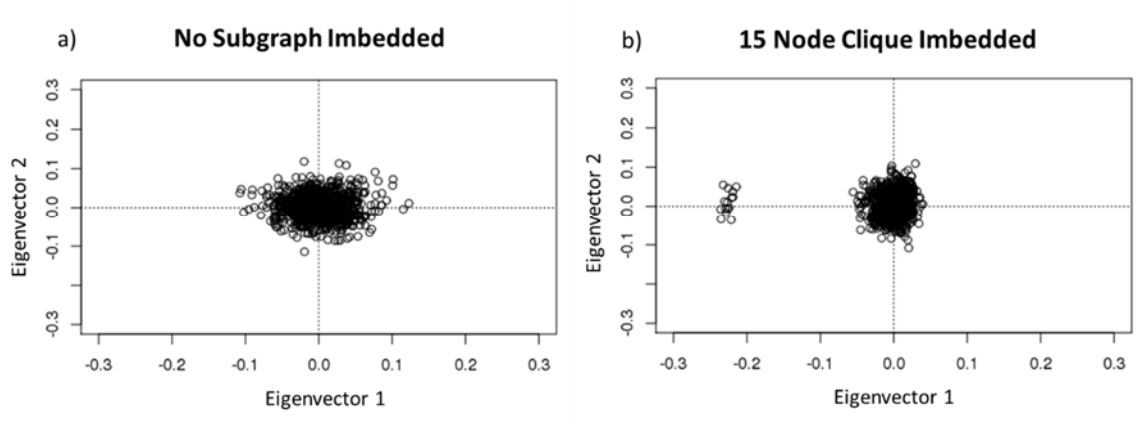

Figure 35. (a) ER Model of 1024 points with no anomalies showing radial symmetry about origin, (b) ER Model of 1024 points with anomalous sub-network present.

Table 8. Count of points in each quadrant from Figure 35.

Figure	Q1	Q2	Q3	Q4	Total
(a)	258	259	251	256	**1024**
(b)	250	247	254	273	**1024**

For the chi-square algorithm, the first step involves obtaining the residual matrix of the network as described in Equation (4). Then we obtain the two eigenvectors, $\mathbf{X}_1$ and $\mathbf{X}_2$ corresponding to the two largest eigenvalues and plot these orthogonal eigenvectors on a Cartesian coordinate system. Next we compute a 2 X 2 contingency table where each cell of the table is the number of points that fall in a particular quadrant. The 2 X 2 contingency table is a matrix **O** with elements $\mathbf{O}_{pq}$. We compute the expected number of points in each cell of the table assuming independence as in Equation (14).

(14)

$$\bar{O}_{pq} = \frac{\left(O_{p1} + O_{p2}\right) + (O_{1q} + O_{2q})}{N}$$

The chi-square statistic is then

(15)

$$\chi^2([x_1 x_2]) = \sum_p \sum_q \frac{(O_{pq} - \bar{O}_{pq})^2}{O_{pq}}$$

Because the non-anomalous case assumes that the points are radially symmetric, rotating the Cartesian plane should not affect the detection statistic result. But an anomaly could project the points in a certain direction so the Cartesian plane is rotated to maximize the detection statistic as in Equation (16).

(16)

$$\chi^2{}_{max} = -\max_{\theta} \chi^2(x_1 x_2 \begin{bmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{bmatrix}^T)$$

Algorithm 2:	*chi-square* algorithm
Input:	Observed network
Output:	Alert for anonymous subgraph detected
	Obtain the two eigenvectors that correspond to the two largest eigenvalues of the residual matrix:
	Plot the two eigenvectors and count the number of points in each quadrant**:
	Standardize the observed counts and calculate the χ^2 statistics using Equation (15)
	Rotate the eigenvectors using Equation (16) and calculate χ^2 statistics**:
	Store the maximum χ^2 statistic value, χ^2_{max} :
	Signal if observed detection statistic crosses specified threshold, i.e, $\chi^2_{max} > K$**:
** :	Identifies serious concerns in critical steps during implementation of algorithm

For these reasons, three main concerns have to be addressed when implementing the chi-square statistic in practice:

- Counting the number of points in each quadrant has some limitations such as accounting for points that lie on an axis or at the origin
- For Equation (16), how small should the grid of θ values used be to ensure the global maximum is captured?
- What is the appropriate cut-off value *K* such that the algorithm signals when $\chi^2_{max} > K$?

Counting of vertices in each quadrant to calculate a detection statistic has some limitations and this will be explored in the following section. Furthermore, (Miller, Beard et al. 2015) imply that the detection statistic follows the chi-square distribution for all network size and background probability combinations. This implies that the detection statistic is (a) independent of the network size and/or, (b) independent of the background probabilities.

Also a signaling value, *K*, is not specified although this is a critical component for detecting an anomaly. **A** practitioner applying the algorithm would need to know at what point the method should signal the presence of an anomaly. The following chapter will investigate these concerns.

6.8 Evaluating statistical properties of algorithms when there is no anomaly

The assumptions from (Miller, Beard et al. 2015) will be confirmed or rejected by observing both the detection statistic results with no anomaly present and when anomalies are present. In this section, we focus on the scenario when no anomalies are present and explore the

performance of the algorithms for the Erdös-Rényi, R-MAT and Chung-Lu models. To investigate the performance of the algorithms for different networks, we will consider the following network sizes, $n = 128, 256, 512, 1024$ and also background probabilities, $p_0 = 0.01, 0.05, 0.1, 0.3$. For brevity, the results that are shown in Table 9,

Table 10, Table 11, and Table 12 only include connectivity $p_0 = 0.05, 0.1, 0.3$ for $n = 128$ and $p_0 = 0.01, 0.1, 0.3$ for $n = 256, 512,$ and 1024. The figures included in this chapter are for $n = 512$ with $p_0 = 0.1$. Additional figures and tables are in the Appendix and follow similar patterns as in the figures below.

We evaluate the statistical properties of both algorithms for the case with no anomaly present by comparing their empirical distributions to the theoretical distributions. We specifically compare the results of the upper quantiles, between 95% - 99%. This is to monitor the performance of false signals as anomalies of interest in our case are 5% of the network or less. Histograms and Q-Q plots are used as visual aids for observing the distributions of the detection statistics along with the expected theoretical distributions.

6.8.1 Statistical properties of eigenvector L1 norm algorithm

In Miller et al., the detection statistic from the L_1 norm algorithm is stated to follow a Gumbel distribution. This distribution depends on two parameters, the location and scaling parameters a_m and b_m respectively. These parameters need to be estimated in order to either standardize the observed detection statistic or convert the standard Gumbel distribution to the observed statistic. Furthermore, it is not discussed in Miller et al., what the effect of the number of eigenvectors m, could have on the detection statistic result. An arbitrary value, m = 100, is used in

the paper without a discussion or validation of the approach. In this section, we will compare two different estimation techniques for a_m and b_m where in one case we use the Method of Moments estimator (MOM) that uses historical data to estimate these parameters as in Equation (12) and (13) and in the second case, we use the Extreme Value Theorem approach in Equation (10) and (11). In addition, we will also compare the effect of the arbitrary set value, m, on the non-anomalous behavior of the L_1 norm statistic. That is, we set $m < n$ in one case and $m = n$ in another. If the algorithm statistic follows the Gumbel distribution, then we should expect a better performance for when $m = n$ as the (a_m, b_m) estimates should be more accurate (Hoisungwan 2015).

6.8.1.1 Estimating a_m and b_m using MOM estimator and setting m < n

In this section, we investigate the behavior of the L_1 norm algorithm for $m < n$ and no anomalies present. The number of eigenvectors, m is arbitrarily set to 30 for networks of sizes 128 and 256. The size, m, is increased to 50 for networks of sizes 512 and 1024. These values approximate the arbitrarily set values in the Miller et al, (Miller, Beard et al. 2015). When estimating the location and scaling parameters, a_m and b_m, we use the MOM estimator as in Equation (12) and (13). These equations require historical data in order to be implemented (Hoisungwan 2015). Historical data in our case involves first running 1000 simulations to estimate both μ_k, σ_k and to calculate the L_1 norms from residuals with no anomalies. The results for the Erdös-Rényi , R-MAT, and Chung-Lu models are shown.

Figure 36 for the Erdös-Rényi model shows that the detection statistic distribution is similar to the theoretical Gumbel distribution, although they are dissimilar at the higher quantiles. The same observation is noted in the plot comparisons for the R-MAT model. Deviations at the higher quantiles reduces the usefulness of the algorithm to a practitioner. It makes setting an effective signaling threshold more difficult. Interestingly, the Chung-Lu model as in Figure 36 has the worst performance compared to the other two models. Although the model in our study uses a Rank 1 approximation of the R-MAT model to generate its probability matrix, there seems to be little similarity in the results for the Chung-Lu model in comparison to the R-MAT model. Table 9 also corroborates our conclusions for this case.

6.8.1.2 Estimating a_m and b_m using the Extreme Value Theorem and setting m < n

In this section, we investigate the non-anomalous behavior of the L_1 norm algorithm for $m < n$ and employ the Extreme Value theorem to estimate the location and scaling parameters, a_m

and b_m. We also set $m = 30$ for the cases where $n = 128, 256$ and $m = 50$ for $n = 512, 1024$. The observations are similar to the scenario for when the MOM is used to estimate the location and scaling parameters, a_m and b_m respectively. For example, the Erdös-Rényi and R-MAT statistic distributions also diverge at the higher quantiles as in Figure 37. There is also a larger variation in the simulation results for this case. For example in

Table 10, the column corresponding to the 99% quantile simulation results, the Erdös-Rényi , R-MAT and Chung-Lu columns have wider ranges (*2.56, 6.57, and 11.83*) in comparison to Table 9 with ranges of (*0.63, 1.92, 9.92*) for the Erdös-Rényi , R-MAT and Chung-Lu model results respectively. The Chung-Lu model performs worse also in this case as seen in Figure 37 and

Table 10 with multiple simulation values either much higher than the theoretical Gumbel distribution or much lower.

6.8.1.3 Estimating a_m and b_m using historical data and setting m = n

In this section, we investigate the non-anomalous behavior of the L_1 norm algorithm for m = n and also when estimating the location and scaling parameters, a_m and b_m, using simulated historical data. In this scenario, only the ER model performs comparatively to the theoretical Gumbel distribution as is observed in Figure 38. The detection statistic values when applied on the Chung-Lu model has the largest variation as seen in Table 11. This implies that this approach depends on the type of network model and the number of eigenvectors, m, used.

6.8.1.4 Estimating a_m and b_m using the Extreme Value Theorem and setting m = n

In this section, we investigate the non-anomalous behavior of the L_1 norm algorithm for m = n and also when estimating the location and scaling parameters, a_m and b_m using the Extreme Value Theorem. We note that for the Erdös-Rényi and R-MAT models, the results are comparable as seen in Figure 39. Both histogram and Q-Q plots are similar although when compared to the case when $m < n$ using the Extreme Value theorem, these simulation values are generally lower. We also notice that the Chung-Lu model results are similar to when $m < n$ as in Figure 39 and Table 12.

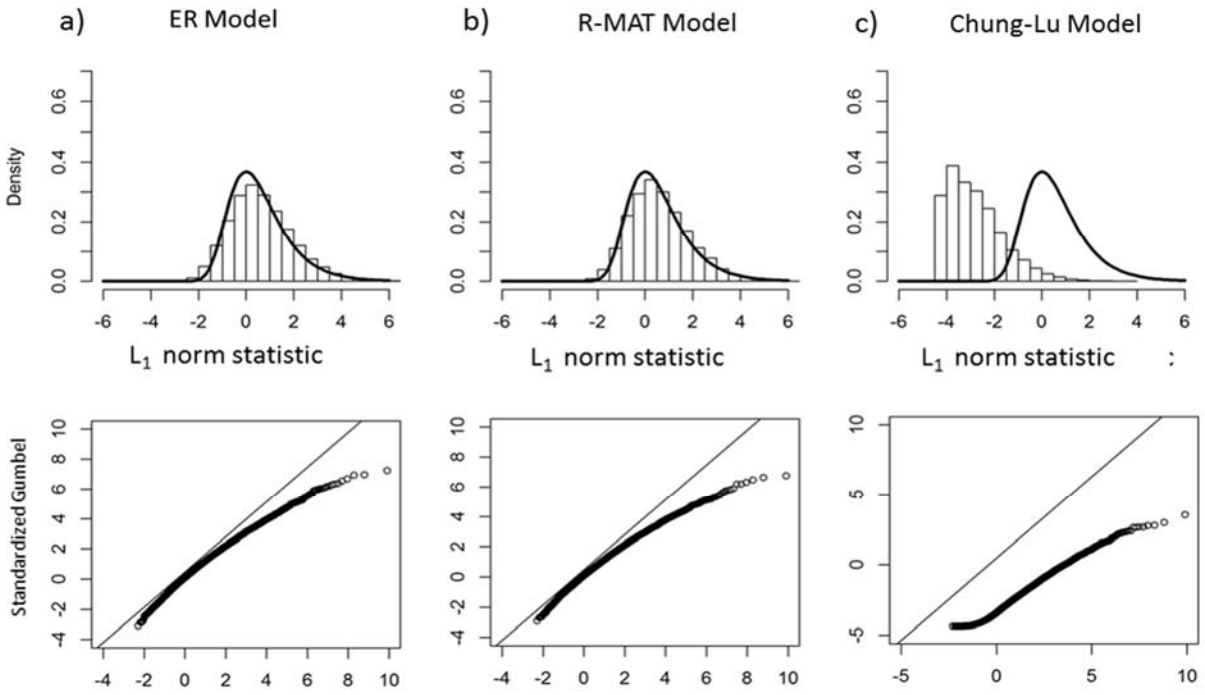

Figure 36. ((a) Erdös-Rényi , (b) R-MAT, and (c) Chung-Lu Model) Top figures are histogram density plots when parameters a_m and b_m are estimated using historical data with $m < n$. Bottom figures are the Q-Q plots of the simulated statistics

Table 9. Percentiles of the L_1 norm based on 10,000 simulations with no anomalous subgraph present. The results are compared to the theoretical Gumbel distribution when $m = 30$ for $n = 128, 256$ and $m = 50$ for $n = 512, 1024$. Scaling parameters a_m and b_m are estimated from historical data using MOM estimators

		ER Model					R-MAT Model					Chung-Lu Model				
Network Size	p_0	*95%*	*96%*	*97%*	*98%*	*99%*	*95%*	*96%*	*97%*	*98%*	*99%*	*95%*	*96%*	*97%*	*98%*	*99%*
Standard Gumbel		**2.97**	**3.20**	**3.49**	**3.90**	**4.60**	**2.97**	**3.20**	**3.49**	**3.90**	**4.60**	**2.97**	**3.20**	**3.49**	**3.90**	**4.60**
128	0.050	2.94	3.13	3.40	3.81	4.33	3.04	3.29	3.62	4.12	5.18	2.63	2.83	3.09	3.57	4.37
128	0.100	2.86	3.04	3.34	3.70	4.17	2.81	3.00	3.24	3.54	4.10	2.19	2.38	2.65	2.95	3.43
128	0.300	3.02	3.22	3.44	3.76	4.33	2.80	2.99	3.22	3.50	3.96	2.12	2.33	2.63	2.95	3.47
256	0.010	2.81	3.02	3.31	3.82	4.51	3.31	3.64	4.02	4.67	5.88	4.41	4.89	5.40	6.18	7.47
256	0.100	2.92	3.11	3.34	3.61	4.17	2.93	3.10	3.35	3.68	4.23	1.71	1.88	2.11	2.40	2.98
256	0.300	3.00	3.15	3.38	3.72	4.27	2.86	3.06	3.32	3.60	4.16	1.29	1.49	1.72	2.03	2.45
512	0.010	3.05	3.22	3.50	3.90	4.49	2.96	3.17	3.48	3.88	4.54	4.55	4.73	4.97	5.34	5.86
512	0.100	3.09	3.27	3.50	3.86	4.39	2.91	3.10	3.33	3.69	4.16	-0.51	-0.30	-0.07	0.25	0.81
512	0.300	2.87	3.05	3.25	3.68	4.30	2.98	3.17	3.43	3.79	4.30	0.20	0.36	0.59	0.84	1.34
1024	0.010	2.74	2.93	3.14	3.52	3.98	2.84	3.02	3.26	3.60	4.03	9.13	9.35	9.65	10.06	10.71
1024	0.100	3.05	3.25	3.50	3.80	4.34	2.90	3.09	3.35	3.69	4.34	0.24	0.31	0.42	0.58	0.79
1024	0.300	3.05	3.26	3.49	3.80	4.34	2.91	3.07	3.31	3.64	4.06	-0.41	-0.18	0.05	0.37	1.05

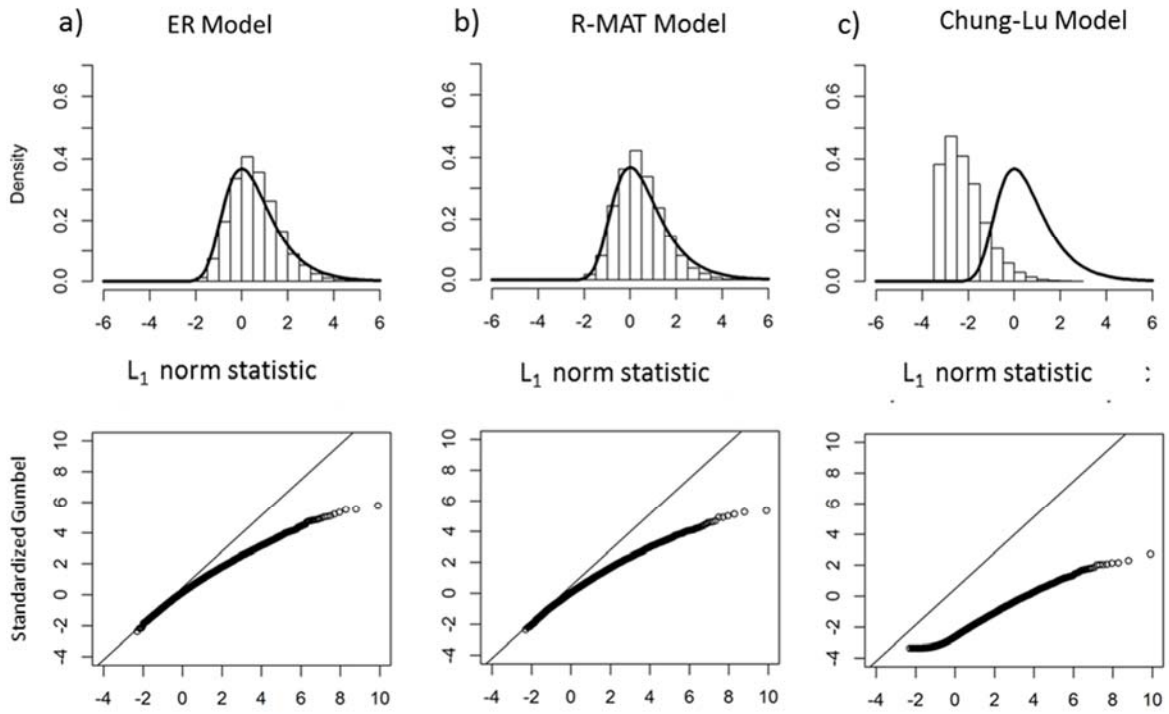

Figure 37. ((a) Erdös-Rényi , (b) R-MAT, and (c) Chung-Lu Model) Top figures are histogram density plots and parameters a_m and b_m are estimated using Extreme Value Theorem with $m < n$. Bottom figures are the Q-Q plots of the simulation

Table 10. Percentiles of the L_1 norm based on 10,000 simulations with no anomalous subgraph present. The results are compared to the theoretical Gumbel distribution when m = 30 for n = 128, 256 and m = 50 for = 512, 1024. Scaling parameters a_m and b_m are estimated from historical data using the Extreme Value Theorem

		ER Model					R-MAT Model					Chung-Lu Model				
Network Size	p_0	*95%*	*96%*	*97%*	*98%*	*99%*	*95%*	*96%*	*97%*	*98%*	*99%*	*95%*	*96%*	*97%*	*98%*	*99%*
Standard Gumbel		**2.97**	**3.20**	**3.49**	**3.90**	**4.60**	**2.97**	**3.20**	**3.49**	**3.90**	**4.60**	**2.97**	**3.20**	**3.49**	**3.90**	**4.60**
128	0.050	2.70	2.87	3.11	3.47	3.94	3.36	3.63	3.98	4.52	5.66	3.08	3.32	3.65	4.24	5.22
128	0.100	2.55	2.70	2.96	3.27	3.67	2.47	2.63	2.84	3.10	3.58	1.89	2.06	2.28	2.53	2.95
128	0.300	2.35	2.50	2.66	2.91	3.34	2.17	2.32	2.50	2.72	3.08	1.70	1.87	2.10	2.35	2.75
256	0.010	3.66	3.94	4.32	4.99	5.90	5.49	6.02	6.61	7.66	9.56	7.30	8.10	8.97	10.26	12.43
256	0.100	2.41	2.56	2.74	2.95	3.40	2.21	2.35	2.54	2.80	3.22	1.31	1.43	1.61	1.83	2.27
256	0.300	2.28	2.39	2.57	2.82	3.23	2.01	2.17	2.37	2.57	2.99	1.05	1.21	1.39	1.63	1.96
512	0.010	3.02	3.19	3.46	3.84	4.41	3.07	3.29	3.60	4.01	4.67	4.07	4.23	4.45	4.78	5.26
512	0.100	2.50	2.64	2.82	3.11	3.52	2.28	2.44	2.62	2.91	3.28	-0.40	-0.24	-0.07	0.18	0.61
512	0.300	2.37	2.52	2.68	3.02	3.52	2.20	2.35	2.55	2.82	3.22	0.17	0.30	0.47	0.67	1.06
1024	0.010	3.02	3.19	3.46	3.84	4.41	2.50	2.65	2.87	3.16	3.54	7.94	8.12	8.38	8.74	9.30
1024	0.100	2.50	2.64	2.82	3.11	3.52	2.28	2.43	2.64	2.92	3.43	0.20	0.25	0.33	0.44	0.60
1024	0.300	2.37	2.52	2.68	3.02	3.52	2.31	2.45	2.65	2.91	3.26	-0.26	-0.10	0.06	0.29	0.75

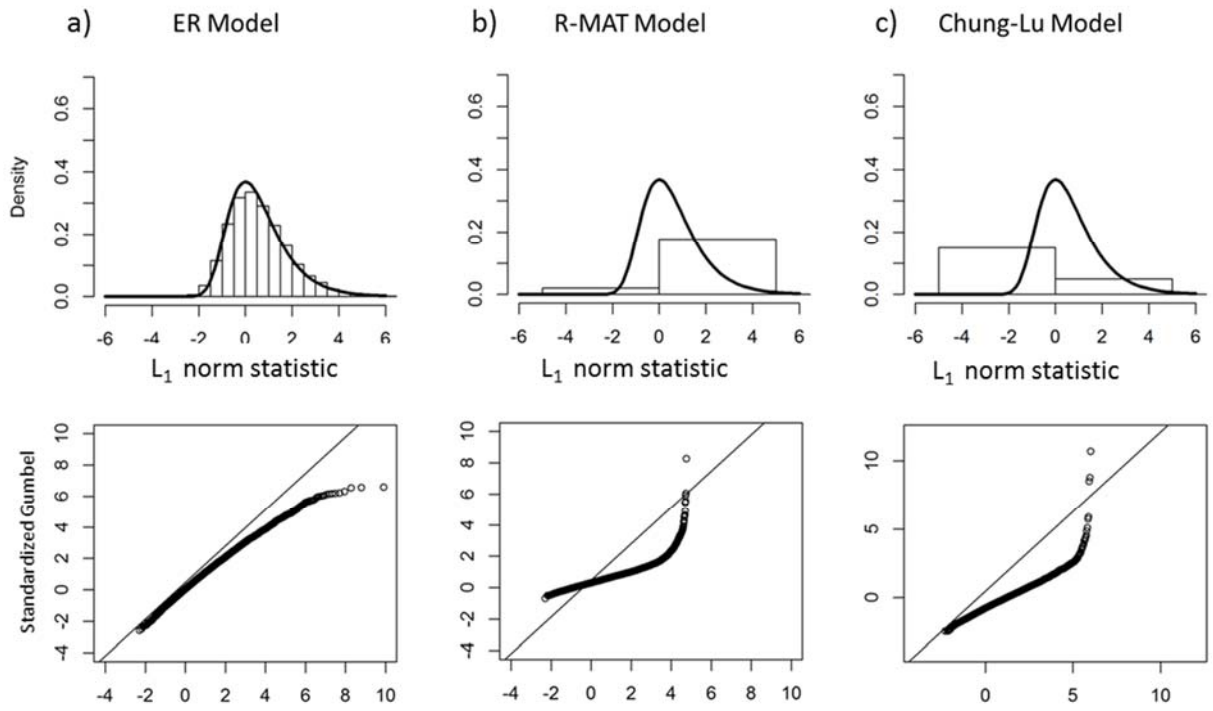

Figure 38. ((a) Erdös-Rényi, (b) R-MAT, and (c) Chung-Lu Model) Top figures are histogram density plots and parameters a_m and b_m are estimated using MOM estimators with m = n. Bottom figures are the Q-Q plots of the simulation

Table 11. Percentiles of the L_1 norm based on 10,000 simulations with no anomalous subgraph present. The results are compared to the theoretical Gumbel distribution when m = n. Scaling parameters a_m and b_m are estimated using the MOM estimator based on historical data

		ER Model					R-MAT Model					Chung-Lu Model				
Network Size	p_0	*95%*	*96%*	*97%*	*98%*	*99%*	*95%*	*96%*	*97%*	*98%*	*99%*	*95%*	*96%*	*97%*	*98%*	*99%*
Standard Gumbel		**2.97**	**3.20**	**3.49**	**3.90**	**4.60**	**2.97**	**3.20**	**3.49**	**3.90**	**4.60**	**2.97**	**3.20**	**3.49**	**3.90**	**4.60**
128	0.050	1.90	2.17	2.74	4.70	5.55	3.84	4.33	4.61	6.26	7.05	13.71	13.86	13.89	13.95	14.19
128	0.100	3.10	3.32	3.59	3.95	4.72	3.35	4.18	4.18	4.73	9.80	16.52	16.52	16.52	16.52	18.35
128	0.300	3.12	3.34	3.57	3.91	4.43	2.02	2.16	2.32	2.59	2.97	2.13	2.30	2.57	2.84	3.34
256	0.010	3.76	4.06	4.49	5.03	5.37	3.01	3.19	3.42	3.96	4.59	6.36	6.36	6.43	6.52	6.74
256	0.100	3.12	3.33	3.59	3.93	4.62	1.57	1.72	4.55	4.91	6.60	16.32	37.62	39.67	39.67	42.61
256	0.300	3.04	3.25	3.56	3.89	4.57	3.02	3.22	3.49	3.87	4.56	2.13	2.31	2.58	2.98	3.56
512	0.010	3.39	3.47	3.72	3.95	6.74	4.32	4.83	5.40	7.02	8.09	15.46	15.48	15.53	15.58	15.72
512	0.100	3.02	3.23	3.45	3.80	4.42	1.41	1.52	1.72	2.18	3.76	1.06	1.20	1.40	1.70	2.20
512	0.300	3.03	3.27	3.52	3.88	4.51	3.10	3.32	3.64	4.01	4.60	2.17	2.37	2.62	2.98	3.58
1024	0.010	1.71	2.06	2.87	4.53	6.94	3.87	4.58	5.46	6.75	15.87	32.81	32.94	33.18	33.36	33.40
1024	0.100	3.07	3.30	3.55	3.89	4.43	3.17	3.41	3.71	4.16	5.28	0.76	0.93	1.22	1.54	2.08
1024	0.300	2.96	3.17	3.46	3.84	4.54	3.09	3.30	3.52	3.82	4.44	2.01	2.19	2.44	2.81	3.36

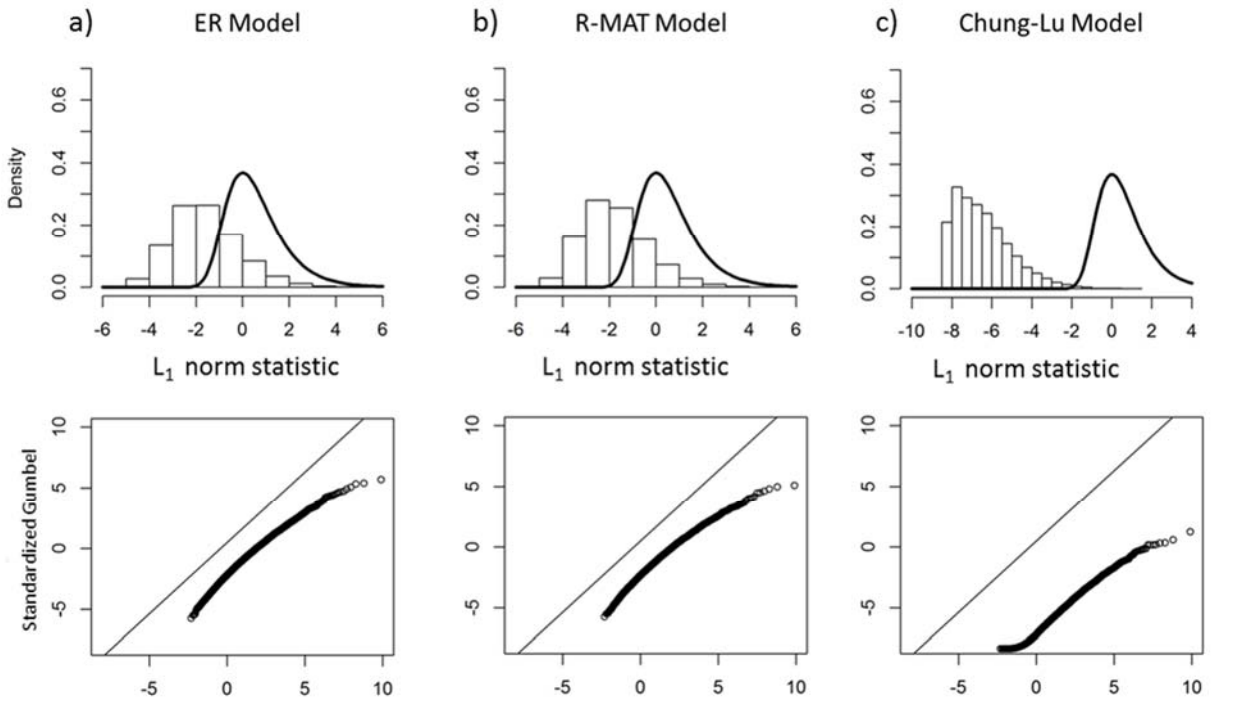

Figure 39. ((a) Erdös-Rényi, (b) R-MAT, and (c) Chung-Lu Model) Top figures are histogram density plots and parameters a_m and b_m are estimated using Extreme Value Theorem with m = n. Bottom figures are the Q-Q plots of the simulation

Table 12. Percentiles of the L_1 norm based on 10,000 simulations with no anomalous subgraph present. The results are compared to the theoretical Gumbel distribution when m = n. Scaling parameters a_m and b_m are estimated using the Extreme Value Theorem

		ER Model					R-MAT Model					Chung-Lu Model				
Network Size	p_0	*95%*	*96%*	*97%*	*98%*	*99%*	*95%*	*96%*	*97%*	*98%*	*99%*	*95%*	*96%*	*97%*	*98%*	*99%*
Standard Gumbel		**2.97**	**3.20**	**3.49**	**3.90**	**4.60**	**2.97**	**3.20**	**3.49**	**3.90**	**4.60**	**2.97**	**3.20**	**3.49**	**3.90**	**4.60**
128	0.050	2.15	2.38	2.69	3.17	3.78	3.01	3.37	3.84	4.55	6.05	2.64	2.97	3.4	4.18	5.47
128	0.100	1.95	2.15	2.49	2.9	3.43	1.85	2.06	2.33	2.68	3.3	1.09	1.3	1.6	1.93	2.47
128	0.300	1.68	1.88	2.1	2.42	2.99	1.45	1.64	1.89	2.17	2.65	0.83	1.05	1.36	1.69	2.21
256	0.010	3.11	3.51	4.07	5.04	6.35	5.77	6.53	7.4	8.91	11.68	8.4	9.56	10.81	12.68	15.83
256	0.100	1.29	1.51	1.78	2.09	2.74	1.01	1.21	1.49	1.86	2.48	-0.3	-0.12	0.14	0.46	1.1
256	0.300	1.1	1.27	1.52	1.89	2.49	0.72	0.95	1.23	1.54	2.14	-0.67	-0.45	-0.18	0.17	0.64
512	0.010	1.85	2.07	2.46	3	3.79	1.92	2.22	2.66	3.24	4.17	3.37	3.62	3.96	4.49	5.23
512	0.100	1.11	1.31	1.57	1.97	2.54	0.81	1.03	1.29	1.68	2.21	-3.67	-3.41	-3.14	-2.75	-2.08
512	0.300	0.94	1.13	1.36	1.85	2.54	0.7	0.9	1.18	1.56	2.12	-2.76	-2.56	-2.29	-1.98	-1.37
1024	0.010	1.03	1.31	1.63	2.2	2.9	0.54	0.77	1.09	1.54	2.11	9.49	9.81	10.24	10.85	11.79
1024	0.100	0.39	0.62	0.92	1.28	1.92	0.21	0.43	0.74	1.16	1.95	-3.58	-3.49	-3.35	-3.17	-2.91
1024	0.300	0.22	0.46	0.72	1.07	1.69	0.25	0.46	0.76	1.16	1.68	-4.35	-4.08	-3.81	-3.43	-2.64

6.9 Statistical properties of chi-square algorithm with no anomalies present

In (Miller, Beard et al. 2015), there is an implicit assumption that the values derived from the chi-square algorithm follow the chi-square distribution. We want to investigate this assumption by comparing the empirical distribution of the chi-square statistic to the theoretical chi-square distribution with *df = 1*. We will therefore investigate the distribution for the chi-square detection statistic by observing its behavior for the case when no anomalous subgraphs are embedded in the network. This is done for multiple network sizes and connectivity combinations.

For our approach, we will first generate 10,000 non-anomalous simulated networks of different sizes and average degree combinations. Then calculate the chi-square statistic for each simulation as described in (Miller, Beard et al. 2015). Next we will compare the statistic to the theoretical chi-square distribution using histogram and Q-Q plots. We also compare the quantiles of the observed chi-square detection statistic and the theoretical chi-square distribution.

6.9.1 Histogram and q-q plots of simulation results

We show in this chapter the network sizes and background probabilities that yielded the most interesting results. The other scenarios we explored are available in the Appendix. The theoretical chi-square distribution is overlaid on the histogram plots to compare their distributions. Alongside, we include the Q-Q plot to better understand how much the algorithm deviates from the theoretical chi-square statistic especially at the higher quantiles.

For the ER Model, we see that for all the network size combinations, the non-anomalous simulation results do not follow the chi-square distribution. Both the histogram plot and the Q-Q plot in Figure 40 reflect this difference.

In Figure 40, we notice that in general, the R-MAT model follows the chi-square distribution better than the ER and Chung-Lu model. This is partly due to the inherent nature of how the R-MAT model is generated. Because edge assignments are based on the Kronecker product, a large number of iterations will yield edge assignment matrices where most cells have practically a zero probability of receiving an edge. This therefore skews the distribution of degrees towards the right as the popular edges dominate in this context. This is appropriate in this case as the chi-square distribution with *df = 1* is also skewed to the right. Whereas, the ER model and Chung-Lu model allows for a more uniform distribution of degrees across the network.

For the Chung-Lu model in Figure 40, the histogram distribution appears to be similar to the ER Model. In the Chung-Lu model, using the Rank-1 approximation of the R-MAT model spreads out the distribution of popular nodes. We also notice the same phenomena when we observe the R-MAT model in Figure 33 as compared to the Chung-Lu model in Figure 34.

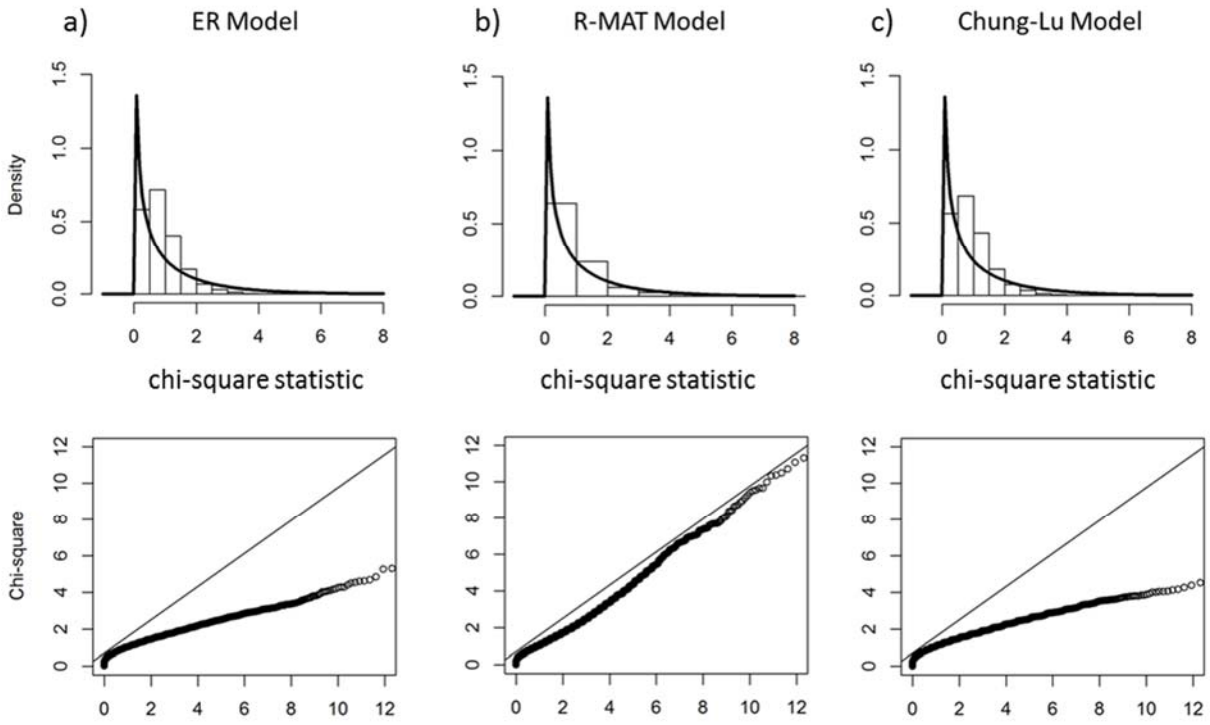

Figure 40. ((a) Erdös-Rényi, (b) R-MAT, and (c) Chung-Lu Model) Top figures Histogram density plots of 10,000 simulations with chi-square distribution, df = 1, overlaid. n = 512. Bottom figures are the Q-Q plots of the simulation

Table 13. (Chi-square distribution) 10,000 non-anomalous simulations are run and the results compared to the χ^2 with df = 1.

		ER Model					R-MAT Model					Chung-Lu Model				
Network Size	p_0	*95%*	*96%*	*97%*	*98%*	*99%*	*95%*	*96%*	*97%*	*98%*	*99%*	*95%*	*96%*	*97%*	*98%*	*99%*
χ^2 with *df = 1*		**3.84**	**4.22**	**4.71**	**5.41**	**6.63**	**3.84**	**4.22**	**4.71**	**5.41**	**6.63**	**3.84**	**4.22**	**4.71**	**5.41**	**6.63**
128	0.050	57.92	59.69	61.25	63.87	67.66	36.11	37.27	38.55	40.11	43.18	20.97	21.70	22.50	23.58	25.39
128	0.100	3.72	3.94	4.28	4.73	5.49	4.97	5.28	5.65	6.23	7.14	4.98	5.19	5.52	5.92	6.53
128	0.300	2.18	2.31	2.44	2.77	3.16	1.99	2.11	2.34	2.57	3.16	2.05	2.30	2.38	2.68	3.02
256	0.010	22.98	24.17	25.80	27.86	31.63	37.04	38.08	39.26	40.76	43.73	25.02	25.79	26.68	28.20	30.68
256	0.100	2.60	2.72	2.94	3.24	3.74	2.70	2.90	3.18	3.51	4.10	2.79	2.91	3.15	3.47	3.93
256	0.300	2.17	2.31	2.45	2.68	3.19	2.36	2.60	2.89	3.22	3.86	2.17	2.33	2.53	2.74	3.22

512	0.010	9.76	10.44	11.39	12.76	15.05	30.03	30.97	32.09	33.89	36.66	21.13	21.86	23.16	24.59	26.97
512	0.100	2.40	2.56	2.76	3.08	3.44	3.08	3.33	3.66	4.12	5.29	2.67	2.82	3.02	3.33	3.81
512	0.300	2.17	2.31	2.48	2.69	3.07	3.33	3.60	4.06	4.89	6.35	2.27	2.34	2.52	2.75	3.15
1024	0.010	6.69	7.23	7.96	9.02	10.96	21.65	22.46	23.23	24.58	27.04	17.22	17.98	19.15	20.23	22.69
1024	0.100	2.28	2.42	2.58	2.81	3.18	3.97	4.44	5.20	6.25	8.52	2.60	2.76	3.01	3.30	3.83
1024	0.300	2.17	2.29	2.43	2.70	3.16	4.54	5.11	5.94	7.11	9.32	2.23	2.41	2.58	2.86	3.32

We also want to observe how the chi-square detection statistic compares with the chi-square theoretical distribution for multiple node and background probability combinations. In (Miller, Beard et al. 2015) it is implicitly assumed that the detection statistic follows the chi-square distribution. We take a look at the observed quantiles after 10,000 simulations with the theoretical quantiles for the chi-square distribution. The tables show that chi-square algorithm detection statistic is dependent on both the network size and background probability. We see that sparse networks, $p_0 < 0.05$, have detection statistic values much higher than the chi-square theoretical quantile values.

This chapter also includes the numerical results in Table 13 that show the differences in results for both the Erdös-Rényi, R-MAT, and Chung-Lu model. It is observed that only a few of the network size and background connectivity combinations we introduced yields quantiles that align with the theoretical chi-square distribution. This again emphasizes our observation that the chi-square detection statistic does not follow the chi-square distribution and a detection value K based on the chi-square theoretical distribution will yield unpredictable results. The model that emulated the chi-square distribution the best is the R-MAT model although it also performs poorly with sparse networks. This is critical to a practitioner as setting the signaling threshold is dependent on the background connectivity of the observed network. Some recommendations for improving its performance is devising a better way to assign points to each quadrant, particularly for sparse networks. This improvement will be explored in the next chapters.

6.10 Evaluating algorithms with anomaly present

The detection and false alarm rates for different network sizes and anomalous subgraph combinations is also explored. We compared these rates for the simple Erdös-Rényi model, the R-MAT model, and the Chung-Lu model and for both the L_1 norm and chi-square algorithm.

We consider two evaluation metrics for the detection rule.

1. False alarm rate (FAR) is P[*signal* | *no anomaly*], i.e., the proportion of cases where no anomaly is present but the detection rule incorrectly signals an anomaly.

2. Detection rate (DR) is P[*signal* | *anomaly*], i.e., the proportion of cases where anomaly is present and the detection rule correctly signals an anomaly.

The resulting confusion matrix for calculating the detection rate (DR) and false alarm rate (FAR) is shown in Table 14

Table 14. Confusion matrix

	Actual Yes	**Actual No**
Predicted Yes	TP	FP
Predicted No	FN	TN

We have

$$DR = \frac{TP}{TP + FN}$$

$$FAR = \frac{TP}{TP + FN}$$

For our approach, we ran 500 simulations for each network size $n = 128, 256, 512, 1024$. For the network size $n = 128$, the background connectivity is chosen to be $p_0 = 0.05$ and for other network sizes, $p_0 = 0.01$. Selecting a higher background connectivity of $p_0 = 0.05$ for the network of size $n = 128$, ensures that the majority of the nodes are indeed connected. If $p_0 = 0.01$, then the average degree of the network would be 1.28 which would result in a network of mostly isolated nodes. For $n = 128$ and 256 network sizes, we randomly imbed subgraphs of 3%, 4%, 5%, and 6% of the network size into 250 of the 500 simulations. For $n = 512$ and 1024 network sizes, we randomly imbed subgraphs of 1%, 2%, 3%, and 4% of the network size into 250 of the 500 simulations. For brevity, only the results for $n = 256$ and $n = 512$ are shown in this section as the other network sizes led to similar results. Each detection and false alarm calculation is performed for the case where $\alpha = 0.05$.

6.10.1 Performance with anomalous subgraph present for eigenvector L_1 norm algorithm

To evaluate the behavior for when an anomaly is embedded of the L_1 norm algorithm, we compare the performance of different L_1 norm calculations, in particular, using the extreme value theorem to estimate the parameters a_m and b_m for different values of m. We are therefore highlighting the performances for when $m < n$ and also when $m = n$. The use of the MOM estimators is ignored in this section as the performance is also comparable to using the extreme

value theorem. Also, to a practitioner, the extreme value theorem is more useful as it does not require historical data of previous L_1 norms in order to estimate the location and scaling parameters, a_m and b_m.

The accompanying Figure 41 and Table 15 illustrate our observations. We see that L_1 norm methodologies perform comparably well in detection and false alarm rates. For all networks observed, the L_1 norm algorithm has false alarm rates that are close to the expected false alarm rate, which is the dashed black line in Figure 41. Furthermore, the detection rates are relatively high for all network sizes investigated and connectivity.

6.10.2 Performance for anomalous subgraph present chi-square algorithm

To compare the anomalous subgraph present behavior - when an anomaly is imbedded - of the chi-square algorithm, we evaluate the performance of theoretical chi-square quantiles. For an $\alpha = 0.05$, the cut of value is 3.841. We notice that the chi-square algorithm results in higher false alarm rates, dashed black line, for all cases. This implies that in practice, a larger proportion of networks will signal that an anomaly is present despite no anomaly being present.

6.10.2.1 Observations

Figure 41 shows the detection and false alarm rates for different network size combinations of the Erdös-Rényi, R-MAT and the Chung-Lu model. For the chi-square algorithm, we notice that in all the cases explored, the false alarm rate from using the chi-square statistic is significantly higher than the expected false alarm rate of 0.05 which is the thick dashed black lines. Although the detection rate is high, having significantly higher false alarm rates than expected results in an algorithm that is difficult to implement in practice. This again highlights that the chi-square distribution does not provide the appropriate detection value for use in anomaly detection. Instead, some method for improving the algorithm is needed.

We observe the same scenario for the R-MAT model in Table 15. For the χ^2 value of 3.84 corresponding to the 95% theoretical chi-square distribution with $df = 1$, the false alarm rates are inconsistent for different network size and background probability combinations. It emphasizes again that the algorithm statistic detection value selected is dependent on the network model being investigated. This is also the same for the Chung-Lu model in Figure 41 as well as Table 15. That is, the chi-square detection value for $\alpha = 0.05$ produces a false alarm rate (FAR) that exceeds the desired FAR rate of 5% in all cases.

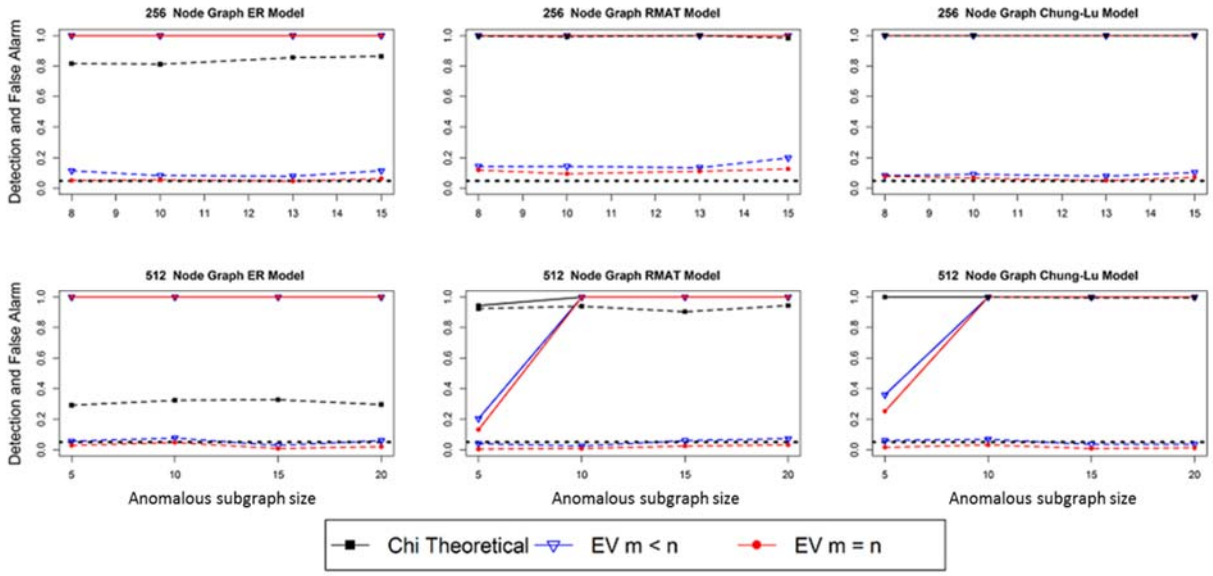

Figure 41. (Erdös-Rényi, R-MAT, and Chung-Lu Model) Detection and False alarm rates with n = 256 and 512. Number of anomalous subgraph varies from 3%, 4%, 5%, and 6% for n = 256 and 3%, 4%, 5%, and 6% for n = 512. Detection rates are solid lines while false alarm rates are dashed lines. Background connectivity, $p_0 = 0.01$

Table 15. Detection and False Alarm Rates. Background probability, p0 = 0.01 and foreground probability, with clique present, is $p_1 = 1$. We perform 500 simulations for each network size and connectivity combination with an anomalous subgraph randomly embedded in 250 of 500 simulations

Erdös-Rényi Model		Detection Rate			False Alarm Rate		
Network Size	Subgraph Size	χ^2Ther. 95% (%)	L_1 EV m < n 95% (%)	L_1 EV m = n 95% (%)	χ^2Ther. 95% (%)	L_1 EV m < n 95% (%)	L_1 EV m = n 95% (%)
256	8	100.00	100.00	100.00	81.60	11.60	5.20
256	10	100.00	100.00	100.00	81.20	8.40	5.60
256	13	100.00	100.00	100.00	85.60	8.00	4.80
256	15	100.00	100.00	100.00	86.40	11.60	6.40
512	5	100.00	100.00	100.00	29.20	5.60	2.80
512	10	100.00	100.00	100.00	32.40	7.60	4.80
512	15	100.00	100.00	100.00	32.80	2.80	0.80
512	20	100.00	100.00	100.00	29.60	6.00	2.00
R-MAT Model		Detection Rate			False Alarm Rate		
256	8	100.00	100.00	100.00	99.60	14.40	12.00
256	10	100.00	100.00	100.00	99.20	14.40	9.60
256	13	100.00	100.00	100.00	100.00	13.60	11.20
256	15	100.00	100.00	100.00	98.40	20.00	12.80
512	5	94.40	20.40	13.20	92.40	4.00	0.40
512	10	100.00	100.00	100.00	94.00	2.40	0.80

512	15	100.00	100.00	100.00	90.40	6.00	2.40
512	20	100.00	100.00	100.00	94.40	7.20	3.20
Chung-Lu Model		Detection Rate			False Alarm Rate		
256	8	100.00	100.00	100.00	100.00	8.40	8.00
256	10	100.00	100.00	100.00	100.00	9.20	6.80
256	13	100.00	100.00	100.00	100.00	8.00	5.20
256	15	100.00	100.00	100.00	100.00	10.40	7.20
512	5	100.00	36.00	25.20	100.00	6.00	1.60
512	10	100.00	100.00	100.00	100.00	6.80	3.20
512	15	100.00	100.00	100.00	99.60	3.60	0.80
512	20	100.00	100.00	100.00	99.60	3.60	1.20

6.11 Special cases and recommendations for improvement

We propose in this section some ideas for improving both the chi-square and L_1 norm algorithms.

6.11.1 Improving the L_1 norm algorithm

One of the concerns when applying the L_1 norm algorithm is determining the number, m, of eigenvectors required for calculating the detection statistic. This becomes a tuning parameter that needs to be accounted for as we observed that this can have an effect on the performance of the algorithm. In the previous simulation results, selecting an m that is too large could lead to higher than expected statistic cut off values for certain network and background connectivity combinations. Table 12, shows the result of using the entire eigenspace, that is, letting m = n where n refers to the network size. These extreme differences between the simulated quantiles and the theoretical distribution are more pronounced in the Chung-Lu model as in Table 12.

Another major concern when implementing the L_1 norm algorithm is the requirement to have historical data in order to estimate the following parameters: location and scaling parameters a_m and b_m, as well as the mean μ_k and standard deviation σ_k. In (Miller, Beard et al. 2015), there is no discussion on how these parameters should be estimated, especially as historical data is needed. However, we suggest estimating the location and scaling parameter, a_m and b_m, using the extreme value theorem as it does not require historical data. To estimate the parameters: μ_k and σ_k, we developed an approach that only requires the current static network.

The L1 norm proposed by (Miller, Beard et al. 2015) is as follows:

$$L = -\min_{1 \le k \le m} \frac{|X_k|_1 - \mu_k}{\sigma_k}$$

where μ_k and σ_k are the mean and standard deviation of the k^{th} eigenvector of the residual matrix. Those are estimated using historical data where no anomalies are known to be present.

As commented in this study, the implementation of this algorithm is impractical for most static networks since historical data is needed. To overcome this issue, we analyze how this statistic performs if $|\mathbf{X}_k|_1$ is standardized using only the *m* eigenvectors of the current network. We analyzed three different standardization approaches and they are as follows:

1. Using mean and standard deviation of (*m*) L_1 norms of the current observation:

(17)

$$L = -\min_{1 \le k \le m} \frac{|X_k|_1 - \mu_m}{\sigma_m}$$

2. Using median and IQR (Inter Quartile Range) of (*m*) L_1 norms of the current observation:

(18)

$$L = -\min_{1 \le k \le m} \frac{|X_k|_1 - M_m}{IQR_m / k_1}$$

where M_m is the median of *m* L_1 norms for the current network, the constant k_1=1.3489, assuming that the eigenvectors of the residual matrix follow a normal distribution.

3. Using median and mad (median absolute deviation) of (m) L_1 norms of the current observation:

(19)

$$L = -\min_{1 \le k \le m} \frac{|X_k|_1 - M_m}{MAD_m / k_2}$$

where M_m is the median of *m* L_1 norms for the current network, the constant k_2 = 0.67449, assuming that the L_1 norms of the eigenvectors follow a normal distribution.

Of these three approaches, using the median, M_m, and inter-quantile range, IQR_m, to estimate the L_1 norm statistic or using, μ_m, and standard deviation, σ_m, worked sufficiently. Using the median and IQR performed the best and this is illustrated in the histogram and Q-Q plots in Figure 42. That is, if an anomalous subgraph is present, the median, M_m, and interquartile range IQR_m, will not be affected which makes this approach appropriate for standardizing the detection

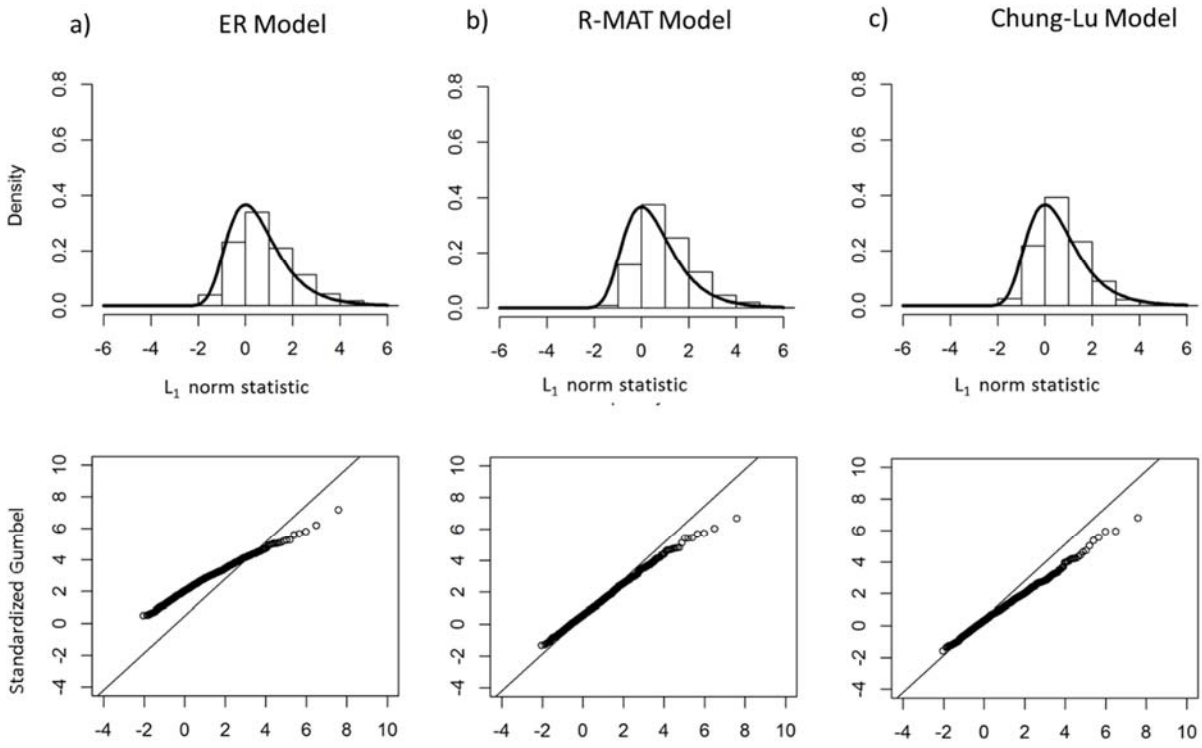

Figure 42. ((a) Erdös-Rényi , (b) R-MAT, and (c) Chung-Lu Model) Top figures Histogram density plots of 10,000 simulations using inter-quantile range, IQR_m, and the median, Mm to standardize detection statistic. Bottom figures are the Q-Q plots of the simulation

statistic. Also, selecting $m < n$ worked the best. We show below the simulation results for both cases as well as the detection and false alarm plots.

Table 16. (L1 norm, m < n, Median and IQR) 10,000 in-control simulations are run and the results compared to the theoretical Gumbel distribution when m = 30 for n = 128, 256 and m = 50 for n = 512, 1024.

		ER Model					R-MAT Model					Chung-Lu Model				
Network Size	p_0	*0.95*	*0.96*	*0.97*	*0.98*	*0.99*	*0.95*	*0.96*	*0.97*	*0.98*	*0.99*	*0.95*	*0.96*	*0.97*	*0.98*	*0.99*
Standard Gumbel		**2.97**	**3.20**	**3.49**	**3.90**	**4.60**	**2.97**	**3.20**	**3.49**	**3.90**	**4.60**	**2.97**	**3.20**	**3.49**	**3.90**	**4.60**
128	0.050	6.55	6.90	7.52	8.56	10.09	3.81	4.05	4.45	4.71	5.14	4.25	4.49	4.92	5.39	6.36
128	0.100	4.30	4.50	5.03	5.86	6.74	3.16	3.39	3.80	4.01	4.49	3.58	3.78	4.07	4.27	5.17
128	0.300	3.26	3.48	3.88	4.26	5.31	3.51	3.70	4.02	4.46	5.40	3.35	3.52	3.88	4.26	5.03
256	0.010	7.37	7.60	8.28	8.79	9.77	4.26	4.41	4.71	5.13	6.14	4.69	4.96	5.30	5.82	6.62
256	0.100	4.13	4.35	4.65	5.06	6.11	2.41	2.56	2.77	3.03	3.86	3.35	3.53	3.76	4.12	4.88
256	0.300	3.53	3.87	4.34	4.90	5.87	3.31	3.43	3.79	4.33	4.77	3.60	4.01	4.26	4.72	5.44
512	0.010	10.53	10.73	11.53	12.18	13.16	3.14	3.27	3.41	3.59	3.83	5.43	5.69	5.94	6.52	7.02
512	0.100	3.17	3.39	3.68	4.08	4.68	3.50	3.68	3.93	4.35	4.76	2.77	2.98	3.26	3.66	4.22
512	0.300	3.18	3.47	3.79	4.16	4.87	4.24	4.57	4.71	5.01	5.76	3.14	3.27	3.51	4.08	4.74
1024	0.010	8.91	9.70	10.21	11.15	13.36	1.48	1.58	1.68	1.82	2.01	3.85	3.98	4.22	4.60	5.02
1024	0.100	3.44	3.65	3.92	4.36	5.09	8.05	8.81	9.42	9.99	10.71	2.45	2.60	2.81	3.36	3.84
1024	0.300	3.29	3.58	3.83	4.30	4.78	7.73	7.98	8.19	8.89	9.43	3.15	3.41	3.67	4.06	4.56

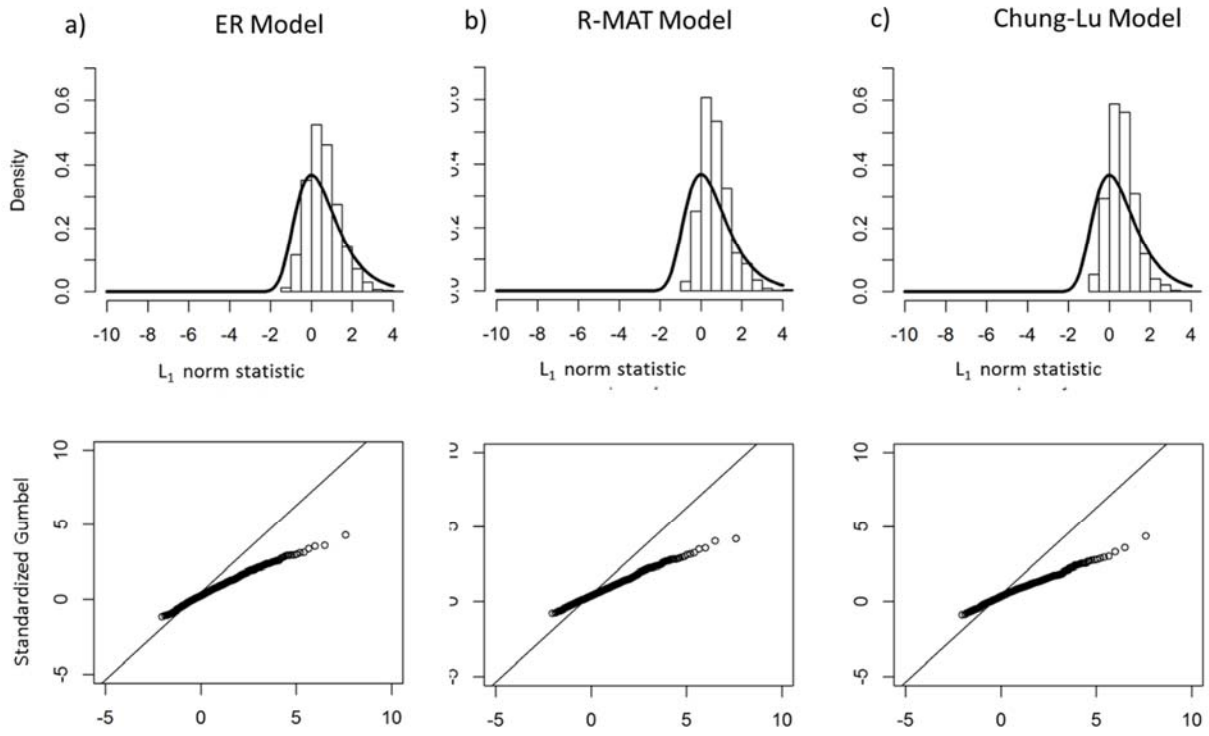

Figure 43. ((a) Erdös-Rényi , (b) R-MAT, and (c) Chung-Lu Model) Top figures Histogram density plots of 10,000 simulations using mean, μm, and the standard deviation, σm, to standardize detection statistic. Bottom figures are the Q-Q plots of the simulation

Table 17. (L1 norm, m < n, Mean and SD) 10,000 in-control simulations are run and the results compared to the theoretical Gumbel distribution when m = 30 for n = 128, 256 and m = 50 for n = 512, 1024.

		ER Model					R-MAT Model					Chung-Lu Model				
Network Size	p_0	*0.95*	*0.96*	*0.97*	*0.98*	*0.99*	*0.95*	*0.96*	*0.97*	*0.98*	*0.99*	*0.95*	*0.96*	*0.97*	*0.98*	*0.99*
Standard Gumbel		**2.97**	**3.20**	**3.49**	**3.90**	**4.60**	**2.97**	**3.20**	**3.49**	**3.90**	**4.60**	**2.97**	**3.20**	**3.49**	**3.90**	**4.60**
128	0.050	3.08	3.21	3.36	3.59	3.81	1.88	1.97	2.06	2.15	2.41	2.05	2.12	2.18	2.25	2.56
128	0.100	2.31	2.38	2.51	2.80	3.16	1.69	1.82	1.94	2.07	2.37	1.90	2.01	2.10	2.25	2.46
128	0.300	1.87	1.97	2.08	2.23	2.33	1.85	1.94	2.07	2.23	2.36	1.72	1.82	2.04	2.21	2.45
256	0.010	3.20	3.30	3.44	3.62	3.83	2.23	2.37	2.48	2.68	3.20	2.36	2.47	2.54	2.71	3.02
256	0.100	2.07	2.20	2.46	2.61	2.89	1.37	1.44	1.51	1.67	1.88	1.72	1.79	1.90	2.12	2.28
256	0.300	1.91	2.00	2.16	2.34	2.65	1.38	1.48	1.67	1.84	2.35	1.72	1.84	2.05	2.27	2.50
512	0.010	4.85	5.00	5.13	5.30	5.80	1.69	1.76	1.83	1.93	2.12	2.97	3.04	3.10	3.20	3.37
512	0.100	2.09	2.19	2.36	2.51	2.91	2.20	2.27	2.36	2.65	2.84	1.77	1.85	2.07	2.40	2.67
512	0.300	2.02	2.10	2.23	2.62	3.12	0.66	0.73	0.82	0.91	1.07	1.84	1.97	2.13	2.27	2.74
1024	0.010	4.84	5.00	5.15	5.43	5.98	1.19	1.24	1.37	1.45	1.54	2.47	2.53	2.67	2.85	3.12
1024	0.100	2.26	2.42	2.49	2.70	3.11	4.25	4.37	4.51	4.64	4.83	1.65	1.73	1.82	1.96	2.22

	1024	0.300	2.00	2.12	2.31	2.53	2.84	0.57	0.60	0.63	0.69	0.74	1.92	2.11	2.34	2.43	2.73

The results of the detection and false alarm rates for all three approaches are also shown in Figure 44. From our analysis, using an *m* between 30 to 50 provides the best results in most of the network combinations we explored where $m = 30$ applies to smaller networks ($n < 257$) and $m = 50$ is suggested for larger networks ($1024 > n > 257$). Also approximating the L_1 norm statistic using the eigenvectors of a single network performs sufficiently as shown in Figure 44.

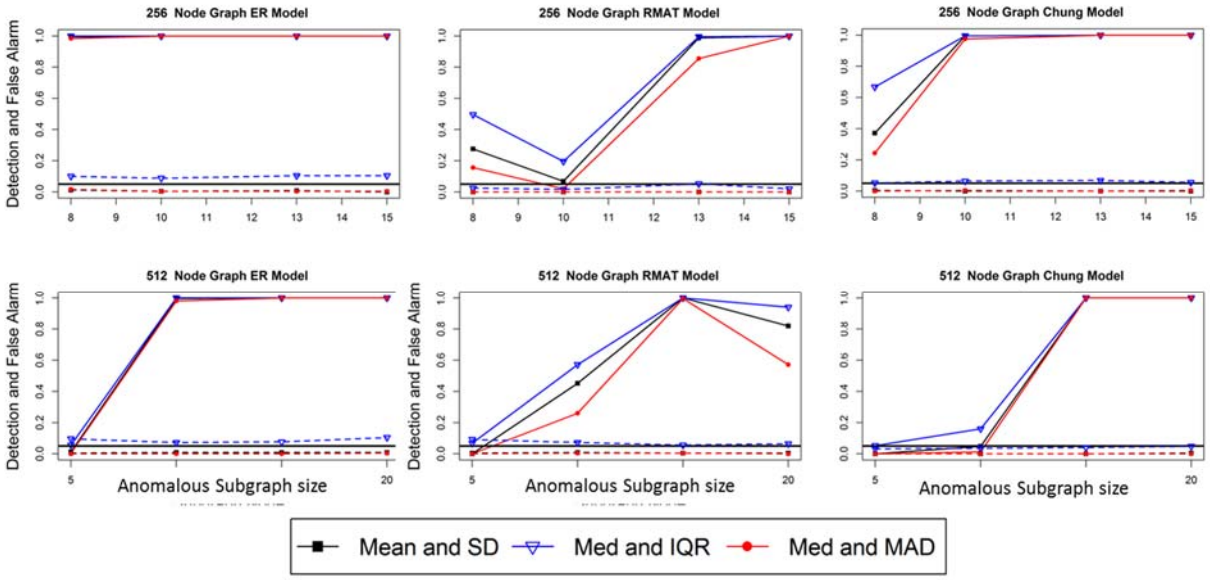

Figure 44. (Erdös-Rényi, R-MAT, and Chung-Lu Model) Detection and False alarm rates with n = 256 and 512. Number of anomalous subgraph varies from 3%, 4%, 5%, and 6% for n = 256 and 3%, 4%, 5%, and 6% for n = 512. Detection rates are solid lines while false alarm rates are dashed lines. Background connectivity, $p_0 = 0.01$

6.12 Improving the Chi-square algorithm

One of the noticeable concerns with the chi-square algorithm proposed in (Miller, Beard et al. 2015) is its poor performance with sparse networks. We observed that in Table 13, the chi-square algorithm particularly has very high statistic values for sparse networks. For $p_0 < 0.01$, the detection static values are about an order of magnitude larger than the theoretical values. We hypothesize that this is due to how points are assigned to a quadrant. In sparse networks, the first two principal components of the residual matrix have a higher proportion of values close to zero. So when plotted, although radial symmetry is maintained, a significant number of points end up near or on the origin. Figure 45a and Figure 46a illustrate this phenomena. In these figures, some

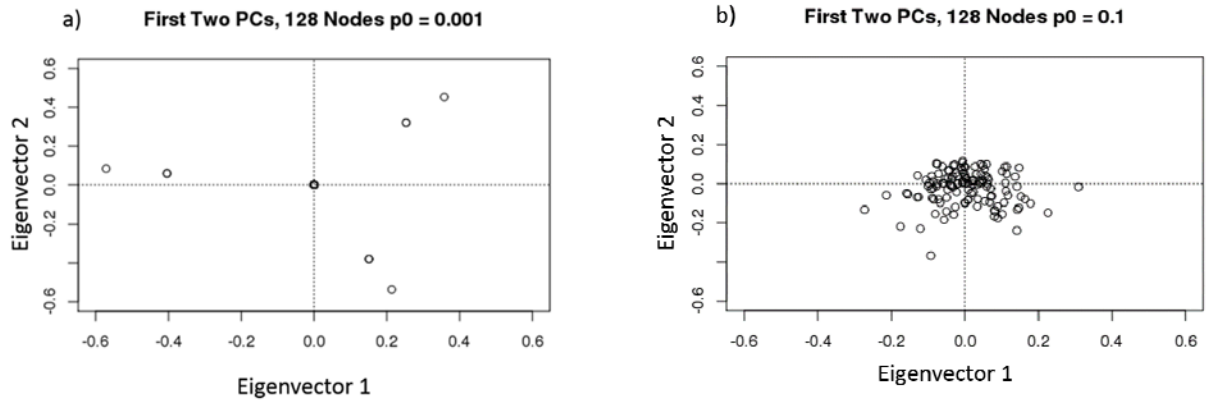

Figure 45. Figure (a) Sparse network with N = 128 and p_0 = 0.001, ER Model. There are 128 points in the plot although most are at the origin. Figure (b) Dense network with p_0 = 0.1 and we observe radial symmetry.

points are in fact on the origin but due to the computational limitations of some spectral decomposition calculations, these values are actually approximations. One result of this is an abundance of points that end up in one particular quadrant. Furthermore, for points that end up right on the origin or one of the axis, there's no methodology to ensure these points are appropriately accounted for. Hence, when assigning points to the 2 X 2 table as the algorithm proposes, there is a tendency for a particular quadrant to be over-represented.

As an example of how the quadrant count is affected, Table 18 and Table 19 show the results when the graph is sparse versus when it is more connected. In Table 18, Q2 and Q4 are over-represented and in Table 19, we notice that Q2 and Q3 are over-represented.

Table 18. Count of points in each quadrant for Figure 45

p_0	**Q1**	**Q2**	**Q3**	**Q4**	**Total**
(a)	5	27	6	90	**128**
(b)	30	32	31	35	**128**

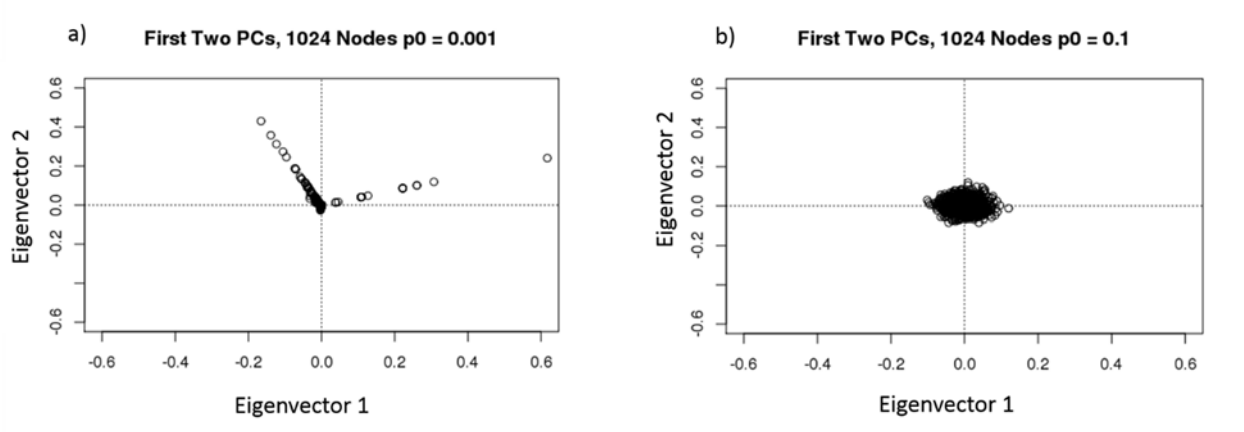

Figure 46. Figure (a) Sparse network with N = 1024 and p_0 = 0.001, ER Model. There are a total of 1024 points in the figure although most are centered at the origin. Figure (b) Dense network with p_0 = 0.1 and we observe radial symmetry.

Table 19. Count of points in each quadrant for Figure 46

p_0	Q1	Q2	Q3	Q4	Total
(a)	15	81	928	1	**1024**
(b)	246	238	285	255	**1024**

It should be noted that this behavior is network size dependent. That is, for the same background connectivity value, the plot of the first two principal components of a larger network tends to be relatively more compact as compared to a smaller network. We observe this in Figure 45 and Figure 46. To verify this, we ran multiple simulations with no anomalies present and observed that the distance of points from the origin is inversely proportional to the square root of the network size. In particular, $\propto \frac{k}{\sqrt{n}}$. Also, we observed that this distance, d, is also inversely proportional to the connectivity of the graph, p_0, that is $d \propto \frac{k}{p_0}$. This relationship is relatively weak when compared to the effect network size has on the average distance of a point from the origin.

What this implies is that we can improve on the performance of the chi-square statistic by allocating points that are close to the origin equally to all four quadrants. We can do this by

specifying that points that are a distance, d from the origin should be approximately equally distributed to all four quadrants. This distance d should be adjusted to compensate for smaller and larger networks. In our improvement, we specify d based on calculating the distances of every point from the origin. Using the relationship that, $d \propto \frac{k}{\sqrt{n}}$, the best performing k value that was observed through simulation results was when $k = 0.35$. This was the k value that worked for the Erdös-Rényi, R-MAT, and Chung-Lu models. This approach also resolves one of the concerns with points lying on an axis. Figure 45 and Figure 46 and empirical observations showed that points a significant distance away from the origin rarely lie on one of the axes.

The top rows of Table 20 shows the simulation results for both the Erdös-Rényi, R-MAT, and Chung-Lu models with no improvements made to the detection statistic. The bottom rows of Table 20 shows the simulation results for both the Erdös-Rényi, R-MAT, and Chung-Lu models with our improved methodology. It is observed that for the improved version, the behavior of having significantly higher detection statistics than expected from the theoretical distribution is limited. This is more apparent for the R-MAT and Chung-Lu models.

Table 20. Simulation results compared to the theoretical chi-square distribution. Results only show the sparse networks for $p_0 = 0.05$ when $n = 128$ and $p_0 = 0.01$ for other network sizes. Includes both the statistics without any improvements, top rows, and algorithm results with improvement

No improvements added		ER Model					R-MAT Model					Chung-Lu Model				
Network Size	p_0	*95%*	*96%*	*97%*	*98%*	*99%*	*95%*	*96%*	*97%*	*98%*	*99%*	*95%*	*96%*	*97%*	*98%*	*99%*
χ^2 *with df = 1*		**3.84**	**4.22**	**4.71**	**5.41**	**6.63**	**3.84**	**4.22**	**4.71**	**5.41**	**6.63**	**3.84**	**4.22**	**4.71**	**5.41**	**6.63**
128	0.050	3.72	3.94	4.28	4.73	5.49	4.97	5.28	5.65	6.23	7.14	4.98	5.19	5.52	5.92	6.53
256	0.010	22.98	24.17	25.80	27.86	31.63	37.04	38.08	39.26	40.76	43.73	25.02	25.79	26.68	28.20	30.68
512	0.010	9.76	10.44	11.39	12.76	15.05	30.03	30.97	32.09	33.89	36.66	21.13	21.86	23.16	24.59	26.97
1024	0.010	6.69	7.23	7.96	9.02	10.96	21.65	22.46	23.23	24.58	27.04	17.22	17.98	19.15	20.23	22.69
Improvement added		ER Model					R-MAT Model					Chung-Lu Model				
χ^2 *with df = 1*		**3.84**	**4.22**	**4.71**	**5.41**	**6.63**	**3.84**	**4.22**	**4.71**	**5.41**	**6.63**	**3.84**	**4.22**	**4.71**	**5.41**	**6.63**
128	0.050	3.61	3.80	4.17	4.58	5.34	3.10	3.25	3.56	3.91	4.41	2.55	2.70	2.86	3.14	3.57
256	0.010	11.58	12.47	13.49	14.95	16.88	6.06	6.41	6.97	7.68	8.69	3.87	4.14	4.39	4.84	5.63
512	0.010	9.02	9.70	10.50	11.89	14.04	6.68	7.03	7.48	8.14	9.21	5.65	6.02	6.39	7.02	8.00
1024	0.010	6.43	7.06	7.75	8.77	10.45	6.59	6.91	7.34	8.03	9.06	8.47	9.11	9.75	10.65	12.28

6.12.1 Performance with anomalous subgraph present

For this section, we ran 500 simulations where 250 out of the 500 simulations have an anomalous subgraph embedded. We also compare the performance of the chi-square algorithm with the revised algorithm. Figure 47 and Table 21 show that the improved chi-square algorithm retains the same detection power while significantly reducing the false alarm rates. This is more apparent in the R-MAT and Chung-Lu models.

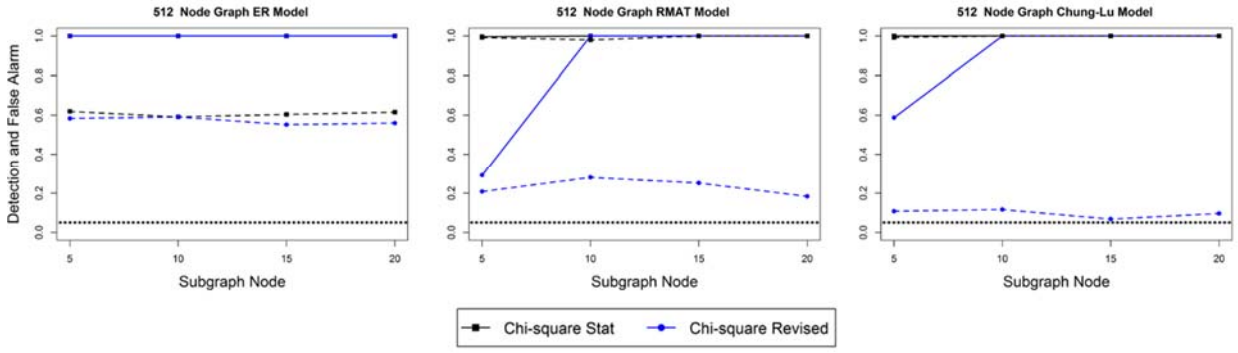

Figure 47. (Erdös-Rényi, R-MAT, and Chung-Lu Model) Number of anomalous subgraph varies from 1%, 2%, 3%, and 4% for n = 512. Detection rates are solid lines while false alarm rates are dashed lines. Background connectivity, $p_0 = 0.01$. A comparison of the traditional detection statistic and the improved version

Table 21. Detection and False Alarm Rates, Erdös-Rényi Model. Background probability, $p_0 = 0.05$ for $n = 128$ and $p_0 = 0.01$ for other network sizes. Foreground probability is $p_1 = 1$. We perform 500 simulations for each row with an anomalous subgraph randomly embedded in 250 of 500 simulations

ER Model		Detection Rate		False Alarm Rate	
	Subgraph Size	χ^2Ther. 95% (%)	χ^2 Revised 95% (%)	χ^2Ther. 95% (%)	χ^2 Revised 95% (%)
512	5	100	100	62	58.4
512	10	100	100	59.2	59.2
512	15	100	100	60.4	55.2
512	20	100	100	61.6	56
R-MAT Model		Detection Rate		False Alarm Rate	
	Subgraph Size	χ^2Ther. 95% (%)	χ^2 Revised 95% (%)	χ^2Ther. 95% (%)	χ^2 Revised 95% (%)
512	5	99.6	29.2	99.2	20.8
512	10	100	100	98	28
512	15	100	100	100	25.2
512	20	100	100	100	18.4
Chung-Lu Model		Detection Rate		False Alarm Rate	
	Subgraph Size	χ^2Ther. 95% (%)	χ^2 Revised 95% (%)	χ^2Ther. 95% (%)	χ^2 Revised 95% (%)
512	5	100	58.8	99.2	10.8
512	10	100	100	100	11.6
512	15	100	100	100	6.8
512	20	100	100	100	9.6

6.13 Discussions and future works

In our work, we evaluated two algorithms proposed in Miller et al., for anomaly detection in binary, static networks. These are the chi-square algorithm and L_1 norm algorithm. It is assumed in (Miller, Beard et al. 2015) that the chi-square algorithm has detection statistic values that follow the χ^2 distribution with $df = 1$, while the L_1 norm algorithm has detection statistic values that follow the Gumbel distribution. We show that this is not the case by comparing the quantiles obtained from multiple simulation results where we compare the detection statistic values to their respective theoretical distributions. Specifically, we show that the distribution of these values are affected by the connectivity of the network, the network size, and the network model.

These inconsistencies, such as the different behaviors when applied to different network sizes, network models, and connectivity means the algorithms are impractical to a practitioner. For example, it is difficult to establish a signaling detection value, K, due to these inconsistencies. We also show that because the L_1 norm algorithm requires historical data for implementation, it is unsuitable for a practitioner to use for most static networks.

Many of these concerns are addressed in our work. We introduce improvements to the chi-square algorithm that improve its performance in sparse networks. The resulting detection and false alarm rates after our improvements are added show that our recommendations are advantageous over the current algorithm. We also proposed a way of standardizing the L_1 norm statistic that requires only the currently observed network. We compare the effects of our improvements to the theoretical statistic distributions and show that they perform sufficiently.

Finally, we extend these algorithms to count networks, an area of importance to practitioners but something not investigated in Miller et al. The algorithms along with our improvements perform sufficiently when applied to count networks and the same conclusions were obtained. We conclude with the comment that statistical evaluation of anomaly detection methods is an important research area where little work has been done, and we encourage more work in this important direction. Future research will look into further extending these algorithms to dynamic networks.

7 DEVELOPED A SIDE-CHANNEL TECHNIQUE FOR ANOMALY DETECTION IN A MANUFACTURING PROCESS

7.1 Overview of chapter

This work presents a novel approach to non-destructive evaluation by mounting a PZT to a fixture and the fixture-part combination is excited. Two other mounting methods are also explored in this work, (1) Directly mounting the PZT to the part, and (2) Magnetic mounting. Tests were conducted on parts with varying degrees of simulated changes and the performance of different mounting methods compared using both traditional damage metrics and a newly developed location based damage metric. Results show that normal parts have a quality loss metric of 14.29% in comparison to quality loss metrics of 73.81%, 98.81%, and 70.24% for the three altered groups. These results show that mounting a PZT to a fixture has potential for non-

destructive evaluation in manufacturing settings. It should be noted that a paper on this topic has been submitted to the Journal of Manufacturing Systems and reviewer comments addressed (Komolafe *et. al 2018*)

7.2 Significance of PZT research to manufacturing

There are two main contributions from this study. First, a novel mounting method is investigated that uses an instrumented fixture for non-destructive evaluation (NDE) of manufactured parts. A PZT is directly mounted to a fixture and the fixture is interrogated with different parts affixed to it. This allows for multiple interrogations of different parts using the same measurement system as parts can be interchanged in the fixture. This study also compares the effectiveness of this method to other mounting methods. The benefits and drawbacks of all mounting methods are investigated and discussed. This study shows that attaching PZTs directly to a fixture has the potential to be used in current NDE processes. Second, a new quality loss metric is introduced for NDE. This novel metric is compared to other widely accepted quality loss metrics used in impedance-based non-destructive evaluation. This study shows that the new quality loss metric performs better in distinguishing between different types of structural changes.

The remainder of this chapter is organized as follows. The experimental design, materials, and methods are introduced in the next section. Following this, a new quality loss metric is developed and its performance compared to other traditional quality loss metrics. The results of the experiments are presented. Finally, potential improvements to the fixturing design are discussed and future research directions are provided in the final section.

7.3 Materials and methods

In this section, the part design, experimental setup, and fixture design are discussed. The part design includes the type of material used to manufacture each of the parts and the overall structural changes added to each part group. In the experimental set up section, the different mounting methods are discussed and the materials used for each study. Finally, the fixture design section describes the key design considerations of the fixture.

7.3.1 Part design

To compare the performance of different mounting methods, various levels of mass change are introduced to the reference part design. Tests were performed on twelve 4.00x2.00x1.00 in^3 cold-finished steel blocks. The twelve machined blocks are separated into four groups **A**, B, C, and D, corresponding to the amount of mass change. Based on the relationship between the mass

of the structure its stiffness, and the impedance of the structure, Z_s, as described in (Liang, Sun et al. 1997), there should be differences between the impedance measurements, $Z_{s,}$, in the different groups. That is, for each percentage mass change introduced in each part group, the parts mass and stiffness will also be affected. Therefore the observed mechanical impedance of the structure will also change. These differences are detected using an impedance analyzer as the impedance, $Z_{s,}$ is directly related to the electric admittance, $Y(\omega)$, as in equation (1).

To induce this mass change, slots of varying depths are introduced to some of the machined blocks. The machined blocks weigh 2.28lbs when no slots are added. The percent mass changes are shown in Table 1.

Table 22. Groups of parts and description

Group (3X)	Feature depth (*in*)	Slot width (*in*)	Mass Change (%)
A	0	0.5	0
B	0.1	0.5	2.5
C	0.2	0.5	5
D	0.3	0.5	7.5

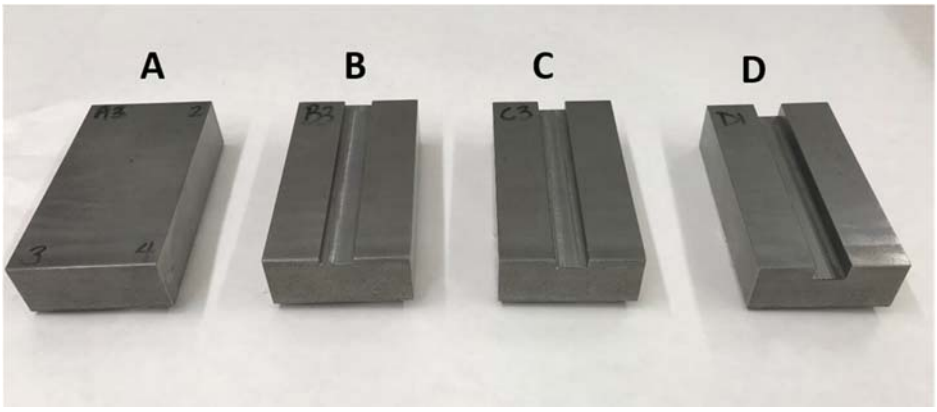

Figure 48. Four groups of machined blocks were created, A - no slot, B - 0.1" slot, C - 0.2" slot, D - 0.3" slot

7.3.2 Experimental setup

The piezoelectric transducer is 1.00in diameter and 0.08in thick and obtained from American Piezoelectric International, Ltd. . Fifteen transducers in total are used in this study: twelve for direct mounting onto the twelve parts, one for magnetic mounting, and two for mounting onto the fixture. For direct-mounting experiments, the consistency of measurements between the transducers is a main concern hence only transducers with similar profiles are used. Inconsistencies

are mainly due to variations in material properties and manufacturing tolerances. Soldering and bonding (for those cases where PZTs are directly mounted on parts) also introduce some minor discrepancies.

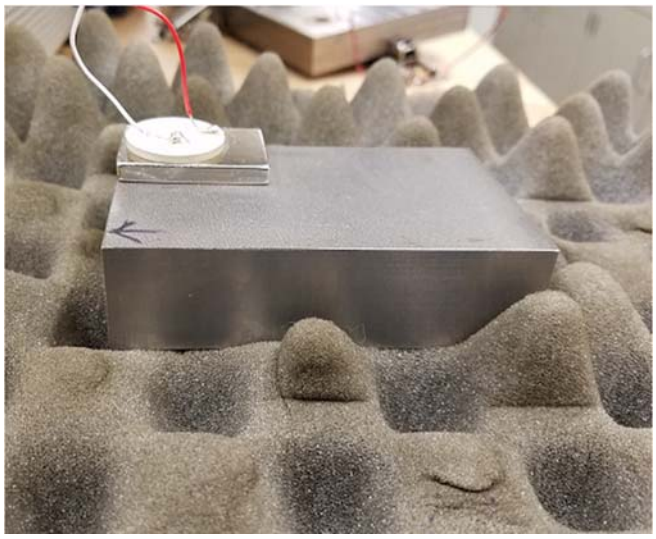

Figure 49. Magnet with PZT mounted onto part. System placed on foam to better approximate free boundary conditions

For direct mounting onto the parts, the PZTs were bonded using a thin film of cyanoacrylate glue. This is a common mounting technique for PZTs in NDE (da Silveira, Campeiro et al. 2017). The PZT is mounted to the same corners of all 12 blocks to ensure consistency in measurements.

To mount using magnets, a 35 Neodymium magnet with a pull force of 23.25lb and dimensions of 1.50x1.00x0.18 in^3. The PZT is bonded using a thin film of cyanoacrylate glue. This allows reuse of the PZT for multiple tests. An illustration of the mounting technique is shown in Figure 49.

7.3.3 Fixture design

The third mounting method explored in this study is an instrumented fixture. The goal is to detect the structural changes on the part from the fixture and part combination. To achieve a highly repeatable and reproducible measurement system, the following attributes are considered in the fixture design:

a) **Steel**: The top and bottom plates of the fixture are made from 1018 cold rolled steel to achieve high rigidity. This minimizes bending of the plates once the draw latches are engaged. An added benefit of this steel type is it is suitable for machining operations.
b) **Crevice**: **A** crevice is added to the bottom plate to ensure parts are positioned at the same location for all measurements.
c) **Guiding Pins**: Four guiding pins ensure the top and bottom plates remained aligned for all measurements.

d) **Bearings**: Four bearings are installed on the guiding pins to improve usability of the fixture. The bearings also reduce the tension placed on the guiding pins from the top plate being moved up and down.

e) **Tight-hold Draw Latch**: Four draw latches made from stainless steel exert equal and sufficient force on the top plate.

The different attributes can be seen in

Figure 50a and

Figure 50b below. The dimension of the top and bottom plates are both 8.00x6.00x0.75 in^3. The fixture weighs 25.12lbs without the part loaded.

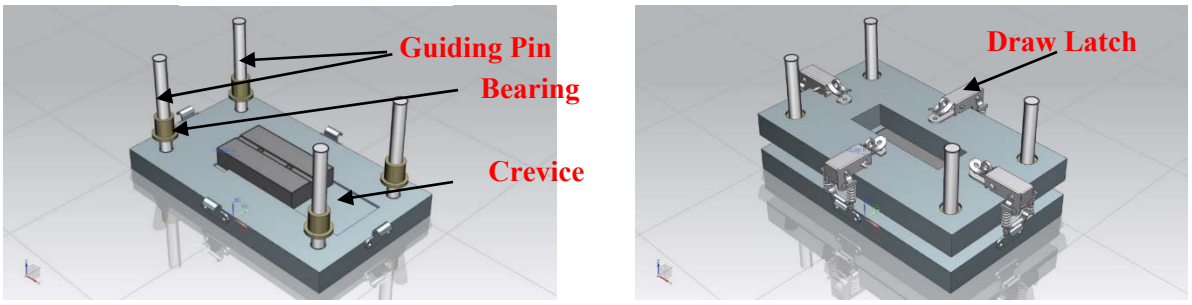

Figure 50. Computer generated model of fixture.

Figure 50a (left) Model of bottom plate of fixture with part sitting in crevice.

Figure 50b (right) Model of entire fixture with springs engaged

An HP Agilent Impedance analyzer, model no. E4990A, is used for excitation of the PZTs and to measure the impedance signatures. Figure 51 shows the analyzer connected to the PZTs attached to the fixture. PZTs are excited from 10 kHz to 100 kHz. Excitation beyond 100 kHz is not considered in this study as the signal is dominated by either the dynamic response of the fixture itself or more affected by soldering/bonding imperfections (Park, Cudney et al. 2000, Wong, Du

et al. 2015). These regions are excluded since the aim is to capture global effects, such as mass changes between different parts.

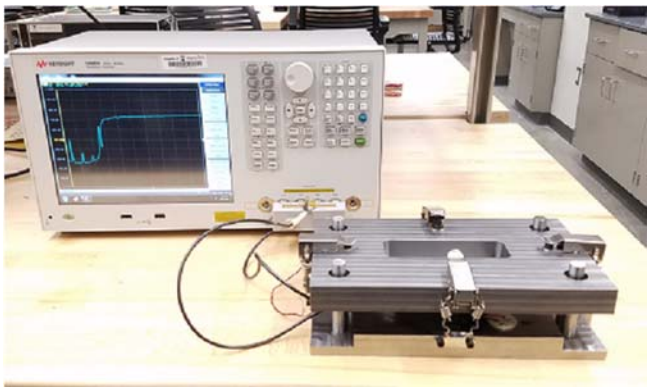

Figure 51. Impedance analyzer and Fixture

7.4 Peak location based damage metric

The common quality-loss metrics used for damage quantification in impedance-based NDE practices are Root Mean Square Deviation (RMSD) and correlation coefficients (Bray and Stanley 1996, Park, Cudney et al. 2000). Both methods treat each frequency of excitation with equal importance including frequencies that could be attributed to variations within the measurement system (Bray and Stanley 1996, Park, Cudney et al. 2000). In this application, the ability to distinguish between different parts is critical. Therefore a quality loss metric is needed that limits the influence of variation due to the measuring system and further highlights differences attributable to the individual parts. To address these concerns the novel quality loss function uses the peak locations in the calculation of the damage metric. Previous research has shown that the location of the peaks (Carden and Fanning 2004) is critical for quality loss evaluation in impedance-based NDE.

A number of researchers have used overall frequency shifts for quality loss measurements in PZT applications (Carden and Fanning 2004). Kessler et al. looked into the ability to use frequency shifts for detecting multiple types of damage to the same part (Kessler, Spearing et al. 2002). De Roeck et al. showed that overall stiffness degradation of a bridge could signal after a frequency shift of as low as 1% (De Roeck, Peeters et al. 2000). Carden et al. applied a Autoregressive Moving Average (ARMA) time series approach to tracking the deterioration of a structure over time (Carden and Brownjohn 2008). A more in-depth discussion of the use of frequency shifts for damage quantification can be found in (Carden and Fanning 2004).

Other techniques attempt to incorporate the peak location. Winston et al. obtained peak locations, width, and amplitude of the baseline structure. For an induced damage to the structure, variation statistics where then calculated for corresponding peaks, which might have shifted due to the induced damage (Winston, Sun et al. 2000). Albakri et al. used peak locations to model the location of a simulated damage on a part (Albakri and Tarazaga 2017). Khajeh et al. used the peak location and the number of peaks to replicate the type of damage introduced to a structure (Khajeh and Koma 2007).

In the proposed approach, the peak locations are the only defining metric used for quality loss. A peak is defined as an impedance measurement with a value above a specified threshold from its neighboring points. Two tuning parameters are introduced: the threshold, **T**, and window size, **W**. The threshold, **T**, is how much greater the impedance at that location must be in comparison to the average of its neighboring points. The window, **W**, is how many neighboring points to consider. To account for smooth peaks, neighboring points are points that are at least two measurements away from the peak point. These two parameters are tuned for the different mounting methods used in this study.

The specific frequency range of excitation is also required for each of the mounting methods. This frequency range is selected to minimize effects that could be attributed to soldering or bonding variations. Also, this application focuses on frequency ranges where the baseline parts have similar peak locations. Similar peak locations are obtained by calculating the quality loss metrics for the control parts and identifying frequency ranges where the responses of the parts are similar.

For *m* identified peaks in the baseline structure and *n* identified peaks in the newly interrogated structure, the formula for peak location quality loss quantification is given by:

$$\boldsymbol{Q_{loss}} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} |\boldsymbol{F_{base,i}} - \boldsymbol{F_{x,i}}| + \sum_{\boldsymbol{j=1}}^{\boldsymbol{n}} |\boldsymbol{F_{base,j}} - \boldsymbol{F_{x,j}}| \qquad (2)$$

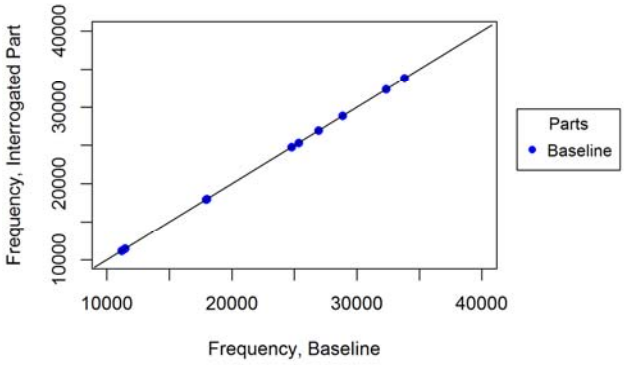

Figure 52. Plot of baseline which is Group A, Replicate 1, versus its two other measurements.

where $\boldsymbol{F_{base,i}}$ is the i^{th} peak of the baseline structure and $\boldsymbol{F_{x,i}}$ is the closest peak to the baseline's i^{th} peak. The interrogated structure has a peak j along the frequency range of interest and this is compared to the location of the closest baseline peak. The metric accounts for instances when the interrogated structure might have multiple peaks at the window of interest, as in, when $m \neq n$.

To illustrate the benefit of observing just the peak locations, a comparison of the location of peaks for the different groups of parts is shown. If peak locations of the groups of parts perfectly align, then a plot of their locations should align perfectly on the y = x line. The peak locations for the baseline, which for these comparisons is Group A, replicate 1, measurement 1 is shown in Figure 52. It also includes measurement 1 in comparison to measurements 2 and 3. It is shown in Figure 53a that the peak locations for all parts in group A, have peak locations that align with the baseline. For the other parts, when a slot is added, these peak locations deviate from the y = x line as in Figure 53b, Figure 53c, and Figure 53d.

To further illustrate the benefit of using peak locations for quality loss calculations in manufactured parts, the performance of the various quality loss metrics is compared. In particular, the results of using the peak location, RMSD, and correlation coefficient are evaluated when the PZT is directly mounted onto the part. This is the data set that should show the clearest distinction between the different part groups as the impedance signal consists entirely of the part and the PZT.

Figure 6a

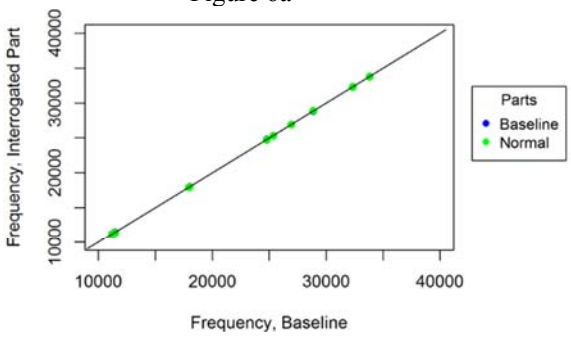

Figure 6b

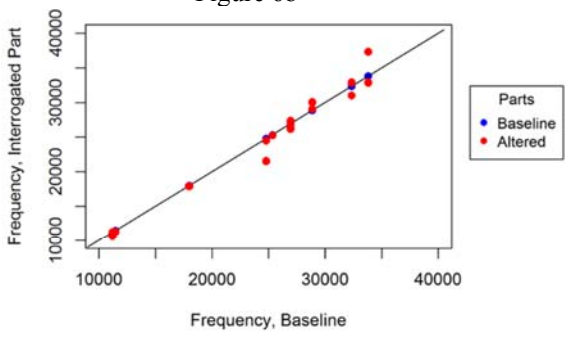

Figure 6c

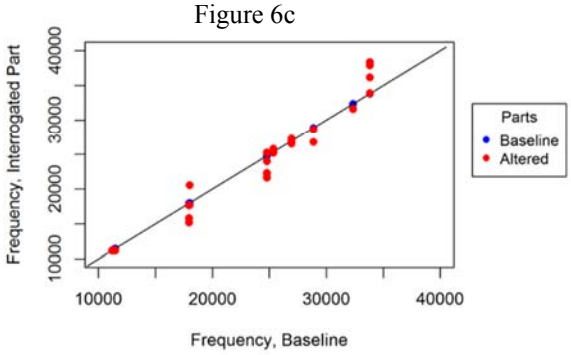

Figure 6d

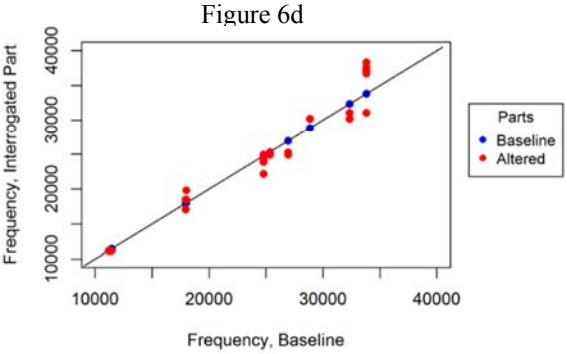

Figure 53. Comparison of peak locations for all groups. Points will overlap exactly on the y = x line if they are similar to the baseline measurement. Top left (Figure 53a) is the overlay of replicates in Group A with the baseline. Top right (Figure 53b) is the overlay of replicates in Group B with the baseline. Bottom left (Figure 53c) is the overlay of replicates in Group C with the baseline. Bottom right (Figure 53d) is the overlay of replicates in Group D with the baseline.

Figure 8a

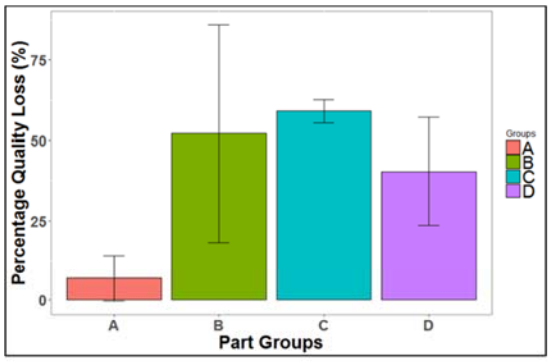

Figure 8b

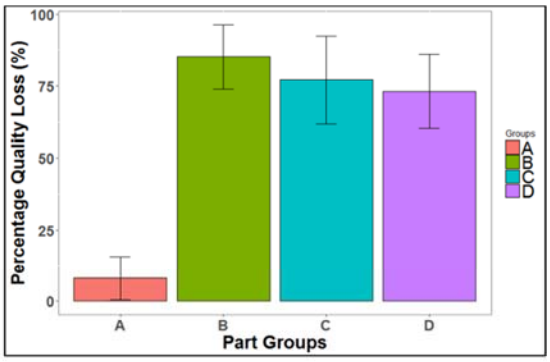

Figure 8c

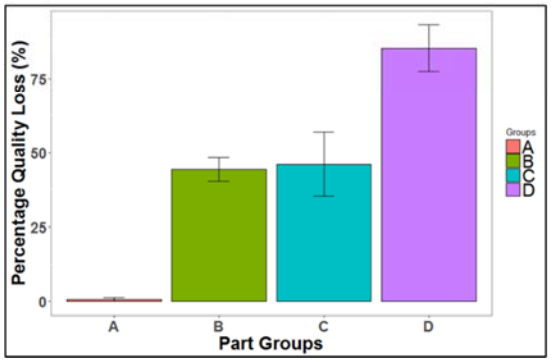

Figure 54. Bar graphs of percentage quality loss for each part group. Figure 54a (top left) result when RMSD is used. Figure 54b (top right) result when correlation coefficient is used. Figure 54c (bottom left) result when peak location metric is used

The bar graphs in Figure 54a, Figure 54b, and Figure 54c reflect the percent quality loss for each group for the RMSD, Correlation Coefficient (corr coeff), and peak location metrics, respectively. The quality loss for each measurement is normalized using the maximum observed quality loss metric for all groups. This normalization is done to ensure that different quality loss metrics can be compared to each other (da Silveira, Campeiro et al. 2017). The black lines above the bars correspond to the standard deviation within each group. This study uses a single part's measurement rather than an average of all nominal parts as the baseline profile. If the average of all nominal parts was used as the baseline, it could affect the measured variances within the nominal parts by reducing them. Hence, nominal parts will appear more similar in the observed quality loss metrics than they are in practice. The baseline profile used for the following quality loss calculations is the Group A, replication 1, measurement 1. Each group contains three replicates and three separate measurements were taken for each replicate.

It is observed from the bar graphs in Figure 54a, Figure 54b, Figure 54c that there is a clear separation between the baseline part - (parts A), and defective ones. This is observed in both the

height of the bar graphs for each group, and the standard deviation within the different groups – black line. For the RMSD metric, the nominal parts, Group A, have an average normalized quality loss of 6.81% with a standard deviation of 7.18%. In using the correlation coefficient metric, the nominal parts have an average normalized quality loss of 8.03% with a standard deviation of 7.57%. Using the peak location metric gives an average normalized quality loss of 0.71% with a standard deviation of 0.64%. This is a significantly better performance in terms of the distinction from the other groups and variation within the group. Also the average standard deviation within each group is 15.41%, 11.73%, and 5.81% for the RMSD, corr-coeff, and peak location metrics, respectively. Table 23 below shows quality loss metric values. This means that for this application, the peak location metric results in a smaller variance within the nominal parts in comparison to the other methods which is ideal. An added benefit is a lower variation in the quality loss metric within each group using the peak location method. These observations provide justification for using the peak location metric for quality loss calculations and will be used for the remainder of the study.

7.5 Results and discussions

The results from the multiple experiments conducted are discussed in this chapter. It includes the results from directly mounting the PZT to the part, magnetic mounting, and using the fixture and part combination. The advantages and disadvantages of each mounting methods are contrasted.

Table 23. Normalized results for all metrics investigated

	RMSD				***Correlation Coefficient***				***Peak Location***			
Part & Meas.	**A**	**B**	**C**	**D**	**A**	**B**	**C**	**D**	**A**	**B**	**C**	**D**
Rep 1 Meas. 1	0	90.57	54.29	40.29	0	87.95	61.45	74.7	0	39.29	33.82	100
Rep 1 Meas. 2	0.29	100	54.57	40	0	100	62.65	71.08	0	42.63	34.13	84.96
Rep 2 Meas. 1	4.86	25.43	60.86	21.43	7.23	71.08	73.49	57.83	0.71	43.75	60.96	78.48
Rep 2 Meas. 2	4.29	27.14	60	22.29	7.23	74.7	73.49	60.24	0.71	42.03	53.47	78.48
Rep 3 Meas. 1	16	38.29	62	63.43	16.87	93.98	95.18	90.36	1.52	49.11	46.84	83.49
Rep 3 Meas. 2	15.43	31.14	62.57	54.86	16.87	83.13	96.39	84.34	1.32	49.22	46.84	85.97

7.5.1 PZT Directly Mounted to Part

The twelve selected PZTs are mounted onto the four (4) groups which are differentiated by their part design. Each group has three replications and each part replication has two

measurements (or observations). Each part replicate undergoes two randomized measurements. The results show that groups' A are distinguishable from the other groups as shown in the box plot in Figure 54c and also Table 2.

7.5.2 PZT mounted to magnet

When the PZT is onto a magnet and the magnet is attached to the part, a higher variance between parts is observed. For example, in Figure 55a, replication A1, (plotted in red) has a significantly different impedance signal than replications A2 and A3. The same is observed for part replication C1 (plotted in red) in Figure 55c and part replication D3 (plotted in blue) in Figure 55d. Parts which showed significantly different signatures from their counterparts typically had a relatively smoother machined surface at the corner where the magnet makes contact. This suggests that the smoother machined surface affects the contact area between the magnet and the part, especially during excitation.

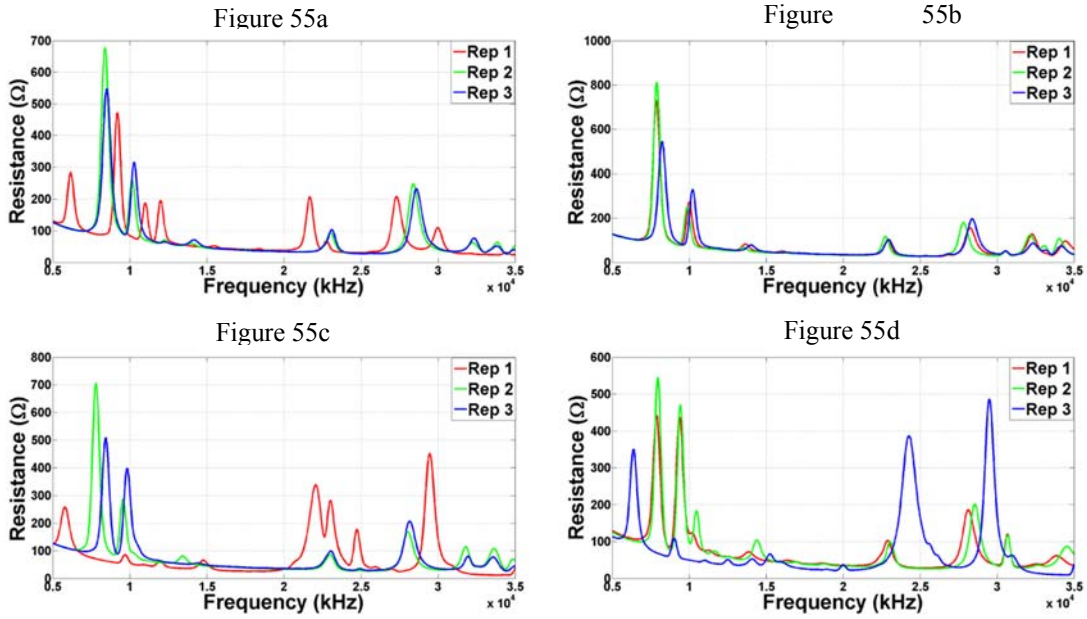

Figure 55. Signature profiles for parts when magnet is used. Figure 55a (top left) for Group A, rep 1, rep 2, rep 3. Figure 55b (top right) for Group B, rep 1, rep 2, rep 3. Figure 55c (bottom left) for Group C, rep 1, rep 2, rep 3. Figure 55d (bottom right) for Group D, rep 1, rep 2, rep 3.

The profiles in Figure 55a show that replicates A2 and A3 have peaks at locations that are different from replicate A1. The bar graph in Figure 56 as well as Table 240 reflects these observations. The standard deviation for group A is relatively large due to the differences within

the nominal parts. The peak location quality loss metric shows a significant difference within the group since A1 is used as the baseline in this measurement. The standard deviations for groups B, C, and D are lower than for group A because the peak location metric relies only on the peaks derived from the baseline. The results from this chapter show that the use of a magnetic mounting method, while promising, is dependent on the surface finish of the part. When a PZT is excited, it exerts a shear force in the axial direction (Lalande, Rogers et al. 1996, Wong, Du et al. 2015). This shear force is transferred to the magnet as the PZT is bonded directly to the magnet. In this study, the surface finish of similar part types were different in some cases which then affects the coupling between the magnet and part. This is reflected on the measured impedance signature.

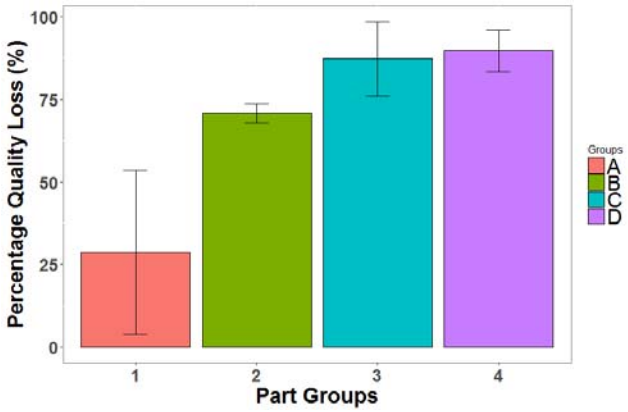

Figure 56. Bar graph of percentage quality loss for each part group when peak location metric is used. This is for the case of the magnetic mounting of PZTs.

7.5.3 PZT mounted to fixture and part combination

The comparison of impedance profiles for this mounting setup is shown in Figure 57a, Figure 57b, and Figure 57c. The profiles show a comparison between the nominal parts, group A, and the altered parts. This mounting method is also able to differentiate between the normal group, group A, and the other groups B, C, and D, which have a slot. The differences are less evident in comparison to the other methods. This is due to a smaller

Table 24. Quality loss metric for when magnets are mounted onto 12 parts. The results from each group is averaged. Part A Rep 1 is used as the baseline

	Groups			
Part & Meas.	**A**	**B**	**C**	**D**
Rep 1	0 (Baseline)	67.53	82.63	82.35
Rep 2	44.72	72.8	100	93.69
Rep 3	41.16	72.09	79.14	93.04

span of frequency ranges used in calculating the quality loss metrics. An obfuscating effect is observed from the fixture that limited the viable frequency ranges.

Figure 57a

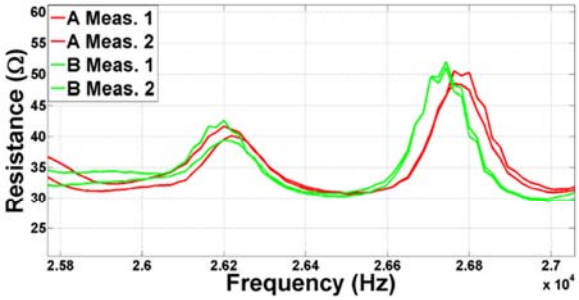

Figure 57b

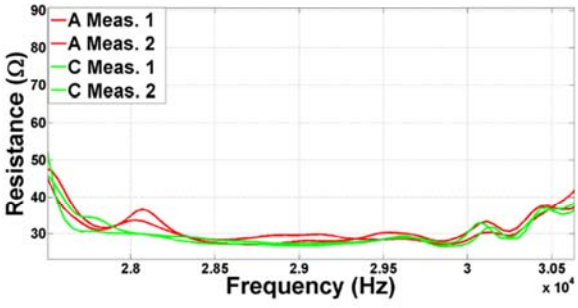

Figure 57c

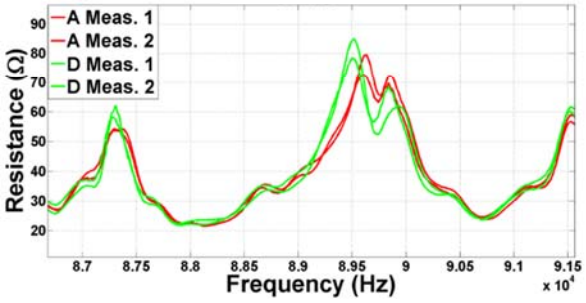

Figure 57. An example of an impedance peak of parts where A is the red line and altered parts are plotted in green. Figure 57a (top left) an example of an impedance peak of parts A and B. Figure 57b (top right) an example of an impedance peak of parts A and C. Figure 57c (bottom left) an example of an impedance peak of parts A and D.

Certain peaks are associated to local fixture modes and these will be minimally affected by the part under test. Whereas, some other peaks would reflect the dynamics of the coupled structure, where the effects of the part under test become more evident. Some examples of frequency ranges that were used are 12.5 kHz - 15.01 kHz, 18.40 kHz – 19.06 kHz, and 87.34 kHz – 87.40 kHz. Because of these obfuscating effects, better performance is achieved when treating all parts in a particular group as the same population and aggregating their profiles. For example, profiles of replicates A1, A2, and A3 are added together and group A measurement 1 used as the baseline. Two measurements are conducted for each group so for the nominal case, measurement two is compared to the baseline, measurement 1. For example, in Figure 57a, Figure 57b, and Figure 57c, the peaks of the different groups of parts have similar peak locations.

Table 25. Quality loss metric for when fixture is used. The results from each group is averaged. Rep 1 is used as the baseline.

	Groups			
Part & Meas.	**A**	**B**	**C**	**D**
Rep 1	0 (Baseline)	61.91	97.62	71.43
Rep 2	14.29	85.71	100	69.05

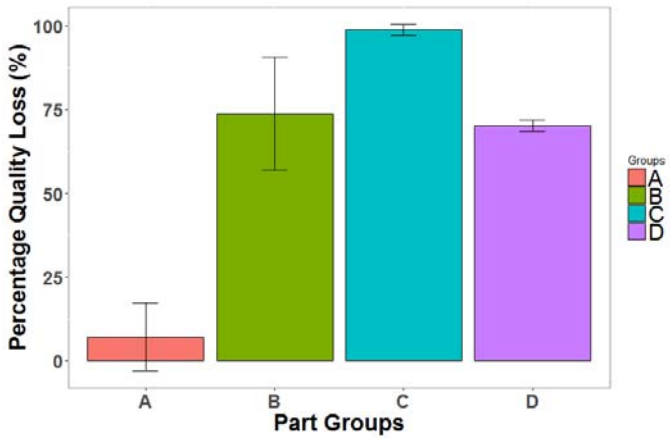

Figure 58. Percentage quality loss for each part group with the part affixed to a fixture

Although, there are some locations which show a divergence between groups of parts. In Figure 57a, group A (red) and B (green) have divergent peak locations at the frequency 26.8 KHz. In Figure 57a, group A (red) has a peak at the frequency location 28.2 KHz while parts in group C do not. Finally, it is observed in Figure 57b that groups A and D differ at the frequency location at 89.7 KHz. These are just

some of the examples of frequency ranges used for discriminating between groups of parts when the fixture mounting method is used.

The results are shown in the bar graph in Figure 58. Table 25 shows the values for the quality loss metric for each group.

7.6 Conclusion

This study shows that impedance based monitoring methodologies can be integrated into current manufacturing processes for NDE purposes. A new quality loss metric is introduced for this application. Multiple mounting methods are investigated. The method of interrogating a PZT directly mounted to a fixture is able to distinguish between a nominal part and those which have a geometric alteration. This mounting method was able to distinguish between the baseline part with no mass change and the other groups which have different levels of mass and stiffness change.

When the PZT is bonded to the magnet, some inconsistencies are observed between similar parts. Upon further investigation, these inconsistencies are likely due to the surface finish of the parts where the magnet made contact. Parts with a smoother surface finish had impedance signals significantly different from their counterparts. For this reason, if magnets are to be used in a manufacturing setting as an alternate mounting method, the manufacturing process must be reliable enough to maintain tight surface finish tolerances. As the goal of this study is to integrate impedance based NDE at all stages of the manufacturing process, including possibly roughing operations, the reliability of using magnets in all aspects of the manufacturing process diminishes.

The PZT-Fixture-Part combination provides us the consistency between interrogations of different parts. However, there are some masking effects from the fixture. That is, the signatures of the different parts appeared similar along multiple frequency locations. Therefore, one of the challenges in using this approach is identifying the frequency ranges where the signature is most sensitive to the dynamics of the part under test and not the fixture itself. Hence using only these frequency ranges for quality loss metric calculations. In this study, frequency ranges were selected via a manual scanning process.

Future research should consider automating the process of finding the frequency ranges of interest. One option would be to apply machine learning techniques to automate the frequency window identification. Another research path is to study the effect of fixture weight and other fixture designs on the impedance signature. In particular, the part weight to fixture weight ratio should be explored to understand its effect on the sensitivity of detecting quality loss. Fixture

designs that vary the contact area between the part and the fixture should also be investigated. In the context of fixture designs, other opportunities for research include looking into other material types for designing the fixture and comparing their performances.

Additive manufacturing is gaining prominence as a viable alternative for manufacturing parts. Performance of the PZT fixture NDE method needs to be verified for additively produced parts. This leads to a multitude of possibilities such as extending the types of alterations that can be investigated as well as fields that are affected.

8 SUMMARY OF FINDINGS AND FUTURE DIRECTION

The proposed work here in this field of big data analytics investigates different steps in the data analytics framework and addresses unique challenges at each step. Preprocessing data, specifically free text data, is difficult in the era of big data analytics as the quality of the data (veracity), quantity of data (volume), and forms of data (variety) require novel approaches to preprocessing the data. In the next step of the big data analytics framework, statistical learning, there are two objectives (1) statistical modeling and (2) anomaly detection. Two key contributions are introduced for each of these objectives.

8.1 Contributions in pre-processing of healthcare related text data

The main contribution of the work presented in this research area is creating a web based application that pre-processes text data using machine learning methodologies. Specifically, this web application makes it possible for frontline staff and external stakeholders to parse raw text data through this application and the result is disambiguated text documents. In this work, I introduce some specific pre-processing methodologies catered to healthcare related data such as handling of terms related to medication dosage, patient information, time event occurred, location of event, among others. This preprocessing application is critical as it allows different domain experts or users to interact with a common data source for future natural language processing analysis.

8.2 Contributions in statistical modeling of text data using community extraction algorithms

Aggregating text data is an open research area due to the variety of data sources, grammatical and syntax variations, among other unique challenges text data presents. The main

contribution of this work in this research area is developing a statistical modeling methodology to aggregate text documents into intuitive themes using a community extraction paradigm. This research area is called topic modeling. A limitation of many topic modeling applications is they are designed with the assumption that a majority of documents in a corpus belong to a particular topic. Furthermore, some of these methods require that the analyst know beforehand the number of topics contained in the corpus of documents. These are critical shortcomings that community extraction addresses and details of the methodology are addressed in Section 4 of this manuscript. I demonstrate the efficacy of the community extraction methodology in a real world application where 1773 healthcare related documents are aggregated into intuitive topics. Themes generated by frontline staff reporting these 1773 healthcare incidents are compared to themes generated autonomously using the community extraction methodology. The community extraction methodology is consistent in the partitioning of documents into communities across simulations. In addition, the community extraction methodology outperforms expert human taggers in the similarity between documents in a community. With this methodology, organizations collecting text data, i.e., product reviews, customer surveys, blogs, product descriptions, incident reports, can aggregate similar documents into topics. This way, general trends in key metrics that would frequently remain hidden in many organizations such as customer satisfaction, critical safety metrics, and seasonal concerns, among others are autonomously aggregated and highlighted.

8.3 Contributions in statistical modeling of overall health outcomes using newly defined social determinants of health

Another open research area, in statistical modeling, is using an individual's neighborhood demographic data to infer health outcomes for that particular individual. The main contributions in this work are summarized as follows: (1) as with previous studies, this work uses similar features for defining our social determinants of health (SDOH), however, we are getting the data from the federal government, large healthcare organizations, and other credible organizations which are both reliable and statistically sound (2) this work uses a sophisticated hexagonal binning to create local neighborhoods which are more intuitive as they have similarly shared infrastructure and SDOH limitations (3) on the medical side, this work looks at general risk factors and behavioral patterns which are not restricted to a set of diseases (4) and finally this work uses very different approaches in connecting SDOH factors with general health outcomes. I apply machine learning algorithms, Spearman correlation and other statistical tools to identify SDOH variables and the

general health outcomes with which they are paired. The final statistical models are therefore prospective, not just retrospective. This is very impactful as it allows practitioners to assign a health risk score to a geography or neighborhood. Some immediate benefits are: better allocation of resources in implementing intervention plans, improved identification of affected cohorts for clinical trials or pharmaceutical research, targeted government policies, and realistic stratification of healthcare costs across geographies, among many others.

8.4 Contributions in statistical evaluation of a suite of anomaly detection methodologies

Although numerous algorithmic techniques for anomaly detection have been introduced, however, to date, little work has been done to evaluate these algorithms from a statistical perspective. This work evaluates the statistical properties of the chi-square algorithm and L_1 norm algorithm. It is important to understand these properties when implementing the algorithms as a practitioner will want to know how the algorithms behave for different network sizes and network types. The main contributions of this work can be summarized in two main points: (1) evaluate the statistical properties of the chi-square algorithm and L_1 norm algorithm and identify critical shortcomings pertaining to their statistical properties as well as implementability and (2) introduce methodological improvements to both algorithms. Specifically providing more practical and appropriate signaling and detection schemes in both algorithms. These goals were achieved in this work. This work defines the attributes that must exist in order for a practitioner to apply the chi-square algorithm or L_1 norm algorithm. In addition, I identify the false alarm rates for both algorithms for different network types, detection rates, and a signaling threshold to indicate when an anomaly is present. Moreover, a suite of improvements are introduced that make the algorithms feasible to a practitioner.

8.5 Contributions in anomaly detection in a manufacturing process using side-channels

Large interconnected manufacturing enterprises also lend themselves to additional vulnerabilities. These could be in the form of malicious cyber-physical attacks that are designed to alter a part or process. However, other forms of vulnerabilities exist such as: unintended alterations that can exist at the intersection of advanced machineries and legacy systems, corrupt or outdated STL files, and alterations in a part due to a misconfiguration of an additive manufacturing (AM) printer, among many others. There are two main contributions from this work. First, a novel mounting method is investigated that uses an instrumented fixture for non-destructive evaluation (NDE) of manufactured parts. A piezoelectric transducer (PZT) is directly

mounted to a fixture and the fixture is interrogated with different parts affixed to it. This allows for multiple interrogations of different parts using the same measurement system as parts can be interchanged in the fixture. This work addresses immediate and possibly future vulnerabilities that are related to the advancements in the digitization of the manufacturing enterprise.

8.6 Future works and final thoughts

Although the contributions posed in this work are significant, with multiple accepted peer reviewed papers, there's still an avenue to advance some of these works. In the work related to word sense disambiguation, there is an opportunity to integrate additional information such as semantic analysis (including information on the tone of the document i.e., positive, neutral or negative), part of speech tagging (identifying nouns, pronouns, verbs, etc.), lemmatization (reducing words to their root) among others to this work. To model text data into communities, this work leveraged insights from the word sense disambiguation application in the pre-processing stage. The additional pre-processing steps listed above would also improve modeling of text data into communities. Furthermore, the work relies solely on latent semantic analysis (LSA) for dimension reduction. There are a range of dimension reduction techniques that can be employed. The extraction algorithm can also be modified to be more efficient, i.e. using a different metaheuristic algorithm for finding a near optimal solution.

I briefly investigated other machine learning algorithms to address challenges in both these works (1) relating social determinants of health to health outcomes and (2) developing an instrumented fixturing for anomaly detection in a manufacturing setting. Future works can look into using machine learning algorithms to identify the key variables driving the observed effects. For the instrumented fixturing work, reducing the weight of the fixture will significantly improve sensitivity of the methodology to anomalies.

Data analytics touches a lot fields and this work investigates unique challenges in two particular fields, manufacturing and healthcare. A recurring challenge in addressing the problems in these two fields is wrangling the raw data into a coherent form. I applied a range of techniques for data pre-processing, data wrangling, and feature extraction, introducing some novel methodologies along the way. There is an opportunity to leverage many of the methodologies explored in this work to other fields, and also subsets within these fields.

Finally, along the framework of big data analytics, there are open research areas in current real world applications. In this work, I successfully implemented statistical methodologies to

address many of these shortcomings and demonstrated their efficacy in solving these current real world problems.

9 REFERENCES

Adler, B. T., L. De Alfaro, S. M. Mola-Velasco, P. Rosso and A. G. West (2011). Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. International Conference on Intelligent Text Processing and Computational Linguistics, Springer.

Akoglu, L., M. McGlohon and C. Faloutsos (2010). "Oddball: Spotting anomalies in weighted graphs." Advances in Knowledge Discovery and Data Mining: 410-421.

Akoglu, L., H. Tong and D. Koutra (2015). "Graph based anomaly detection and description: a survey." Data Mining and Knowledge Discovery **29**(3): 626-688.

Albakri, M. I. and P. A. Tarazaga (2017). "Electromechanical impedance–based damage characterization using spectral element method." Journal of Intelligent Material Systems and Structures **28**(1): 63-77.

Albert, R., I. Albert and G. L. Nakarado (2004). "Structural vulnerability of the North American power grid." Physical review E **69**(2): 025103.

Albright, D., P. Brannan and C. Walrond (2010). Did Stuxnet take out 1,000 centrifuges at the Natanz enrichment plant?, Institute for Science and International Security.

Alley, D. E., C. N. Asomugha, P. H. Conway and D. M. Sanghavi (2016). "Accountable health communities—addressing social needs through Medicare and Medicaid." N Engl J Med **374**(1): 8-11.

America, F. (2011). "Map the meal gap." Feeding America.

Anderson, I., B. Robson, M. Connolly, F. Al-Yaman, E. Bjertness, A. King, M. Tynan, R. Madden, A. Bang and C. E. Coimbra Jr (2016). "Indigenous and tribal peoples' health (The Lancet–Lowitja Institute Global Collaboration): a population study." The Lancet **388**(10040): 131-157.

Aronson, A. R. and F.-M. Lang (2010). "An overview of MetaMap: historical perspective and recent advances." Journal of the American Medical Informatics Association **17**(3): 229-236.

Azarnoush, B., K. Paynabar, J. Bekki and G. Runger (2016). "Monitoring temporal homogeneity in attributed network streams." Journal of Quality Technology **48**(1): 28.

Bader, D. A. and K. Madduri (2008). Snap, small-world network analysis and partitioning: An open-source parallel graph framework for the exploration of large-scale networks. Parallel and Distributed Processing, 2008. IPDPS 2008. IEEE International Symposium on, IEEE.

Beasley, J. E. (1998). "Heuristic algorithms for the unconstrained binary quadratic programming problem." London, England.

Braveman, P., S. Egerter and D. R. Williams (2011). "The social determinants of health: coming of age." Annual review of public health **32**: 381-398.

Bray, D. E. and R. K. Stanley (1996). Nondestructive evaluation: a tool in design, manufacturing and service, CRC press.

Carden, E. P. and J. M. Brownjohn (2008). "ARMA modelled time-series classification for structural health monitoring of civil infrastructure." Mechanical systems and signal processing **22**(2): 295-314.

Carden, E. P. and P. Fanning (2004). "Vibration based condition monitoring: a review." Structural health monitoring **3**(4): 355-377.

Cer, R., K. Bruce, D. Donohue, N. Temiz, U. Mudunuri, M. Yi, N. Volfovsky, A. Bacolla, B. Luke and J. Collins (2012). "Searching for non-B DNA-forming motifs using nBMST (non-B DNA motif search tool)." Current protocols in human genetics: 18.17. 11-18.17. 22.

Cer, R. Z., K. H. Bruce, D. E. Donohue, A. N. Temiz, A. Bacolla, U. S. Mudunuri, M. Yi, N. Volfovsky, B. T. Luke and J. R. Collins (2011). "Introducing the non-B DNA Motif Search Tool (nBMST)." Genome biology **12**(1): P34.
Chakrabarti, D., C. Faloutsos and Y. Zhan (2007). "Visualization of large networks with min-cut plots, A-plots and R-MAT." International Journal of Human-Computer Studies **65**(5): 434-445.
Chakrabarti, D., Y. Zhan and C. Faloutsos (2004). R-MAT: A recursive model for graph mining. Proceedings of the 2004 SIAM International Conference on Data Mining, SIAM.
Chang, Y.-s. and Y.-H. Sung (2005). "Applying name entity recognition to informal text." Recall **1**: 1.
Chasin, R., A. Rumshisky, O. Uzuner and P. Szolovits (2014). "Word sense disambiguation in the clinical domain: a comparison of knowledge-rich and knowledge-poor unsupervised methods." Journal of the American Medical Informatics Association **21**(5): 842-849.
Chen, J. and Y. Saad (2012). "Dense subgraph extraction with application to community detection." IEEE Transactions on Knowledge and Data Engineering **24**(7): 1216-1230.
Cheng, X., X. Yan, Y. Lan and J. Guo (2014). "Btm: Topic modeling over short texts." IEEE Transactions on Knowledge & Data Engineering(1): 1-1.
Chung, F., L. Lu and V. Vu (2003). "Spectra of random graphs with given expected degrees." Proceedings of the National Academy of Sciences **100**(11): 6313-6318.
da Silveira, R. Z. M., L. M. Campeiro and F. G. Baptista (2017). "Performance of three transducer mounting methods in impedance-based structural health monitoring applications." Journal of Intelligent Material Systems and Structures: 2349-2362.
Dahan, M., L. Sela and S. Amin (2017). "Network Monitoring under Strategic Disruptions." arXiv preprint arXiv:1705.00349.
De Roeck, G., B. Peeters and J. Maeck (2000). "Dynamic monitoring of civil engineering structures." Computational Methods for Shell and Spatial Structures.
DeSmit, Z., A. E. Elhabashy, L. J. Wells and J. A. Camelio "An approach to cyber-physical vulnerability assessment for intelligent manufacturing systems." Journal of Manufacturing Systems.
DeSmit, Z. E., Ahmed
Wells, L.J
Camelio, Jamie (2016). Cyber-Physical Vulnerability Assessment in Manufacturing Systems. NAMRC. Blacksburg, VA, NAMRC.
Divsholi, B. S. and Y. Yang (2014). "Combined embedded and surface-bonded piezoelectric transducers for monitoring of concrete structures." NDT & E International **65**: 28-34.
Dr. Ralf C. Shlaepfer, M. K. (2014). "Industry 4.0 Challenges and solutions for the digital transformation and use of exponential technologies." Retrieved 05/14/2016, 2016, from http://www2.deloitte.com/content/dam/Deloitte/ch/Documents/manufacturing/ch-en-manufacturing-industry-4-0-24102014.pdf.
Dumais, S. T. (2004). "Latent semantic analysis." Annual review of information science and technology **38**(1): 188-230.
Erdos, P. and A. Rényi (1960). "On the evolution of random graphs." Publ. Math. Inst. Hung. Acad. Sci **5**(1): 17-60.
Ertöz, L., M. Steinbach and V. Kumar (2003). Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. Proceedings of the 2003 SIAM international conference on data mining, SIAM.

Falliere, N., L. O. Murchu and E. Chien (2011). "W32. stuxnet dossier." White paper, Symantec Corp., Security Response **5**(6).
Feinerer, I. and K. Hornik (2012). "tm: Text mining package." R package version 0.5-7.1 **1**(8).
Ferrer, R. L. (2018). Social Determinants of Health. Chronic Illness Care, Springer**:** 435-449.
Foundation, R. W. J., P. Braveman and S. Egerter (2008). Overcoming obstacles to health: report from the Robert Wood Johnson Foundation to the Commission to Build a Healthier America, Robert Wood Johnson Foundation.
Garla, V. N. and C. Brandt (2012). "Knowledge-based biomedical word sense disambiguation: an evaluation and application to clinical document classification." Journal of the American Medical Informatics Association **20**(5): 882-886.
Garla, V. N. and C. Brandt (2012). "Ontology-guided feature engineering for clinical text classification." Journal of biomedical informatics **45**(5): 992-998.
Gillam, S. (2008). "Is the declaration of Alma Ata still relevant to primary health care?" BMJ: British Medical Journal **336**(7643): 536.
Glover, F. and M. Laguna (2013). Tabu Search∗. Handbook of combinatorial optimization, Springer**:** 3261-3362.
Gregory, C. A. and A. Coleman-Jensen (2017). Food insecurity, chronic disease, and health among working-age adults, United States Department of Agriculture, Economic Research Service.
Handcock, M. S., A. E. Raftery and J. M. Tantrum (2007). "Model-based clustering for social networks." Journal of the Royal Statistical Society: Series A (Statistics in Society) **170**(2): 301-354.
Haveliwala, T. H. (2003). "Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search." IEEE transactions on knowledge and data engineering **15**(4): 784-796.
Hoisungwan, J. K. T. B. T. K. J. S. P. (2015). Parameter estimation of Gumbel distribution for flood peak data. C. University. Department of Electrical Engineering, Chulalongkorn University.
Honeycutt, J. Tips for Inexperienced Managers-How to Supervise Experienced Staff, University of South Carolina.
Hunter, D. R., S. M. Goodreau and M. S. Handcock (2008). "Goodness of fit of social network models." Journal of the American Statistical Association **103**(481): 248-258.
Jackson, J., A. Qiao and E. P. Xing "Scaling HDBSCAN Clustering with kNN Graph Approximation."
Jin, Y. and Y. Makris (2008). Hardware Trojan detection using path delay fingerprint. Hardware-Oriented Security and Trust, 2008. HOST 2008. IEEE International Workshop on, IEEE.
Kavanagh, K. T., D. M. Saman, R. Bartel and K. Westerman (2017). "Estimating Hospital-Related Deaths Due to Medical Error: A Perspective From Patient Advocates." Journal of patient safety **13**(1): 1-5.
Kessler, S. S., S. M. Spearing, M. J. Atalla, C. E. Cesnik and C. Soutis (2002). "Damage detection in composite materials using frequency response methods." Composites Part B: Engineering **33**(1): 87-95.
Khajeh, A. and A. Y. Koma (2007). Structural health monitoring using peak of frequency response. Proceedings of the 9th WSEAS Int. Conf. on Automatic Control, Modeling & Simulation, Istanbul, Turkey.
Kim, J.-D., T. Ohta, Y. Tateisi and J. i. Tsujii (2003). "GENIA corpus—a semantically annotated corpus for bio-textmining." Bioinformatics **19**(suppl_1): i180-i182.
Klinenberg, E. (2016). "Social isolation, loneliness, and living alone: Identifying the risks for public health." American journal of public health **106**(5): 786.

Koh, H. K., J. J. Piotrowski, S. Kumanyika and J. E. Fielding (2011). "Healthy people: a 2020 vision for the social determinants approach." Health Education & Behavior **38**(6): 551-557.
Köhler, B., T. Gaul, U. Lieske and F. Schubert (2016). "Shear horizontal piezoelectric fiber patch transducers (SH-PFP) for guided elastic wave applications." NDT & E International **82**: 1-12.
Kuruvilla, S., J. Schweitzer, D. Bishai, S. Chowdhury, D. Caramani, L. Frost, R. Cortez, B. Daelmans, A. d. Francisco and T. Adam (2014). "Success factors for reducing maternal and child mortality." Bulletin of the World Health Organization **92**: 533-544.
Kushel, M. B., R. Gupta, L. Gee and J. S. Haas (2006). "Housing instability and food insecurity as barriers to health care among low-income Americans." Journal of general internal medicine **21**(1): 71-77.
Lalande, F., C. A. Rogers, B. W. Childs and Z. A. Chaudhry (1996). High-frequency impedance analysis for NDE of complex precision parts. 1996 Symposium on Smart Structures and Materials, International Society for Optics and Photonics.
Landauer, T. K., P. W. Foltz and D. Laham (1998). "An introduction to latent semantic analysis." Discourse processes **25**(2-3): 259-284.
Lee, R. M., M. J. Assante and T. Conway (2014). "German steel mill cyber attack." Industrial Control Systems **30**.
Leskovec, J., D. Chakrabarti, J. Kleinberg and C. Faloutsos (2005). Realistic, mathematically tractable graph generation and evolution, using kronecker multiplication. PKDD, Springer.
Liang, C., F. Sun and C. Rogers (1997). "Coupled electro-mechanical analysis of adaptive material systems-determination of the actuator power consumption and system energy transfer." Journal of intelligent material systems and structures **8**(4): 335-343.
Locatelli, S. M., L. K. Sharp, S. T. Syed, S. Bhansari and B. S. Gerber (2017). "Measuring health-related transportation barriers in urban settings." Journal of applied measurement **18**(2): 178.
Makary, M. A. and M. Daniel (2016). "Medical error-the third leading cause of death in the US." BMJ: British Medical Journal (Online) **353**.
Marmot, M. (2005). "Social determinants of health inequalities." The lancet **365**(9464): 1099-1104.
Marmot, M., J. Allen, R. Bell, E. Bloomer and P. Goldblatt (2012). "WHO European review of social determinants of health and the health divide." The lancet **380**(9846): 1011-1029.
Marmot, M. and J. J. Allen (2014). Social determinants of health equity, American Public Health Association.
Marmot, M., S. Friel, R. Bell, T. Houweling and S. Taylor (2009). "Closing the gap in a generation: health equity through action on the social determinants of health." Child Care, Health And Development **35**(2): 285.
Mazrae Farahani, E., R. Baradaran Kazemzadeh, R. Noorossana and G. Rahimian (2016). "A Statistical Approach to Social Network Monitoring." Communications in Statistics-Theory and Methods(just-accepted).
McNamara, C. L., M. Balaj, K. H. Thomson, T. A. Eikemo and C. Bambra (2017). "The contribution of housing and neighbourhood conditions to educational inequalities in non-communicable diseases in Europe: findings from the European Social Survey (2014) special module on the social determinants of health." The European Journal of Public Health **27**(suppl_1): 102-106.
Mehrotra, R., S. Sanner, W. Buntine and L. Xie (2013). Improving lda topic models for microblogs via tweet pooling and automatic labeling. Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, ACM.

Meltzer, R. and A. Schwartz (2016). "Housing affordability and health: evidence from New York City." Housing Policy Debate **26**(1): 80-104.
Meyers, A., D. Cutts, D. A. Frank, S. Levenson, A. Skalicky, T. Heeren, J. Cook, C. Berkowitz, M. Black and P. Casey (2005). "Subsidized housing and children's nutritional status: data from a multisite surveillance study." Archives of pediatrics & adolescent medicine **159**(6): 551-556.
Miller, B., N. Bliss and P. J. Wolfe (2010). Subgraph detection using eigenvector L1 norms. Advances in Neural Information Processing Systems.
Miller, B. A., M. S. Beard, P. J. Wolfe and N. T. Bliss (2015). "A spectral framework for anomalous subgraph detection." IEEE transactions on signal processing **63**(16): 4191-4206.
Moon, S., B. McInnes and G. B. Melton (2015). "Challenges and practical approaches with word sense disambiguation of acronyms and abbreviations in the clinical domain." Healthcare informatics research **21**(1): 35-42.
Na, S., R. Tawie and H.-K. Lee (2012). "Electromechanical impedance method of fiber-reinforced plastic adhesive joints in corrosive environment using a reusable piezoelectric device." Journal of Intelligent Material Systems and Structures **23**(7): 737-747.
Na, W. S. (2017). "Possibility of detecting wall thickness loss using a PZT based structural health monitoring method for metal based pipeline facilities." NDT & E International **88**: 42-50.
Nadarajah, S. and S. Kotz (2004). "The beta Gumbel distribution." Mathematical Problems in Engineering **2004**(4): 323-332.
Naessens, E. (2017). "Medical Error: a Patient's Perspective." Clinical Oncology **29**(10): 667-668.
Newman, M. (2016). "Community detection in networks: Modularity optimization and maximum likelihood are equivalent." arXiv preprint arXiv:1606.02319.
Newman, M. E. (2006). "Finding community structure in networks using the eigenvectors of matrices." Physical review E **74**(3): 036104.
Newman, M. E. and M. Girvan (2004). "Finding and evaluating community structure in networks." Physical review E **69**(2): 026113.
Nowicki, K. and T. A. B. Snijders (2001). "Estimation and prediction for stochastic blockstructures." Journal of the American statistical association **96**(455): 1077-1087.
Oberlander, J. and M. J. Laugesen (2015). "Leap of faith—medicare's new physician payment system." New England Journal of Medicine **373**(13): 1185-1187.
Organization, W. H. (2008). "Final report of the Commission on Social Determinants of Health." Closing the gap in a generation. Health equity through action on the social determinants of health. Geneva: WHO.
Organization, W. H. (2008). "Our cities, our health, our future: acting on social determinants for health equity in urban settings. Report to the WHO Commission on Social Determinants of Health from the Knowledge Network on Urban Settings." Our cities, our health, our future: acting on social determinants for health equity in urban settings. Report to the WHO Commission on Social Determinants of Health from the Knowledge Network on Urban Settings.
Pakhomov, S., T. Pedersen and C. G. Chute (2005). Abbreviation and acronym disambiguation in clinical discourse. AMIA Annual Symposium Proceedings, American Medical Informatics Association.
Papadimitriou, S., H. Kitagawa, P. B. Gibbons and C. Faloutsos (2003). Loci: Fast outlier detection using the local correlation integral. Data Engineering, 2003. Proceedings. 19th International Conference on, IEEE.
Park, G., H. H. Cudney and D. J. Inman (2000). "Impedance-based health monitoring of civil structural components." Journal of infrastructure systems **6**(4): 153-160.

Park, G., H. H. Cudney and D. J. Inman (2001). "Feasibility of using impedance-based damage assessment for pipeline structures." Earthquake engineering & structural dynamics **30**(10): 1463-1474.
Park, G., H. Sohn, C. R. Farrar and D. J. Inman (2003). "Overview of piezoelectric impedance-based health monitoring and path forward."
Paul, S. M. P. and C. Jayaguru (2016). "Health monitoring of Concrete specimens using smart aggregates." i-Manager's Journal on Structural Engineering **5**(3): 25.
Pedersen, T. (2000). A simple approach to building ensembles of Naive Bayesian classifiers for word sense disambiguation. Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference, Association for Computational Linguistics.
Pickett, K. E. and M. Pearl (2001). "Multilevel analyses of neighbourhood socioeconomic context and health outcomes: a critical review." Journal of Epidemiology & Community Health **55**(2): 111-122.
Pink, B. and P. Allbon (2008). The health and welfare of Australia's Aboriginal and Torres Strait Islander peoples, Commonwealth of Australia Canberra.
Priebe, C. E., J. M. Conroy, D. J. Marchette and Y. Park (2005). "Scan statistics on enron graphs." Computational & Mathematical Organization Theory **11**(3): 229-247.
Procter, J. B., J. Thompson, I. Letunic, C. Creevey, F. Jossinet and G. J. Barton (2010). "Visualization of multiple alignments, phylogenies and gene family evolution." Nature methods **7**: S16-S25.
Raju, V., G. Park and H. H. Cudney (1999). Impedance-based health monitoring of composite reinforced structures. Ninth International Conference on Adaptive Structures and Technologies.
Ralston, P. A., J. H. Graham and J. L. Hieb (2007). "Cyber security risk assessment for SCADA and DCS networks." ISA transactions **46**(4): 583-594.
Ranshous, S., S. Shen, D. Koutra, S. Harenberg, C. Faloutsos and N. F. Samatova (2015). "Anomaly detection in dynamic networks: a survey." Wiley Interdisciplinary Reviews: Computational Statistics **7**(3): 223-247.
Raulf-Heimsoth, M., Z. Chen, H. Rihs, H. Kalbacher, V. Liebers and X. Baur (1998). "Analysis of T-cell reactive regions and HLA-DR4 binding motifs on the latex allergen Hev b 1 (rubber elongation factor)." Clinical and Experimental Allergy **28**(3): 339-348.
Šaltenis, V. (2004). "Outlier detection based on the distribution of distances between data points." Informatica **15**(3): 399-410.
Sampson, L., S. S. Martins, S. Yu, A. D. P. Chiavegatto Filho, L. H. Andrade, M. C. Viana, M. E. Medina-Mora, C. Benjet, Y. Torres and M. Piazza (2018). "The relationship between neighborhood-level socioeconomic characteristics and individual mental disorders in five cities in Latin America: multilevel models from the World Mental Health Surveys." Social psychiatry and psychiatric epidemiology: 1-14.
Savage, D., X. Zhang, X. Yu, P. Chou and Q. Wang (2014). "Anomaly detection in online social networks." Social Networks **39**: 62-70.
Savova, G. K., J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler and C. G. Chute (2010). "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications." Journal of the American Medical Informatics Association **17**(5): 507-513.
Schlossberg, M. (2003). "GIS, the US census and neighbourhood scale analysis." Planning, Practice & Research **18**(2-3): 213.

Singh, N., B. A. Miller, N. T. Bliss and P. J. Wolfe (2011). Anomalous subgraph detection via sparse principal component analysis. Statistical Signal Processing Workshop (SSP), 2011 IEEE, IEEE.
Smith, J., K. Griffiths, J. Judd, G. Crawford, H. D'Antoine, M. Fisher, R. Bainbridge and P. Harris (2018). "Ten years on from the world health organization commission of social determinants of health: Progress or procrastination?" Health Promotion Journal of Australia **29**(1): 3-7.
Steinbach, M., G. Karypis and V. Kumar (2000). A comparison of document clustering techniques. KDD workshop on text mining, Boston.
Sun, F. P., Z. Chaudhry, C. Liang and C. Rogers (1995). "Truss structure integrity identification using PZT sensor-actuator." Journal of Intelligent material systems and structures **6**(1): 134-139.
Sun, F. P., Z. A. Chaudhry, C. A. Rogers, M. Majmundar and C. Liang (1995). Automated real-time structure health monitoring via signature pattern recognition. Smart Structures & Materials' 95, International Society for Optics and Photonics.
Sun, J., H. Qu, D. Chakrabarti and C. Faloutsos (2005). Neighborhood formation and anomaly detection in bipartite graphs. Data Mining, Fifth IEEE International Conference on, IEEE.
Syed, S. T., B. S. Gerber and L. K. Sharp (2013). "Traveling towards disease: transportation barriers to health care access." Journal of community health **38**(5): 976-993.
Team, R. C. (2014). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2014.
Tufféry, S. (2011). Data mining and statistics for decision making, Wiley Chichester.
Turney, P. D. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. European Conference on Machine Learning, Springer.
Venes, D. (2017). Taber's cyclopedic medical dictionary, FA Davis.
Vincent, H., L. Wells, P. Tarazaga and J. Camelio (2015). Trojan Detection and Side-Channel Analyses for Cyber-Security in Cyber-Physical Manufacturing Systems. 43rd North American Manufacturing Research Conference (NAMRC), Charlotte, NC, Elsevier B.V.
Vincent, H., L. Wells, P. Tarazaga and J. Camelio (2015). "Trojan Detection and Side-channel Analyses for Cyber-security in Cyber-physical Manufacturing Systems." Procedia Manufacturing **1**: 77-85.
Wang, G., S. Xie, B. Liu and P. S. Yu (2012). "Identify online store review spammers via social review graph." ACM Transactions on Intelligent Systems and Technology (TIST) **3**(4): 61.
Wang, X., M. Tehranipoor and J. Plusquellic (2008). Detecting malicious inclusions in secure hardware: Challenges and solutions. Hardware-Oriented Security and Trust, 2008. HOST 2008. IEEE International Workshop on, IEEE.
Wells, L. J. and J. A. Camelio (2013). "A bio-inspired approach for self-correcting compliant assembly systems." Journal of Manufacturing Systems **32**(3): 464-472.
Wells, L. J., J. A. Camelio, C. B. Williams and J. White (2014). "Cyber-physical security challenges in manufacturing systems." Manufacturing Letters **2**(2): 74-77.
Wilkinson, R. G. and M. Marmot (2003). Social determinants of health: the solid facts, World Health Organization.
Willett, W. C. (2002). "Balancing life-style and genomics research for disease prevention." Science **296**(5568): 695-698.
Williams, D. R., M. V. Costa, A. O. Odunlami and S. A. Mohammed (2008). "Moving upstream: how interventions that address the social determinants of health can improve health and reduce disparities." Journal of public health management and practice: JPHMP **14**(Suppl): S8.

Winston, H. A., F. Sun and B. S. Annigeri (2000). Structural health monitoring with piezoelectric active sensors. ASME Turbo Expo 2000: Power for Land, Sea, and Air, American Society of Mechanical Engineers.
Wolpert, R. L. (2014). "Extremes." Department of Statistical Science,(4).
Wong, Y.-R., H. Du and X. Pang (2015). "Real-time electrical impedance resonance shift of piezoelectric sensor for detection of damage in honeycomb core sandwich structures." NDT & E International **76**: 61-65.
Woodall, W. H., M. J. Zhao, K. Paynabar, R. Sparks and J. D. Wilson (2017). "An overview and perspective on social network monitoring." IISE Transactions **49**(3): 354-365.
Xu, P. and F. Jelinek (2004). Random Forests in Language Modeling. EMNLP.
Zachary, W. W. (1977). "An information flow model for conflict and fission in small groups." Journal of anthropological research **33**(4): 452-473.
Zhao, Y., E. Levina and J. Zhu (2011). "Community extraction for social networks." Proceedings of the National Academy of Sciences **108**(18): 7321-7326.