CEC Theses and Dissertations

College of Engineering and Computing

2019

Adaptable Privacy-preserving Model

Emily Elizabeth Brown

Nova Southeastern University, ebrown91@gmail.com

This document is a product of extensive research conducted at the Nova Southeastern University College of Engineering and Computing. For more information on research and degree programs at the NSU College of Engineering and Computing, please click here.

Follow this and additional works at: https://nsuworks.nova.edu/gscis_etd

Part of the Computer Sciences Commons

Share Feedback About This Item

Adaptable Privacy-preserving Model

by

Emily E Brown

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in
Information Assurance

College of Engineering and Computing
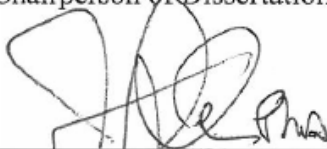Nova Southeastern University

2019

We hereby certify that this dissertation, submitted by Emily Brown, conforms to acceptable standards and is fully adequate in scope and quality to fulfill the dissertation requirements for the degree of Doctor of Philosophy.

_____
James D. Cannady, Ph.D.
Chairperson of Dissertation Committee

01/16/2019
Date

_____
James L. Parrish, Ph.D.
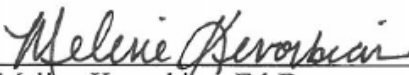Dissertation Committee Member

01/16/2019
Date

_____
Steven R. Terrell, Ph.D.
Dissertation Committee Member

1/16/2019
Date

Approved:

_____
Meline Kevorkian, Ed.D.
Interim Dean, College of Engineering and Computing

1/16/2019
Date

College of Engineering and Computing
Nova Southeastern University

2019

An Abstract of a Dissertation Submitted to Nova Southeastern University
in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

Adaptable Privacy-Preserving Model

by
Emily Brown
2019

Current data privacy-preservation models lack the ability to aid data decision makers in processing datasets for publication. The proposed algorithm allows data processors to simply provide a dataset and state their criteria to recommend an xk-anonymity approach. Additionally, the algorithm can be tailored to a preference and gives the precision range and maximum data loss associated with the recommended approach. This dissertation report outlined the research's goal, what barriers were overcome, and the limitations of the work's scope. It highlighted the results from each experiment conducted and how it influenced the creation of the end adaptable algorithm. The xk-anonymity model built upon two foundational privacy models, the k-anonymity and l-diversity models. Overall, this study had many takeaways on data and its power in a dataset.

Acknowledgements

My profound gratitude goes out to my family and Dr. Cannady. Many sacrifices occurred to allow me time to focus on my studies. There were many late nights and random rambling my family could only nod and smile. These side discussions, constant support, and continuous demand for me never to give up, pushed me through the unknowns and blockades.

Dr. Cannady was my first professor in the Information Assurance doctoral program. His demand in excellence and solid research model led me to publish my first research project, completed in his course. From there my passion for data knowledge and processing grew, and Dr. Cannady supported me in continuing this direction of study. He constantly delivered insightful feedback.

I would also like to thank Dr. Parrish and Dr. Terrell for being on my dissertation committee, Dr. Smith in widening my outlook on the research's potential, and Dr. Rosenbaum for helping broaden my understanding of information privacy. Finally, I would like to thank my co-workers and bosses that supported me and listened to my research progress.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Background/Introduction

Privacy-preservation approaches hinder the quantity of data available for publication (Brown, 2017). In order for a successfully designed privacy-preservation method, the model must incorporate many key factors. First, a model needed to understand the types of attributes in a dataset as it is critical when looking to preserve information privacy. Second, a model needed to understand the necessary anonymization techniques to preserve privacy. Third, a model needed to harmonize the anonymization techniques used on the attributes for a successful de-identified dataset. By building upon foundational privacy models' approaches (Yang, Li, Zhang, & Yu, 2013), a strong platform was created for future improved methods.

There are four commonly accepted categories of attributes: explicit, quasi-identifier, sensitive, and non-sensitive (Nagendrakumar, Aparna, & Ramesh, 2014). An attribute is explicit if it directly identifies an individual, such as social security number. Quasi-identifiers are frequently published attributes that could detect a person's sensitive information if the data-mined together with other published datasets. Common quasi-identifiers are zip codes and year of birth (YOB). Classified as sensitive attributes, these values include details that are particular to a person but also share the same value with other people. For example, health conditions and salaries are sensitive attributes. When attributes do not apply to the other three categories, they are non-sensitive. By properly classifying the attributes, data processors can successfully transform their datasets with a privacy-preserving model.

Throughout all research on privacy-preservation models, alterations were required. Anonymization techniques apply modifications to a dataset for publication eligibility. Two common anonymization techniques, generalization and suppression, are published in several relevant models (Sweeney, 2002b; Machanavajjhala, Kifer, Gehrke, & Venkatasubramanian, 2007; Liu, Luo, & Huang, 2011; Angiuli & Waldo, 2016; Brown, 2017). Other techniques used for anonymization include bucketization, data falsification, and synthesizing datasets (Li, Li, Zhang, & Molloy, 2012; Brown, 2017; Dwork, 2009). Modifications of datasets hinder information precision and can change attributes' relationships to one another (Angiuli & Waldo, 2016). A strong privacy model needed to balance anonymization techniques to maintain a high precision and attribute correlation.

There are four commonly accepted privacy-preservation methods. Sweeney (2002a) created the *k*-anonymity privacy protection model to de-identify an individual from its record. The *k*-anonymity theory stated a record is anonymous if there are at least *k*-1 records matching the same criteria. The *l*-diversity model expanded on *k*-anonymity, requiring there be at least *l*-diverse group of sensitive values for records to be unidentifiable (Machanavajjhala et al., 2007). The *t*-closeness model addressed weaknesses from *k*-anonymity and *l*-diversity where the total population had to be in a specific range for a record to be published (Li, Li, & Venkatasubramanian, 2007). Differential privacy model took a statistical approach to anonymize a dataset (Dwork, McSherry, Nissim, & Smith, 2006). These four models provided future researchers a foundation upon which to build a secure privacy model that meets at least these minimum standards.

To create a quality privacy-preservation model, a model needed to account for a series of critical elements. It was crucial to classify attributes correctly in the dataset for accurate processing. Data decision makers should understand how the application of anonymization techniques affect the dataset for publication. Organizations have to set guidelines based on privacy-preservation models for effective anonymization of their datasets. Attribute classification, anonymization techniques, and de-identification models are all key elements to maintain successful modification of dataset's information for privacy-preservation.

Problem Statement

There were no adaptable privacy-preservation methods available to handle diverse datasets. Past failures with organizations protecting sensitive data, like Netflix's anonymized dataset failure (*InfoLawGroup*, 2013), demonstrated the need for an adaptable tool to apply k-anonymity and l-diversity requirements to de-identify a dataset. In order to modify personal information appropriately, for publication, quasi-identifiers and sensitive attributes must be reviewed (Yang et al., 2013). As the quantity of quasi-identifiers increased, the loss in data exponentially increased, but Yang et al. (2013) did not propose a way to handle the quasi-identifier variety to reduce data loss. Sensitive attributes require l-diversity, so quasi-identifier pairs do not have a single sensitive value that would result in directly identifying personal information (Machanavajjhala et al., 2007). The l-diversity model tightened the requirements for records to pass for publication, but the model negatively affected the quantity of publishable data. The *1/2k* theory aided in publishing higher quantity of data (Brown, 2017); however, research limitations only measured a single fraction and did not measure precision rate.

Organizations see the accumulating need for anonymization techniques when working with datasets containing personal information (Garfinkel, 2015). No previous research adjusted the *k* fraction for quasi-identifiers and reviewed sensitive attribute suppression. This research study proposed an adaptable solution to privacy-preservation.

Dissertation Goal

The object of any privacy model is to design a method that is effective without decreasing security (Aggarwal & Yu, 2008). The goal of this dissertation research was to create an algorithm flexible to handle different quantities of quasi-identifiers. Additionally, it analyzed a single sensitive value to understand the impact a diversity requirement has on the overall publishable data rate. The model reviewed a dataset's composition to determine the best combination of anonymization techniques to maintain precision and attribute correlations. The study investigated four privacy-preservation aspects of data processing: the impact on the number of different quasi-identifiers in a dataset, the set condition of *l* for *l*-diversity on a sensitive attribute, the difference in the quasi-identifier pair values have on the overall dataset, and the rate of precision in a dataset at different fractions of *k*. This work accomplished finding the appropriate *k* fraction to process a dataset to meet *k*-anonymity and *l*-diversity with a certain degree of precision. To measure success in this research, the algorithm can take a dataset of diverse sizes and complete the lowest required modifications on the records to publish a dataset that has a set precision range. The user is only required to provide three details: the dataset, the quasi-identifier (which attributes), and the sensitive value (which single attribute and the diversity value). For example, a user having a 500,000 record dataset, with five quasi-identifiers, and a single sensitive attribute with a diversity of three can

have the algorithm calculate those user details to determine the best approach to anonymize the dataset. The model predicts the precision range and the best corresponding fraction of k. The end algorithm provided all elements of a strong privacy model to ensure it protects sensitive information, maintains similar features to the original dataset, and preserves the correlation between attributes for data-mining (Gkoulalas-Divanis & Verykios, 2009). The research proposed a better solution than previous models in versatility for processing diverse dataset sizes and analyzing the significance of a quasi-identifier pair.

Research Questions

  The goal of this research required the production of an algorithm to handle a combination of different quasi-identifiers, a single sensitive value at diverse levels, and a variety of dataset sizes. Research questions focused on the different aspects of the study, which aided in the creation of the adaptable algorithm. Each research question addressed at least one of the studied aspects for privacy-preservation.

- Can a single quasi-identifier affect the precision rate of the overall publishable dataset by more than 2%?

- What impact does different size datasets have on precision rate?

- Does processing a dataset with a lower fraction level of k improve the precision rate?

- What affects can different quantities of quasi-identifiers have on a dataset of the same size?

- How does l-diversity value affect the overall record publication quantity in a single dataset?

Relevance and Significance

There were two primary reasons this research held relevance and significance. The first significant component was organizations' legal obligations to protect a person's identity (Angiuli & Waldo, 2016). The second component was the need for data in the research environment to develop and improve society (Armer, 1981; Leonard, 2016; Polonetsky & Tene, 2013; van der Sloot & van Schendel, 2016). Data processors legal obligations and researchers' information needs provided two significant and relevant reasons to have this research conducted.

In 1974, the government realized the potential misuse of technology, which exposed privacy issues to society. In response, the government created the Privacy Act of 1974 to protect information privacy (Smith, Dinev, & Xu, 2011). To complicate matters, laws between countries hold privacy at different regulations (Schwartz, 2013). Even within the United States (U.S.), different states have different standards on the requirements for organizations to maintain personal data on a state's resident. For example, California requires a level of security protocols for organizations to implement if they possess Californian residents' information. These legal discrepancies and requirements give a prime reason on why it is relevant and significant to create an algorithm that is adaptable to organizations.

Other researchers and society benefit from a larger, more accurate, anonymized dataset. Currently, alternative models, shown in literature review, attempt to modify datasets for publications but lose valuable content that would remain in an adaptable model. This conducted research aimed to improve on society's ability to learn more information by publishing higher quantities of quality data. It was relevant and significant

for this research to be done as data-mining is a key part in discovering new knowledge (Brown, 2017).

Barriers and Issues

The research overcame three main barriers and issues. The first barrier was the contrasting goals between privacy and utility (Sedayao, Bhardwaj, & Gorade, 2014). When increasing privacy in a dataset, the utility of the data decreases. Conflicts in privacy and utility is an ongoing cause of problems between data publication value and an individual's right to privacy (Armer, 1981). In *Privacy: A Survey,* Armer (1981) used the struggle of a person's privacy versus society's right to learn. The study showed the need for more information to learn how to advance in society, but it came at a cost to an individual's privacy. This maintained a barrier with the growing usage of Big Data (Tene & Polonetsky, 2013). Big Data is powering innovation, but it comes at a cost to individuals' privacy. Data are more necessary than ever before in society (Bertino, 2016); the balance of utility and privacy is difficult to maintain. To solve this barrier, the model educates users on how much precision and data loss the algorithm anticipates when processing a dataset at a certain criterion.

Jurisdictional obligations set a minimum standard for organizations to process data (Sedayao et al., 2014). The European Union (E.U.) implemented comprehensive regulations to set a standard on all data; whereas, the U.S. adopted a sectoral approach. This approach requires handling certain attributes differently depending on the industry dispensing the information (Munir, Yasin, & Muhammad-Sukki, 2015). In the U.S. there are the Health Insurance Portability and Accountability Act (HIPAA) and the Family Educational Rights and Privacy Act (FERPA) which require organizations to remove

identifiable information before making a dataset publicly available (Angiuli & Waldo, 2016). How HIPAA processes attributes are different from how FERPA requires the data to be processed. To remove this obstacle, the proposed algorithm recommended an approach that met all requirements of k-anonymity and l-diversity to ensure de-identification of a record. The algorithm also is flexible in handling diverse requirements so HIPAA and FERPA principles are eligible for processing.

The third issue this research addressed is the adaptability dilemma of current privacy-preservation models. Yang et al. (2013) researched when implementing different quasi-identifiers, the data loss increased. To conquer this issue, the study encompassed many datasets with different amounts of quasi-identifiers. This research overcame the adaptability barrier by creating a model that handles a variety of attribute types, quasi-identifier pairs, and the diversity requirements of a sensitive value.

Assumptions and Limitations

As in many research projects, there were assumptions and limitations that were taken into account. In this study, it assumed the data processed for anonymization was in a structured format and the data processors have knowledge of what classification applies to each attribute for proper processing. This research limited the review of quasi-identifiers to be at a maximum of five. The k value applied to the study remained constant at five. This value is higher than the U.S. Department of Education's requirements and based off previous research that used the numerical value of five (Angiuli & Waldo, 2016; Brown, 2017). The study limited a dataset to have only up to one sensitive attribute, processed with three different potential diversity values (2, 3, or 4). Data processors are required to rank their quasi-identifiers in order of importance and

understand the change in utility if including cell-based suppression or sensitive suppression. These assumptions and limitations allowed the research to focus on a set of parameters.

Research Takeaways

In conducting this research, there are several takeaways learned about datasets. This study explored the relationship between quasi-identifier pairs and the overall total dataset. It found the precision change when using different fractions of k to create false records. This research addressed how to compensate for Yang's et al. (2013) research findings, which showed the exponential data loss when there is an increase in quasi-identifiers. It discovered how to best approach the inclusion of a sensitive value. Overall, the study learned more about the core features in a dataset. It learned the variation in publication eligibility depending on the fraction, the amount of quasi-identifiers, the dataset size, and the sensitive value diversity. This research confidently took away several important details on how the classification of attributes affect the end anonymization of a dataset.

The study's proposed algorithm looked at many elements in a dataset for anonymization. It required two pieces of information on top of the dataset itself: the number of quasi-identifiers and the diversity value, if applicable. In cases where the datasets included a sensitive value the processor could opt-in to include the anonymization technique sensitive suppression. This allowed the research to be more versatile for users input on anonymization techniques depending on the dataset's need. In addition, the experiments showed the impact one anonymization technique has when it is

and is not applied to the process. The end-result created an adaptable algorithm that enables decision makers to process their dataset that best suits their needs.

Even if the proposed algorithm is not widely used, the study discovered a lot of quality information. First, the research illustrated the impact a single cell has on an overall dataset. It reviewed the effects attributes classified as a quasi-identifier or sensitive have on a dataset. The research learned the influence of quasi-identifier pairs and a sensitive value diversity when comparing the original dataset and the post-processed dataset. Expanding on the *1/2k* fraction resulted in a better understanding of the power different fractions have on a dataset's publication ability. This research is full of quality discoveries in addition to the end proposed adaptable algorithm. The study produced a detailed investigation on data and its sway on a dataset.

Definition of Terms

Attribute Disclosure	New knowledge associated with a record is discovered based on the published attributes (Li et al., 2007)
Background Attack	Attackers knows quasi-identifier(s) value(s) to deduce who is associated to record (Li et al., 2007)
Cell-Based Suppression	For the purpose of this study, cell-based suppression is defined as the quasi-identifier (from ranking order) elected for suppression
Counterfactual Reasoning	Determination that new information will be discovered about a person independent of the dataset being published or not (Machanavajjhala et al., 2015)
Differential privacy	Anonymization model that produces synthetic datasets from the original dataset (Dwork & Roth, 2009)
Explicit Attributes	Directly identifiable information on an individual (Nagendrakumar et al., 2014)

Generalization	For the purpose of this study, generalization is a cell modified from a specific value to a broader category range of values
Homogeneity Attack	Sensitive information exposed based on all similar records having the same value (Machanavajjhala et al., 2007)
Identity Disclosure	Associated person is identified to the record (Li et al., 2007)
Inferential Disclosure	New information discovered with extreme assurance through statistical findings on a record (Ciriani, Vimercati, Foresti, and Samarati, 2007)
k-anonymity	Anonymization model to require at least k amount of similar records in a dataset for a record to be eligible for publication (Sweeney, 2002a)
l-diversity	Anonymization model to require there be at least l amount of other sensitive value options in the k record pairs (Machanavajjhala et al., 2007)
Linkage Attack	Discovers new information about a person through the connection of attributes in multiple datasets (Dwork & Roth, 2014)
Noise	For the purpose of this study, noise is false data that are added to a dataset
Precision	For the purpose of this study, precision is the number of original cells divided by the total amount of end cells.
Quasi-identifiers	Commonly published attributes (Nagendrakumar et al., 2014)
Quasi-identifier pair	For the purpose of this study, quasi-identifier pair is all quasi-identifiers in a record linked together
Sensitive Attributes	Values that are special and/or delicate to an individual (Nagendrakumar et al., 2014)
Sensitive Suppression	For the purpose of this study, sensitive suppression is sensitive attribute eligible for suppression if l requirement is not fulfilled
Simulatability Approach	Ability to make the statistical dataset indistinguishable from the original dataset (Machanavajjhala et al., 2015)
Suppression	For the purpose of this study, suppression is the removal of a cell or record's content

Synthetic Data	Statistical data that are comparable to the original data (Garfinkel, 2015)
xk-anonymity	For the purpose of this study, *xk*-anonymity is anonymization model where a fraction(x) of *k* will determine the minimum amount of records in a pair required to allow for noise to be added, anything under the *xk* will be suppressed
1/2k theory	Anonymization model that requires there be at least *1/2k* amount of records in a pair for noise to be added (Brown, 2017)

List of Acronyms

E.U.   European Union

FERPA   Family Education Rights and Privacy Act

GDPR   General Data Protection Regulation

HIPAA   Health Insurance Portability and Accountability Act

Summary

Chapter 1 provided an overview of the research project. It gave a brief definition of some commonly accepted privacy-preservation models, data classifications, and anonymization techniques. The problem facing privacy-preservation today is there are no published models available to account the necessary fraction of *k* to balance the modification of a dataset. The research questions and goals followed to address how the study planned to solve the problem. Within the chapter, it detailed the significance and relevance for conducting this research, and what barriers it overcame. The research introduction provided assumptions, limitations, and research takeaways. Finally, the chapter ends by providing a list of key terms, definitions, and acronyms used throughout the paper.

Chapter 2

Review of Literature

Data availability has improved society; however, people worry about how information is used (van der Aalst, Blichler, & Heinzl, 2017). Information privacy dates back to 1945 (Smith et al., 2011). During that time, society had limited technology; most of the information collected was from the government and some businesses. Starting in 1961, an increased risk in information privacy disclosures began in the social, political, and legal arenas. By 1980, networked computer systems were able to hold more data than ever before, resulting in the 1984 Privacy Protection Act. In the U.S. today, the majority of systems have transitioned over to data systems connected to the entire world, exponentially increasing privacy concerns. Data-mining technologies enable the discovery of new insights; however, they pose a threat to privacy in today's global society (Thuraisingham, 2015).

Struggles rise as the goal of data usage conflicts with the privacy protection requirements (Garfinkel, 2015). Data are changing how people conduct business, research, socialize, and govern society (van der Aalst et al., 2017). According to Yaseen, Abbas, Anjum, Saba, Malik, ... and Bashir (2018), "Data publishing is obligatory for analysts" (p. 27156). Industries also benefit from the acquisition, sale, and analytical review of data (Garfinkel, 2015). To protect individuals' personal information, data must be pre-processed before publication (Bindahman, Arshad, & Zakaria, 2017).

Personal data are deemed de-identified when information can no longer identify or link an individual to the record (*OAIC*, 2014). Removing the identity of a person from the record is "technically and legally" complicated and requires special treatment across

different industries (Leichty & Leong, 2015, p.1). To de-identify records, all explicit attributes first must be masked or eliminated. Records are then reviewed for potential linkage by non-explicit attributes that could leave an individual vulnerable. Once data decision makers meet the specific requirements for their industry and affiliated countries, the process of de-identification is complete.

There are two usages for privacy-preservation: publication and data-mining (Garfinkel, 2015). Privacy-preserving Data Publishing (PPDP) creates synthetic data or de-identifies a dataset's personal information to satisfy publication requirements. PPDP completes a series of tasks to protect an individual's privacy (Rahmani, Amine, & Hamou, 2015). Privacy-preserving Data-mining (PPDM) is a type of data-mining that seeks to protect sensitive data while still accomplishing the data-mining goal (Chidambaram & Srinivasagan, 2014). In both cases, personal data is at risk of explicit information being exposed, which, in return, demands data protection.

The information gained from the data analyzers needs to measure the overall privacy risk (Li et al., 2007). The difference between the pre-processed dataset and post-processed dataset equals the information gained. The post-processed data should be valuable to learn new information about an entire population; however, it should not expose something unique to one individual record. Privacy rights today allow for overall new information to be discovered from a dataset, but any aspect of an individual's identity needs to be anonymized (Machanavajjhala & Kifer, 2015). Privacy-preservation data models' goals are to protect the individual and allow new information to be gained from a dataset.

14

Foundational Models

Data decision makers face challenges in releasing information without negotiating privacy; without data publication, the demand for data itself could decrease (Sweeney, 2002a). When data-mining two publicly available datasets, Sweeney (2000) discovered private information about an individual through common attributes the datasets shared. Further research found even if *k*-anonymity was fulfilled delicate information could expose an individual's identity, requiring a need to add a diversity requirement to the sensitive attributes (Machanavajjhala et al., 2007). Goswami and Madan's (2017) privacy-preservation model review included the four foundational methods to help future researchers build upon.

k-anonymity

Sweeney (2000) used zip code, birthdate, and gender to link medical data, provided by the Group Insurance Commission and Cambridge Massachusetts voters' registration list to expose the medical records of the governor of Massachusetts. This invasion of privacy to an individual illustrated a major weakness in regulations. Organizations were able to publicize data without concern on how other published datasets could allow data mining to discover an identity and sensitive information about a person. To address this issue, Sweeney (2002a) proposed *k*-anonymity to require a minimum amount of *k* records contain the same set of quasi-identifiers. By implementing this standard, the governor's medical history would have been protected because, in order for an organization to publish his record, at least *k* number of records would have had the same zip code, birthdate, and gender. This data processing requirement enhanced the

privacy of individuals and gave stricter guidelines to organizations before publishing

records.

l-diversity

The proposal of the *l*-diversity model addressed weaknesses in *k*-anonymity

(Machanavajjhala et al., 2007). The *l*-diversity model is one of the most commonly

developed models in privacy-preservation (Liu et al., 2011). The *k*-anonymity set a

standard on how many records were required to have the same quasi-identifier set, but the

model did not address the commonality of sensitive attributes (Machanavajjhala et al.,

2007). Researchers found even if the dataset was *k* anonymized if all the sensitive values

in those records were the same, then private information of an individual would be

exposed. To illustrate this, if *k=3*, the quasi-identifier set was zip code, birthdate, and

gender, and the sensitive attribute was diagnosis, then three records matching zip code=

06010, birthdate= 01/01/1970, and gender= male would meet the *k*-anonymity

requirement to be published. However, if all three of those records had a sensitive value

for cancer, then it would not matter who the specific individual was to each record.

Someone matching the quasi-identifier set would have their privacy invaded by

discovering they have cancer. To solve this invasion of privacy, researchers added the *l*-

diversity model to the *k*-anonymity model requirement to diversify the records by

including at least *l* many sensitive values, which eliminated a weakness to *k*-anonymity.

t-closeness

Limitations of *l*-diversity enabled researchers to propose *t*-closeness (Li et al.,

2007). The researchers' theory measured the privacy of an individual by the information

gained, the difference between post-belief and prior-belief based on the information

published in the dataset. This model required the distribution of sensitive attributes to be *t*-close to the overall distance of the table. The *t*-closeness model differs from *l*-diversity by limiting the difference between two records in the *k* record set. Researchers used the Earth Mover Distance metric to calculate the distance between sensitive values. They developed a hierarchy tree distribution to measure categorical sensitive values. Their goal was to have information learned from the overall table without contributing to the information gained about an individual record.

Differential Privacy

Differential privacy is a statistical approach to de-identify an individual by determining the probability of information being reported if an individual is or is not a part of the dataset (Dwork, 2009). Researchers' attention focused on preventing new risks to a person. Their theory believed an individual could be harmed by the knowledge produced by the dataset, but they would be harmed no matter if they were in the dataset or not (Dwork, 2016). The goal was to create a dataset with high statistical accuracy to the overall original data with a low risk of identifying an individual, minimizing the sensitivity of the published dataset. Adding minimal false information, otherwise referred to as noise, would enable a dataset to have low sensitivity (Dwork et al., 2006). In Dwork's (2016) discussion on differential privacy, the privacy model approach separates learning information about the dataset and the knowledge gained from the particular characteristics about an individual. The researcher noted that judgment calls were needed if attributes could be used to violate the privacy of people. Dwork (2016) produced positive results using differential privacy by studying synthetic datasets. These types of datasets are similar to the original dataset but kept an individual's privacy secure. Years

of research on differential privacy has been conducted to use statistical analysis to create synthetic datasets that protect identities while gaining new information.

Proposed Models

The *k*-anonymity, *l*-diversity, *t*-closeness, and differential privacy models are some of the commonly accepted privacy models in protecting a person's identity (Sei, Okumura, Takenouchi, & Ohsuga, 2017). Other models have been proposed using those models as baseline to improve data de-identification. The *t*-closeness model had improvements suggested by the original group of researchers that proposed the *t*-closeness model (Li, Li, & Venkatasubramanian, 2010). To allow for more utility in a published dataset, slicing was proposed to address restrictions from foundational models (Li et al., 2012). In conjunction with *k*-anonymity, clustering algorithms were used to remove the identification of an individual (Canbay & Sever, 2015 & Bindahman et al., 2017). Brown (2017) took *k*-anonymity requirements and applied a balance of noise and suppression to increase the records ability of publication. Various methods attempted to improve anonymizing datasets.

(n,t)-closeness

Researchers that proposed the *t*-closeness model found a better utility and privacy compromise when further breaking down the dataset, proposing a new closeness model, (*n,t*)-closeness (Li et al., 2010). The *t*-closeness model limited the release of valuable information depending on the overall population distance scale originally proposed. They found that breaking down the overall dataset into subpopulations could produce larger overall likelihood of publication. If there were at least *n* records to a population set that was at least *t* close to one another, then it would be a "natural superset" (Li et al., 2010,

p.5) and would be valid for publication. Natural superset means there are multiple sets of populations in one dataset. Independently, each population set meets the *t*-closeness criteria. To demonstrate where (*n, t*)-closeness = (1000, .1), if there were a 1000 records that held the zip code 061** and 2500 records that were 067** both sets would be valid for publication as long as among their sets individually *t*-closeness would not be higher than .1. In the original *t*-closeness proposal, if 061** did not meet the .1 set, by the rest of the population then the records were suppressed. This additional feature allowed for the release of more information about populations.

Slicing

The slicing approach improved on the restrictions that *k*-anonymity and *l*-diversity approaches did not address (Li et al., 2012). The *k*-anonymity method loses a substantial amount of data and *l*-diversity publishes quasi-identifiers where it is easy for attackers to re-identify individuals. To improve on these weaknesses, a new method focused on keeping highly correlated attributes together for research analysis. Slicing partitioned the dataset vertically and horizontally. To complete a vertical partition, the dataset would have attributes broken up into groups based on their correlation capabilities. Then, the dataset would be separated horizontally to create buckets. The buckets would randomly be re-organized so the two columns could not link an individual to the record. This model allowed no generalization to be completed and maintained a set *l*-diversity criterion. The research's example had eight records that contained three quasi-identifiers and one sensitive attribute: age, sex, zip code, and disease. First, age and sex were separated from zip code and disease. Next, the eight records got divided into two groups of four. After

slicing the dataset, the detail groups were randomly re-sorted to allow the original raw data to be kept with minor re-organization, preserving the data value and utility.

Clustering

Clustering techniques were used in combination with *k*-anonymity to improve data privacy and decrease disclosure risks (Canbay & Sever, 2015; Bindahman et al., 2017). Researchers found that limiting the data processing by *k*-anonymity resulted in less diverse data, but there were more lost data when only clustering information (Canbay & Sever, 2015). To build a stronger model, they combined the two techniques to create a 3-anonymized dataset. This approach used the Self-Organizing Maps algorithm to cluster the records. Canbay and Sever (2015) found when clustering the dataset more diverse groups of information would be available to process for anonymization. This increased the diversity and decreased the data loss. Bindahman et al. (2017) presented the S-Cluster approach to progress on data utility and privacy. This model wanted to eliminate gender and age disclosure risk. By processing the dataset with the S-Cluster, the researchers were able to improve the quality of published data. Both clustering techniques worked to diversify and enhance data utility (Canbay & Sever, 2015; Bindahman et al., 2017).

1/2k Theory

Research was conducted to build on *k*-anonymity, introducing *1/2k* theory (Brown, 2017). The concept of *1/2k* theory defined if the quasi-identifier pair had at least 1/2 the amount of *k* records required, it was more beneficial to create false records than removing the original records (Brown, 2017). This research compared a dataset being processed with *1/2k* and *k*-anonymity. Of the four experiments completed, when the dataset was processed with *1/2k* theory, less record suppression was done, with better

total original record count, and higher quasi-identifier pairs, providing a more diverse dataset. In the first experiment alone, there were major differences between using traditional *k*-anonymity approach and the proposed *1/2k* approach. The *1/2k* approach resulted in 4,761 more original records eligible for publication by only adding 2,284 false records. Between the two approaches there was over 7% difference of original records eligible for publication when implementing the *1/2k* theory. This research introduced a fraction of *k* could increase the publication capability of a dataset.

Attribute Classification in Models

Attribute classification is critically important to privacy-preserving models. A study found that 87% of people in the U.S. could be identified solely on three quasi-identifiers: birthdate, gender, and zip code (Sweeney, 2000). The *k*-anonymity method used quasi-identifiers to help propose protection in re-identify records, like the governor of Massachusetts (Sweeney, 2002a). The *l*-diversity model illustrated how it was not sufficient to incorporate only quasi-identifiers to protect datasets, but to include a required diverse amount of sensitive attributes (Machanavajjhala et al., 2007). Based on the privacy-preserving models available today, all models illustrated the significance of attributes classification to protect privacy.

On top of the standard four classifications of attributes, researchers built a new classification to accommodate another perspective of possible attribute grouping. Shi, Xiong, and Fung (2010) proposed an additional category called quasi-sensitive attributes. These attributes would not necessarily be appropriate to categorize as sensitive on their own but when congregated with other known information could expose an individual. Their example used the attribute *symptoms*. This attribute alone would not be sensitive;

however, when linking the symptoms to a sensitive attribute like *disease*, there is a potential privacy breach. This additional category enables another option for a data processer to identify their attributes.

Attribute classification was highly reliant in Brown's (2017) research on a data privacy model for correlational research. The research examined quasi-identifiers and the effects in using them as a growing pair set. In the study, it was more efficient to introduce the quasi-identifiers gradually than to group them as a whole and process the dataset once. To demonstrate this, it would be better first to process two quasi-identifiers and then continuously re-process the pair by adding another quasi-identifier until all quasi-identifiers were included. Furthermore, the research found that creating a ranking system to quasi-identifiers influenced the overall number of records eligible for publication. This research changed the order of processing between YOB and gender. As a result, when the approach processed gender before YOB more original unmodified records, more total records, and more quasi-identifier unique pairs were publishable. Brown (2017) highlighted how attribute priority and gradual grouping of quasi-identifiers increase publication.

Anonymization Techniques

Privacy-preservation models use anonymization techniques to de-identify records. The two most common techniques include the use of generalization and suppression of data. Generalization modifies a cell from a specific value to a broad range of values (Nagendrakumar et al., 2014). This technique is the basis of *l*-diversity and *t*-closeness (Sei et al., 2017). Data warehousing, data-mining, and machine learning all use the generalization technique (Yaseen et al., 2018). Generalization itself has different

approaches. Some researchers used generalization hierarchies to determine the best fit for an attributes' value (Sweeney, 2002b; Yaseen et al., 2018). Other proposed research approaches generalized quasi-identifiers at a single level grouping (Brown, 2017) and fluctuated the generalization range to make the original data eligible for publication (Angiuli & Waldo, 2016). Overall, several studies included the generalization technique when completing anonymization.

Suppression removes content from the dataset that does not meet anonymization requirements. Some models used the addition of noise to mask the true identity of a record (Sei et al., 2017; Angiuli & Waldo, 2016; Brown, 2017). Sweeney (2002b) used suppression technique in the *k*-anonymity model to protect individual's information when they were easy to identify. Additionally, Sweeney (2002b) suppressed part of an attribute's value to increate record privacy. Brown (2017) approached suppression differently by suppressing records that were under $1/2k$. Throughout numerous studies, there were some kind of suppression incorporated. Using generalization and suppression in excessive quantities decreases the data usability (Nagendrakumar et al., 2014); however, there are other anonymization techniques used to help maintain data utility.

Other anonymization techniques include data falsification, synthesizing, and bucketization (Angiuli & Waldo, 2016; Brown, 2017; Dwork, 2009; Li et al., 2012). Data falsification approaches minimize a dataset's loss of records to suppression by adding false records (Angiuli & Waldo, 2016; Brown, 2017). Some models proposed creating synthetic data to make a comparable dataset to the original (Dwork, 2009). Synthesizing dataset is different as it replaces the original records with statistically proportional information for publication. Bucketization maintains all original records, but horizontally

and vertically re-organizes the data set to preserve high correlation between attributes (Li et al., 2012). Proven as viable option, generalization, suppression, falsification, bucketization, and synthetization are suitable anonymization techniques to de-identify datasets for publication.

Angiuli and Waldo (2016) examined how generalization technique could fulfill *k*-anonymity in their research to publish a larger dataset. Instead of traditional grouping for generalization, they proposed a new "greedy algorithm" (Angiuli & Waldo, 2016, p.592). This algorithm reviewed how many additional records would be required to meet *k* and paired the record(s) with the closest value to achieve *k*-anonymity requirements. Records that already met *k* were not affected in the example they provided. This proposed technique enabled more records to be kept for publication. The second element in their research was instead of suppressing any records they added fake records to bring the record count to acceptable *k* minimum requirements. Their results found that applying the greedy algorithm increased publication ability; however, the noise correlation was negatively impacted.

Follow-up research was done on Angiuli and Waldo's (2016) research to balance out suppression and data falsification with *1/2k* theory (Brown, 2017). This study kept traditional single layered generalization by grouping YOB in five-year increments. Instead of adding noise to all records to remove suppression, the research looked to negotiate suppression by reviewing if there were at least half the records of *k*. If the records qualified as *1/2k* then the records would apply noise. This would avoid suppression and benefit the end original records count. The example applied in the research had *k*=5, so if there were at least three records in the dataset matching the quasi-

identifier pair, false records were added to the dataset. If there were one or two records then the record(s) were suppressed. Brown's (2017) approach adjusted traditional suppression technique to benefit publishing more of the original content.

False and Synthetic Data

Datasets that add noise or change to synthetic datasets for privacy-preservation have adverse effects. One proposed model chose to remove suppression by adding false records to meet *k*-anonymity (Angiuli & Waldo, 2016). The research discovered, however, there was a negative impact on correlating attributes. Noise is a perturbation technique (Agrawal & Srikant, 2000). False record creation enables more original content to be published (Brown, 2017), but it hinders the precision of the dataset. A balance must be made to compromise on additional noise and record suppression.

Synthetic data created from a statistical representation of the original data content (Fung, Wang, Chen, & Yu, 2010). The dataset additionally could be partially or fully synthetic (Dreschsler, Bender, & Rässler, 2008). Partially synthesized datasets hold more utility, but the disclosure risk increases. Information confidence issues in modified datasets empowered fully synthesized datasets (Rubin, 1993). These types of datasets are not practical in some research scenarios as they do meet "truthfulness at the record level" (Fung et al., 2010, p.4). Even though differential privacy, a fundamental model, is a valid option for privacy-preservation, the restrictions on a dataset impedes the usability of the published dataset.

Security Issues

Data-mining threatens privacy (Aggarwal & Yu, 2008). Li et al. (2007) broke down two ways information could be disclosed from privacy-preservation models,

identity or attribute based. Moreover, there are three main disclosures risks (Li et al., 2007; Ciriani et al., 2007). *Identity disclosure* could happen when a record is published and an individual is re-identified to that record (Li et al., 2007). *Attribute disclosure* could happen when there is information gained from an individual based on an attribute being published. Connecting quasi-sensitive attributes along with known knowledge could result in an indirect disclosure of information (Shi et al., 2010). In an identity disclosure event an attribute disclosure could proceed; however, if there is an attribute disclosure there is not necessary an identity disclosure. *Inferential disclosure* threat was an additional disclosure type released by Ciriani et al. (2007). This disclosure type is a serious invasion of people's privacy because attackers are able, with high confidence, use statistics to discover a person's identity. Individuals have the right to control the personal information they release (Tene & Polonetsky, 2013). If datasets are published that reveal any personal details of an individual then their privacy has been violated.

As studies evolved, attacks on privacy have affected identity preservation causing newly proposed models to be formed. There were three attacks addressed based on two of the fundamental privacy models. Sweeney's (2002a) *k*-anonymity model looked to protect individuals from *linkage attack*. This attack occurred when two datasets were combined to learn new information about a person. Machanavajjhala et al. (2007) found weaknesses in *k*-anonymity that could result in homogeneity and background knowledge attack. *Homogeneity attacks* are the result of all records that have the same composition of quasi-identifiers also having the same sensitive value revealing personal information. *Background knowledge attacks* happen when a person knows all quasi-identifiers of another individual and can deduce which sensitive value is linked to that person. Other

research published an attack classified as minimally attack. This type of attack occurs when there is new information gained about an individual because the nominal protection on their personal information (Wong, Fu, Wang, & Pei, 2007). There are challenges in preserving privacy which has allowed attackers to use external and public data sources to obtain information to re-identify a user (Gkoulalas-Divanis & Verykios, 2009). Sweeney's (2002a) and Machanavajjhala's et al. (2007), *k*-anonymity and *l*-diversity models have minimized the risk of major attacks.

In differential privacy, decisions have to be made whether an individual's privacy is at risk due to the information published in their dataset. Counterfactual and simulatability approaches are examined to see if the dataset is causing an individual's privacy breach (Machanavajjhala & Kifer, 2015). *Counterfactual reasoning* based idea showed releasing private information from a database is allowed but divulging information about an individual is an invasion of privacy. A probability test checks to see if there is an equal chance of an individual having the sensitive value. This test also see if an individual's probability does not have the sensitive value. For example, in differential privacy model, the algorithm is like a coin toss, there is a 50/50 chance that the person is or is not a part of the dataset (Dwork, 2009); the same concept applies to a counterfactual test. There is a 50/50 chance the person does or does not have the sensitive attribute. *Simulatability approach* focuses to ensure attackers are unable to identify the statistical dataset from the original dataset (Machanavajjhala & Kifer, 2015). Before data publication, these two tests are available to measure privacy disclosure risks.

Legal Obligations

"Consumer data privacy and security are critical areas of opportunity and concern for industry and policymakers" (Listokin, 2017, p.92). Personal information has become easier to collect with the evolution of technology (Lu, Li, Qu, & Hui, 2014). Legislators have worked to regulate how people's data can be obtained, utilized, and stored. The E.U. in May 2018 began enforcement of a new data protection law, which makes it the most momentous update since the mid-nineties (Arend, 2017). Security and privacy laws in the U.S. are not set at the national level but instead determined by industry and state (Breaux & Gordon, 2013). As laws are implemented and modified, privacy-preservation approaches have to adapt to current regulations, which makes it critical to be knowledgeable about consumer protection laws.

Taken affect earlier this year, the General Data Protection Regulation (GDPR) is the new E.U. law. This law encourages the transformation of personal data so an attribute cannot be linked to a specific person without additional data by pseudonymization (Maldoff, 2016). This means even prior to companies potentially publishing or using collected data, some alterations should be completed. If found not in compliance with GDPR, a company could be fined four percent of its annual revenue, or up to 20 million Euros, whichever returns the higher charge (Arend, 2017). For example, if the Equifax data breach occurred when GDPR was in effect, the company would be obligated to paid up to 126 million dollars (Goldman, 2017). This proved the criticality of privacy-preservation models and the importance of ensuring models are created that meet current and future legislation.

The U.S. does not have a national set of privacy laws for personal information; regulations are left up to states and the industrial sectors (Breaux & Gordon, 2013). FERPA and HIPAA are two common standards where legislation requires individuals' information be safeguarded based on the industry. FERPA prevents disclosure of student records' explicit information in educational institutes that receive government funding (Apricorn, 2016). HIPAA is less ambiguous about how consumers' information is to be protected. The health industry defines two privacy rules: expert determination and safe harbor (HHS, 2012). Expert determination requires qualified personnel apply statistical principles to de-identify health information. Safe harbor lists 19 specific attributes that require suppression prior to publication. Since industries define privacy regulations in the U.S., it would require a great deal of collaboration to define a unified national standard (Breaux & Gordon, 2013). This makes privacy-preservations models more complex because it would require algorithms to standardize features that meet a range of industry criterions.

Summary

Previous research helped provide background knowledge and insight to some expectations for this study. Even though technology has advanced, there are to this day restrictions on how individual's identity can be anonymized from privacy-preservation models (Dwork & Roth, 2014). Commonly accepted privacy approaches like *k*-anonymity, *l*-diversity, *t*-closeness, and differential privacy models have set a standard on how data can be de-identified (Sharma & Rajawat, 2016). A study found ranking quasi-identifiers and introducing them gradually to a de-identification process resulted in an improvement of publishable amount of records (Brown, 2017). De-identification

approaches were re-examined to improve how to apply techniques to anonymize data (Angiuli & Waldo, 2016). Improvements on Angiuli and Waldo research found a way to balance suppression and data falsification (Brown, 2017). Many other proposed models have improved de-identification techniques. Security issues about privacy-preservation models requires continuously assessment and tested to ensure the highest confidence in anonymizing a dataset. Corporations must meet legal obligations, so a privacy-preservation method needs to be adaptable for data processers to set their anonymization to the company's legal requirements. This research was cognitive on potential attacks that can occur on published datasets from weaknesses in current privacy models. Altogether, the review of previous research provided a solid foundation for this study.

Chapter 3

Methodology

Introduction

   To create an adaptable algorithm effectively, several aspects of a dataset required studying. The review of cell-based suppression allowed for the study to better understand the influence one attribute had on the overall dataset. All processed datasets measured the precision rate. This research also expanded on Brown's (2017) *1/2k* theory by applying different fractions to a dataset. The use of fractions measured the change in data precision depending the *xk*-anonymity. The study changed the amount of quasi-identifiers to see the impact on precision at different fraction levels of *k*. Additionally, two of the experiments included *l*-diversity to expand on *k*-anonymity for privacy confirmation. All of the experiments documented the changes to pre- and post-datasets depending on the set criteria. All data collected including the total record count, the quasi-identifier pair, the suppression count, the added records, the cell-based suppression total, and the generalization total aided in the development of the adaptable algorithm.

Research Design

   This experimental dissertation study applied a series of experiments to assist in the creation of the end algorithm. Each experiment addressed one of the study's goals. Experiment 1 measured the influence a single quasi-identifier had on a dataset when suppressed prior to the processing the overall dataset. Experiment 2 analyzed the impact on a dataset publication depending on the dataset size. Experiment 3 demonstrated the influence when processing datasets at different fractions of *k*. Experiment 4 reviewed the fluctuation in post-processed dataset when changing the different amounts of quasi-

identifiers. Experiments 5 and 6 included *l*-diversity to measure the change in data

suppression. Together, all six of these experiments were essential in the accomplishment

of the research's goals and the formation of the proposed adaptable algorithm.

Experiment 1: Cell-Based Suppression

The first experiment analyzed the impact on a single quasi-identifier to see the

end difference in publication eligibility. The MIT and Harvard first year of *edx* course

dataset contains 641,138 published records; the study extracted 250,000 records for a

sample set. Five attributes were marked as quasi-identifiers, with the eligible cell-

suppressible quasi-identifier as YOB. There were four different methods applied to the

250,000 dataset.

1. No Generalization, No Cell-Based Suppression
2. Both Generalization and Cell-Based Suppression
3. Only Cell-Based Suppression
4. Only Generalization

These four methods used a single-processing method completed twice, first using the

1/2k theory and second with traditional *k*-anonymity. The single-processing method

compared all five quasi-identifiers at the same time, unlike Brown's (2017) research that

used a gradual attribute introduction method. The post-processed dataset measured

precision and utility to comprehend the influence the dataset has when using cell-based

suppression alone and combined with generalization.

Experiment 2: Different Size Datasets

The second experiment focused on the precision impact when using different size

datasets. The key for this experiment was to focus on the precision rate change. The

overall objective was to construct an adaptable privacy-preservation algorithm that could apply to diverse style datasets. This experiment showed how the size of a dataset affected the precision rate. From the sample set of 250,000 records, there were five different sub-datasets created: 30,000; 60,000; 120,000; 180,000; and 240,000. By processing a variety of different size datasets, the research expanded from the first experiment to see the change in precision. Furthermore, it measured the percentage of the remaining quasi-identifier pairs against all the other datasets. The same four methods, with five quasi-identifiers, were processed twice using the *1/2k* theory and traditional *k*-anonymity.

Experiment 3: xk-anonymity

The third experiment measured the falsification percentage level of *k*. Different fraction levels were applied on top of Brown's (2017) 1/2*k* fraction. The fractions included 1/4, 1/3, 1/2, 3/4, and 1. The goal of this experiment was to see how different fractions(*x*) of *k* changed the precision of the dataset. The *xk*-anonymity model injected a different fraction to each of the sample datasets. Experiment 3 maintained a single-processed approach for each dataset and applied all four methods. From this experiment, the research began to form the foundational platform on how precision rate changed at different *xk*-anonymity approaches.

Experiment 4: Different Number of Quasi-identifiers

As discussed in Yang's et al. (2013) research, as the number of quasi-identifiers increased, the amount of data loss increased. The fourth experiment introduced a number of different quasi-identifiers. This element incorporated the previous experiments' focuses on cell-based suppression, dataset sizes, and *xk*-anonymity. There were four different quasi-identifiers measured: 2, 3, 4, and 5. This research used Brown's (2017)

quasi-identifier ranking system, from most important to least important, to list the attributes as course ID, forum post, gender, YOB, and country. The first two quasi-identifiers processed were course ID and forum post, as they were the most important attributes. When three quasi-identifiers were processed, it still used course ID and forum post but added gender. The four quasi-identifier pair process added YOB on top of the other previous attributes. Lastly, the study measured all five attributes together. The experiment also processed the datasets using the single-process approach. This experiment showed the end precision change when fluctuating the quasi-identifiers.

Experiment 5: Inclusion of l-diversity

The fifth experiment added the inclusion of *l*-diversity. Machanavajjhala et al. (2007) found *k*-anonymity weaknesses; to address this weakness, they required *l*-diversity to protect individuals' sensitive information from being exposed. This addition proved essential for data de-identification. For Experiment 5, all six datasets required modifications to the grade attribute. The grade attribute applied random false numerical values to measure *l*-diversity. The numerical range included 1, .95, .90, .85, .80, .75, .70, .65, .60, and 0. These values represented a grading scale a professor would give in a classroom environment, where .95 would be the minimum value for an A, .9 = A-, and so forth until 0 = F. The datasets were processed using *l* equaled to three different diversity levels: 2, 3, or 4. This allowed for the study to understand the impact a sensitive value has on the entire dataset, depending on the dataset's size, what fraction of *k* it is being processed by and with different amounts of quasi-identifiers within the dataset.

Experiment 6: Suppression of Sensitive Value

Experiment 6 suppressed sensitive values that did not meet *l*-diversity. This was a crucial experiment as records might not meet *l* requirements, but the quasi-identifier pair meets *xk*-anonymity, and the record could provide quality information. This experiment suppressed any grade cell for records that did not satisfy the *l* demand. When implementing this experiment, the four methods from Experiment 1 increased. In total, there were eight methods processed:

1. No Generalization, No Cell-Based Suppression, No Sensitive Suppression

2. All Three Included Generalization, Cell-Based Suppression, Sensitive Suppression

3. Generalization, Cell-Based Suppression

4. Generalization, Sensitive Suppression

5. Cell-Based Suppression, Sensitive Suppression

6. Only Cell-Based Suppression

7. Only Generalization

8. Only Sensitive Suppression

By adding the four methods, there were more combinations of anonymization techniques to analyze the end impact to the post-processed dataset. Data decision makers could choose any of the eight methods to complete privacy-preservation. The experiment measured the impact on a dataset when including sensitive suppression at different fractions and quasi-identifiers.

Sample Set

This research extracted a sample set from Dataverse *HarvardX-MITx Person-Course Academic year 2013 De-Identified* (*Harvard,* 2014). This publicly available comma-separated value (csv) file contained 641,138 records with 20 different attributes. The study created two sample sets from the published dataset: experiment dataset and test dataset. The experiment dataset contained 250,000 records, and the test datasets extracted 500,000 records from the original dataset. From both sample sets, there were sub-datasets created. To ensure the experiment dataset proportionally retrieved a quality sample of the original dataset, there were two main elements considered: course ID and country. The order of creating the experiment sample set went as followed:

1. Added column U to assign random values, used RAND function to set the record's cell value

2. Segregated dataset by course ID into different tabs

3. Copied column Us' values to column V (to make a static number)

4. Sorted the rows by smallest to largest depending on column V value

5. Started at A1 selected records downward, until right the proportion of records were extracted

6. Copied selected records to new excel workbook titled sampleset_250000

The test sample set used steps one, three, and four to create the dataset. The lowest 500,000 records made up the test sample. This allowed the test to avoid any proportional considerations and to have truly a unique collection of records to verify the authenticity of the algorithm.

The experiment sample extracted a proportional representation of the original dataset. Illustrated in Table 1, the study calculated the value each course had on the overall dataset. For example, out of the 641,138 published records, HarvardX CB22x 2013_Spring had 30,002 published records. That course ID had roughly 4.7% of the total records count, so the study pulled 11,750 records to equal the correct proportion of the 250,000 record sample set. Some records required modified rounding to assist in the

Institution	Course	Term	Published Records	Percentage Of Total	Rounded Thousandths	Final Count	Pulled Records
HarvardX	CB22x	2013_Spring	30002	0.0468	0.047	0.05	11750
HarvardX	CS50x	2012	169621	0.2646^2	0.265	0.26	66000
HarvardX	ER22x	2013_Spring	57406	0.0895^1	0.09	0.09	22250
HarvardX	PH207x	2012_Fall	41592	0.0649	0.065	0.07	16250
HarvardX	PH278x	2013_Spring	39602	0.0618	0.062	0.06	15500
MITx	14.73x	2013_Spring	27870	0.0435^1	0.044	0.04	10750
MITx	2.01x	2013_Spring	5665	0.0088	0.009	0.01	2250
MITx	3.091x	2012_Fall	14215	0.0222	0.022	0.02	5500
MITx	3.091x	2013_Spring	6139	0.0096^2	0.01	0.01	2500
MITx	6.002x	2012_Fall	40811	0.0637	0.064	0.06	16000
MITx	6.002x	2013_Spring	22235	0.0347	0.035	0.04	8750
MITx	6.00x	2012_Fall	66731	0.1041	0.104	0.1	26000
MITx	6.00x	2013_Spring	57715	0.09	0.09	0.09	22500
MITx	7.00x	2013_Spring	21009	0.0328	0.033	0.03	8250
MITx	8.02x	2013_Spring	31048	0.0484	0.048	0.05	12000
MITx	8.MReV	2013_Summer	9477	0.0148	0.015	0.02	3750
			641138	1.0002	1.003	1	250000

Table 1 Course Sample Set Extraction

extraction of the 250,000 records. Identified with a superscript 1, those records had the ten-thousandths value of five, but rounded down to help balance. Identified with superscript 2, those records had a ten-thousandths value equaled to six, one of the two were rounded down to extract the set sample value. After collecting the sample set, the research then crosschecked the sample to the countries' representations. The difference

between the published dataset's country representation and sample set representation was under ten-thousandths of a difference.

Instrument Development & Validation

The study built two Java-based programs to complete each experiment and test the end algorithm. The data processing code enabled a semi-automatic process to anonymize a dataset. There were seven steps assembled to the program: unification, sensitive value diversification, generalization, cell-based suppression, xk-anonymity processing, suppression, and confirmation. Since there were multiple inputs that provided zero information as quasi-identifiers (-, NA, *blank*), the unification step combined all three texts to equal "-". This allowed the quasi-identifier pairs to equally be associated if there were no valid information given about that attribute. The second step needed only to occur when l-diversity requirement was in the dataset. This step handled the suppression of the entire record for traditional l-diversity approaches. This step also handled the opt-in for sensitive suppression. The third step reviewed the quasi-identifiers that were under the xk-anonymity criteria and generalized the YOB cell. In the fourth step, if quasi-identifiers did not meet xk-anonymity, the program suppressed the YOB cell. For the fifth step, xk-anonymity reviewed each quasi-identifier pair, if the criteria met xk-anonymity, but was under k-anonymity, the dataset added false records until the pair met k-anonymity. If the criteria was under xk-anonymity, it marked the associated records for suppression. The sixth step deleted all records marked for suppression. The last step confirmed the end quasi-identifier pairs' total, the records' total, and that zero modification needed to occur for the dataset to meet k-anonymity and l-diversity.

By breaking the data processing program down into seven steps, the study could complete any of the eight anonymization methods. To best highlight this, below are the steps used in each of the eight methods for Experiment 6:

1. No Generalization, No Cell-Based Suppression, No Sensitive Suppression (Steps 1,2,5,6,7)

2. All Three Included Generalization, Cell-Based Suppression, Sensitive Suppression (Steps 1,2,3,4,5,6,7), opt-in Sensitive Suppression on step 2

3. Generalization, Cell-Based Suppression (Steps 1,2,3,4,5,6,7)

4. Generalization, Sensitive Suppression (Steps 1,2,3,5,6,7), opt-in Sensitive Suppression on step 2

5. Cell-Based Suppression, Sensitive Suppression (Steps 1,2,4,5,6,7,), opt-in Sensitive Suppression on step 2

6. Only Cell-Based Suppression (Steps 1,2,4,5,6,7)

7. Only Generalization (Steps 1,2,3,5,6,7)

8. Only Sensitive Suppression (Steps 1,2,5,6,7), opt-in Sensitive Suppression on step 2

Each step writes two csv files, a quasi-identifier pairs' information list and a post-processed dataset. The quasi-identifier pairs' file included all attribute values in the pairs' set, the number of records in the pair, and the diversity of the sensitive value. The creation of each file empowered the study to back-up the step's action on the dataset continuously.

To validate the program's accuracy, a random spot check routinely reviewed the records for proper modifications. The quasi-identifier pair file documented the imported

dataset file. This file is valuable when crosschecking the changes of the pre-processed and post-processed dataset. A spot check reviewed three main elements: if records were under xk-anonymity that they were marked for suppression, over xk-anonymity but under k-anonymity false records were added, and pairs that met k-anonymity records were left alone. In addition, the spot checks confirmed correct generalization and cell-suppression.

The end algorithm's Java-based program translated the proposed algorithm to automatically process. The program first ingested and decoded the given csv dataset to identify the quasi-identifier pairs, the dataset's record count, and the sensitive diversity. It provided questions for a user to outline the process criteria. Based on the user's answers, the program ran all available xk-anonymity options and documented the results. The program then compared the results to see which approach provided the best solution for the dataset with the given requirements. Afterwards, it calculated the precision range and data loss elements of the algorithm. At the end, the program outputted the recommended approach, precision range, and maximum data loss. The end algorithm's Java-based program confirmed the proposed algorithm's accuracy by processing each available approach with and without anonymization techniques through the data processing Java program.

Data Analysis

There is no commonly accepted standard for measuring the utility loss of a de-identified dataset (Garfinkel, 2015). A key element in any research is the ability to measure success. Previous researchers built utility matrices to measure pre- and post-processed datasets (Yang, Li, Zhang, Yu, 2013; *Dataverse*, 2014). To measure utility, the research built correlation matrices with Excel's XLMiner Analysis Toolpak. Then, the

study measured precision with a similar formula to Sweeney's (2002b) generalization and suppression research. First, each cell in the pre-processed data had a set value of one. When the cell required modification, the cell's value reduced. Sweeney's (2002b) research measured values based on their hierarchy level of generalization, more generalized the lower the value. Since this research had one level of generalization, a generalized cell's value reduced a half a point. If the cell required suppression, the cell's value deducted the entire point. At the end of the dataset processing, all cells' values made up the total post-processed value. That value divided by the pre-processed total points resulted in the dataset's precision.

Milestones

There were six phases to this dissertation research project. The first phrase extracted two sample sets of the selected dataset. The second phase developed a Java program and completed all six experiments. After the experiment finalized, there was enough information collected to analyze and build an algorithm in the third stage. Once the initial model completed development, the fourth stage tested the system for certainty with the original experiment datasets and test datasets. The testing results provided feedback to modify the end algorithm. When all testing and modification completed the study transitioned to the final stage.

Phase I	Extract Datasets & Modifications	1 Week
Phase II	Coding & Experiments	19 Weeks
Phase III	Create Algorithm	5 Weeks
Phase IV	Testing	5 Weeks
Stage V	Modify	1 Weeks

Stage VI	Final Report	5 Weeks

Phase II: Experiments

1. Cell-Based Suppression (YOB)
2. Different Size Datasets (30,000, 60,000, 120,000, 180,000, 240,000, 250,000)
3. *xk*-anonymity (1/4, 1/3, 1/2, 2/3, 3/4, 7/8, 1)
4. Different Quasi-identifiers (2, 3, 4, 5)
5. Inclusion of *l*-diversity (Grade)
6. Suppression of *l* (Grade)

Milestones	9/9/17	11/30/17	4/4/18	4/21/18	5/12/18	5/26/18	7/14/18	8/4/18	8/25/18	9/29/18	11/5/18	12/15/19	1/16/19
Idea Paper Approval													
Proposal Paper													
Proposal Defense													
Experiments I													
Experiment II													
Experiment III													
Experiment IV													
Experiment V													
Experiment VI													
Algorithm													
Test/Modify													
Write Report(s)													
Peer Editing													
Submit Research													
Dissertation Defense													
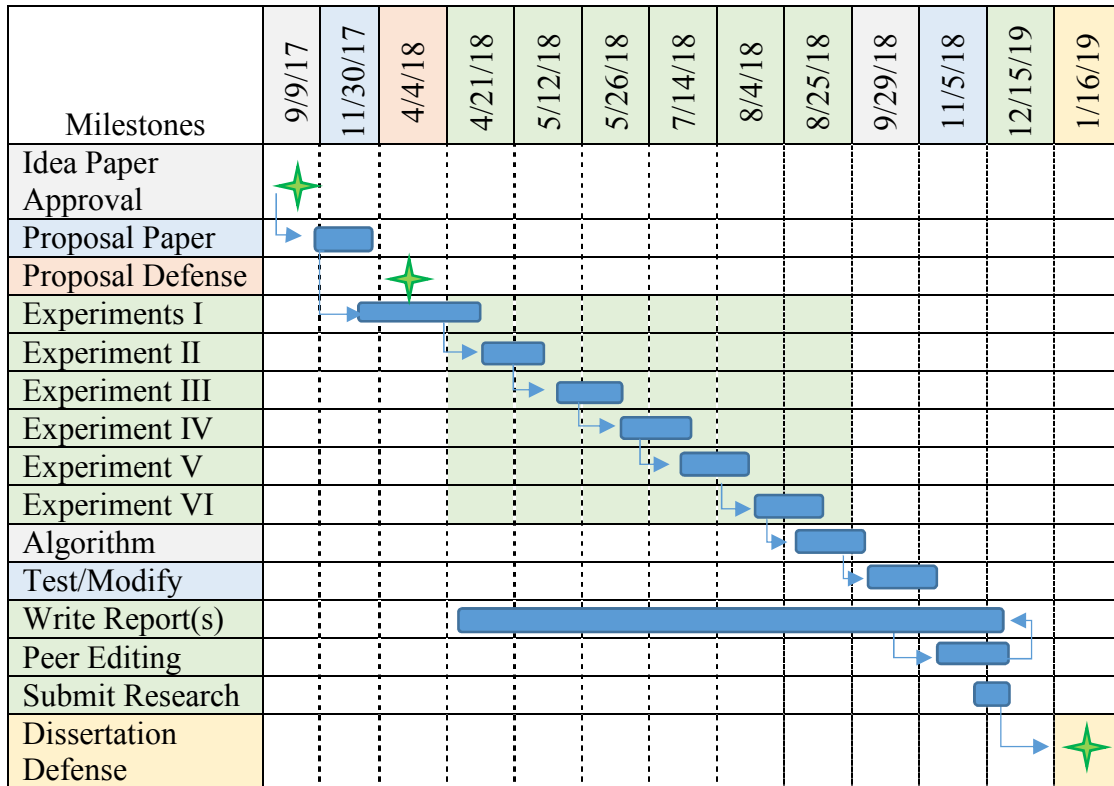
Resources

This dissertation research required several elements. First, the study required the

Dataverse's dataset to conduct the series of experiments, as well as create the test dataset

samples. The adaptable algorithm used foundational elements of the code designed for

Brown's (2017) *1/2k* theory study. Java Eclipse used the Java IDE to design the

programs. A Lenovo ThinkPad laptop with an Intel core i7 conducted all the experiments,

tests, and validations. This laptop hosts a Windows 7 operating system, which has 8GB of memory.

Summary

The methodology chapter outlined the action plan for the dissertation. There were six phases to the study including testing the final proposed algorithm and dissertation report writing process. Phase two divided into six experiments. The experiments broke down the different aspects of the research: quasi-identifiers, sensitive value, cell-based suppression, sensitive suppression, and xk-anonymity. Lastly, this chapter contained details on the required resources, data sampling, and milestones necessary to accomplish this research.

Chapter 4

Results

Data Analysis

Each experiment aided in the development of this study's proposed algorithm. Experiment 1 demonstrated a single quasi-identifier attribute could positively affect the outcome of the post-processed dataset. Experiment 2 showed as the dataset sizes increased the precision also increased. Experiment 3 illustrated different fractions balanced a dataset's suppression and falsification count. Experiment 4 proved when a dataset of the same size has different quasi-identifier amounts, their end balance of suppression and falsification changed. Lastly, Experiments 5 and 6 focused on *l*-diversity; these experiments highlighted when a dataset included sensitive values it can considerably affect the original record outcome if sensitive suppression is not applied. All the data analysis done from these experiments led to the creation of the adaptable algorithm.

Experiment 1: Cell-Based Suppression

The application of anonymization techniques on a dataset proved to have advantages and disadvantages. In both the *1/2k*-anonymity and *k*-anonymity approaches, traditional methods without any anonymization techniques resulted in the highest precision, outlined in Table 2. In contrast, traditional methods had the highest suppression count and lowest amount of original records. When applying both generalization and cell-based suppression on the dataset, the post-processed dataset had the higher total record count, but also had the most added records. Approaches that included cell-based suppression

15268	x--	xGC	xC	xG	T--	TGC	TC	TG
Total	250063	254999	254473	252739	234433	248510	249101	244125
End QP	12298	13126	12773	12821	9172	10852	9876	10446
QP Remain %	80.5476	85.9707	83.6586	83.9730	60.0734	71.0768	64.6843	68.4176
Suppressed	4792	680	490	2761	15567	1490	899	5875
Added	4855	5679	4963	5500				
Precision %	98.0584	97.7148	97.9679	97.8038	100	99.8235	99.7157	99.9007
Original	245208	245339	245349	245208	234433	234894	234936	234433
OwM		249320	249510	247239		248510	249101	244125

Table 2 Experiment 1 Comparison Chart (250k)

had the highest amount of original records and the lowest suppressed records. Between

generalization and cell-based suppression techniques, when applying only generalization,

or when applying generalization before cell-based suppression, the dataset had higher

suppression counts, more added records, and less original records with modifications than

the dataset processed with only cell-based suppression.

This analysis began the process of extracting key pieces of information for the

end algorithm. Primarily, cell-based suppression resulted in the highest original records,

including original modified records and the lowest suppression. This means even if the

traditional method only applied cell-based suppression, there was a higher potential of

saving more records from suppression. Traditional methods provided the maximum

precision value, which means any approach with anonymization techniques applied

decreased the precision. The more details collected, the more versatile the algorithm

became.

Experiment 2: Different Size Datasets

Transitioning from Experiment 1 to Experiment 2, the anonymization pattern remained

the same, which defended Experiment 1's analysis. From the original dissertation

proposal, Experiment 2 hypothesized that it would show how the dataset size affected the

precision. Displayed in Table 3, as the dataset sizes increased the precision increased.

30	--	GC	C	G	Avg	Difference
1/2k	90.8157	87.6195	91.2931	88.1169	89.4613	
k	100.0000	98.8234	98.1361	99.2743	99.0585	

60	--	GC	C	G	Avg	Difference
1/2k	93.5459	91.2700	93.5090	91.7127	92.5094	3.0481
k	100.0000	99.2651	98.7651	99.5039	99.3835	0.3251

120	--	GC	C	G	Avg	Difference
1/2k	96.0089	94.8685	95.9124	95.0355	95.4563	2.9469
k	100.0000	99.6002	99.3191	99.7347	99.6635	0.2800

180	--	GC	C	G	Avg	Difference
1/2k	97.2885	96.6435	97.1861	96.7652	96.9708	1.5145
k	100.0000	99.7383	99.5644	99.8367	99.7849	0.1213

240	--	GC	C	G	Avg	Difference
1/2k	97.9789	97.6067	97.8902	97.6968	97.7932	0.8223
k	100.0000	99.8152	99.6995	99.8934	99.8520	0.0672

250	--	GC	C	G	Avg	Difference
1/2k	98.0584	97.7148	97.9679	97.8038	97.8862	0.0931
k	100.0000	99.8235	99.7157	99.9007	99.8600	0.0079

Table 3 Experiment 2 Precision Comparison

Furthermore, the closer the dataset sizes were to each other and the higher the total record count, the lower the gap between the differences in the average precision. As the dataset size increased, the amount of added records also increased. Table 4 compares 30,000 and 240,000 record datasets; the additional records nearly doubled between dataset sizes. However, the suppression of records did not follow that pattern. In five out of the eight approaches, the 240,000 record dataset had less suppression than 30,000. These findings led the research to review the quasi-identifiers.

30k - 8584	x--	xGC	xC	xG	T--	TGC	TC	TG
Total	24814	32998	31843	30780	17769	28230	28655	25036
End QP	2847	4361	3475	3989	1438	2777	1987	2395
QP Remain %	33.1664	50.8038	40.4823	46.4702	16.7521	32.3509	23.1477	27.9007
Suppressed	7465	882	593	2760	12231	1770	1345	4964
Added	2279	3880	2436	3540				
Precision %	90.8157	87.6195	91.2931	88.1169	100	98.8234	98.1361	99.2743

240k - 15215	1/2--	1/2GC	1/2C	1/2G	T--	TGC	TC	TG
Total	239820	245049	244472	242740	224200	238466	239088	234185
End QP	12081	12960	12575	12649	8957	10672	9668	10273
QP Remain %	79.4019	85.1791	82.6487	83.1351	58.8695	70.1413	63.5426	67.5189
Suppressed	5027	660	465	2795	15800	1534	912	5815
Added	4847	5709	4937	5535				
Precision %	97.9789	97.6067	97.8902	97.6968	100	99.8152	99.6995	99.8934

Table 4 Experiment 2 Comparison 30k & 240k

After Experiment 2's analysis, the study hypothesized it was not the dataset size that affected precision, but the quasi-identifiers meeting the approach's requirements that affected precision. Table 5 compared, at each dataset size, the number of quasi-identifier pairs and amount of records that fulfilled each approach's requirements. From this review, the higher the dataset size, the more total quasi-identifier pairs, which in turn allowed for more quasi-identifiers and records to meet each approach's criteria. As the dataset size increased, the number of quasi-identifier pairs and records that met each criteria increased. For example, the 60,000 record dataset had 45.75% of quasi-identifiers and 85.85% of records meet *1/2k*-anonymity; however, as the dataset size increased to 180,000, the number of quasi-identifiers that met *1/2k*-anonymity were 71.50% and records were 94.42%. This led to the conclusion the higher the dataset size, the more quasi-identifier pairs will meet *1/2k*-anonymity requirement.

At Least Quasi-Identifier Pairs			Meet xk Percent		Difference in Percent	
	1/2	1k	1/2k	1k	1/2k	1k
30	2847	1438	33.17%	16.75%		
60	5186	2963	45.75%	26.14%	12.59%	9.39%
120	8440	5508	61.59%	40.20%	15.84%	14.06%
180	10540	7464	71.50%	50.63%	9.90%	10.44%
240	12081	8957	79.40%	58.87%	7.91%	8.24%
250	12298	9172	80.55%	60.07%	1.15%	1.20%

At Least Records			Meet xk Percent		Difference in Percent	
	1/2	1k	1/2k	1k	1/2k	1k
30	22535	17769	75.12%	59.23%		
60	51512	43951	85.85%	73.25%	10.74%	14.02%
120	112340	102350	93.62%	85.29%	7.76%	12.04%
180	173552	163009	96.42%	90.56%	2.80%	5.27%
240	234973	224200	97.91%	93.42%	1.49%	2.86%
250	245208	234433	98.08%	93.77%	0.18%	0.36%

xk QP Amount				xk Record Amount		
QPs	1/2	>=1		Recs	1/2	>=k
30	870	1438		30	2610	17769
60	1331	2963		60	3993	43951
120	1738	5508		120	5214	102350
180	1761	7464		180	5283	163009
240	1723	8957		240	5169	224200
250	1729	9172		250	5187	234433

Table 5 Experiment 2 QP & Record Comparison

Experiment 3: *xk-anonymity*

Changing the fraction can drastically change the dataset outcome. The original

proposed set of fractions were 1/4, 1/3, 1/2, 2/3, 3/4, 7/8, and 1; however, since multiple

fractions rounded to the same whole number when multiplying x of k, where $k = 5$, the

research limited the fractions to 1/4, 1/3, 1/2, 3/4, and 1. Even with a lower quantity of

different fractions, the study found many factors impacted how the fractions processed

the dataset for anonymity. This experiment found a trend in the post-processed dataset.

As the fractions increased, the difference in precision of original records and the

precision of original records including records with modifications increased. For

example, Table 6 compared *1/3k, 1/2k,* and *3/4k* for the 180,000 record dataset. The

difference in average precisions with and without modifications equaled .3743 for *1/3k,*

180k	-	GC	C	G	Avg	Diff
1/3k O	93.8957	92.5480	92.8612	93.1514	93.1141	
1/3k OwM		93.3191	93.6662	93.4799	93.4884	0.3743
1/2k O	97.2885	93.7431	94.1942	94.9544	95.0451	
1/2k OwM		96.6435	97.1861	96.7652	96.8649	1.8199
3/4k O	99.2246	93.3486	93.3586	94.9589	95.2227	
3/4k OwM		98.8505	98.9732	98.9507	98.9248	3.7021

Table 6 Experiment 3 Precision Trend

1.8199 for *1/2k*, and 3.7021 for *3/4k*. This proved precision range fluctuated depending on the fraction and the anonymization techniques. Another takeaway from this comparison, as fractions increased the importance of applying anonymization techniques to keep original records became more essential. When processing the dataset with *1/3k*-anonymity approach with generalization and cell-based suppression, the difference of the two precisions was .7711; however, in *3/4k*-anonymity approach the difference increased to 5.5019.

From Experiment 4's analysis, the research learned more about the association between dataset sizes and fractions. Table 7 documented the suppression, addition, total records, precision, and original records changes for all the approaches. In addition, it took the absolute difference of addition and suppression and the average of those differences at each fraction level. Based on these results the average lowest absolute difference varied depending on the dataset sizes. The 30,000, 240,000, and 250,000 record datasets' lowest difference average were *1/2k*, whereas, the 60,000, 120,000, and 180,000 record

datasets' were *3/4k*. This highlighted there was not always one fraction that produced the best balance of suppression and addition.

30	1/4--	1/3--	1/3GC	1/3C	1/3G	1/2--	1/2GC	1/2C	1/2G	3/4--	3/4GC	3/4C	3/4G	T--	TGC	TC	TG
Suppressed	0	4009	449	287	1589	7465	882	593	2760	10075	1269	918	3851	12231	1770	1345	4964
Added	23499	7463	10843	7986	10068	2279	3880	2436	3540	539	947	585	844				
Total	53499	33454	40394	37699	38479	24814	32998	31843	30780	20464	29678	29667	26993	17769	28230	28655	25036
Original	30000	25991	26068	20682	25991	22535	22659	22676	22535	19925	20082	20099	19925	12231	17952	17973	17769
Difference	23499	3454	10394	7699	8479	5186	2998	1843	780	9536	322	333	3007	12231	1770	1345	4964
Precision	56.0758	77.6918	72.8756	70.712	73.678	90.8157	87.6195	91.293	88.117	97.3661	95.8762	96.514	96.297	100	98.823	98.136	99.274
Avg	23499	7506.5				2701.75				3299.5				5077.5			

60	1/4--	1/3--	1/3GC	1/3C	1/3G	1/2--	1/2GC	1/2C	1/2G	3/4--	3/4GC	3/4C	3/4G	T--	TGC	TC	TG
Suppressed		3810	436	275	1742	8488	840	526	3032	12481	1401	852	4085	16049	1751	1170	5169
Added	25811	10571	13618	11069	12958	3554	5401	3716	4999	892	1338	923	1249				
Total	85811	66761	73182	70794	71216	55066	64561	63190	61967	48411	59937	60071	57164	43951	58249	58830	54831
Original	60000	56190	56300	56309	56190	51512	51728	51761	51512	47519	47763	47814	47519	43951	44247	44300	43951
Difference	25811	6761	13182	10794	11216	4934	4561	3190	1967	11589	63	71	2836	16049	1751	1170	5169
Precision	69.9211	84.1659	63.8026	84.123	71.669	93.5459	91.27	93.509	91.713	98.1574	97.2139	97.52	97.448	100	99.265	98.765	99.504
Avg	25811	10488.25				3663				3639.75				6034.75			

120	1/4--	1/3--	1/3GC	1/3C	1/3G	1/2--	1/2GC	1/2C	1/2G	3/4--	3/4GC	3/4C	3/4G	T--	TGC	TC	TG
Suppressed	0	2866	394	269	1622	7660	812	483	3051	12874	1264	709	4247	17650	1729	960	5500
Added	23325	11861	1983	12382	13402	4670	6218	4731	5988	1194	1680	1212	1585				
Total	143325	128995	133589	132113	131780	117010	125406	124248	122937	108320	120416	120503	117338	102350	118271	119040	114500
Original	120000	117134	117207	117215	117134	112340	112539	112563	112340	107126	107426	107502	107126	102350	102739	102830	102350
Difference	23325	8995	1589	12113	11780	2990	5406	4248	2937	11680	416	503	2662	17650	1729	960	5500
Precision	83.7258	90.8051	89.4663	90.533	89.806	96.0089	94.8685	95.912	95.036	98.8977	98.3143	98.505	98.465	100	99.6	99.319	99.735
Avg	23325	8619.25				3895.25				3815.25				6459.75			

180	1/4--	1/3--	1/3GC	1/3C	1/3G	1/2--	1/2GC	1/2C	1/2G	3/4--	3/4GC	3/4C	3/4G	T--	TGC	TC	TG
Suppressed	0	1956	364	282	1312	6448	724	445	3065	11731	1136	1321	1678	16991	1514	858	5602
Added	19399	11575	12795	12066	12446	4837	6038	4900	5839	1315	1744	1321	1678				
Total	199399	189619	192431	191784	191134	178389	185314	184455	182774	169584	180608	180639	177202	163009	178486	179142	174398
Original	180000	178044	178091	178093	178044	173552	173719	173746	173552	168269	168595	168642	168269	163009	163448	163535	163009
Difference	19399	9619	12431	11784	11134	1611	5314	4455	2774	10416	608	0	0	16991	1514	858	5602
Precision	90.2713	93.8957	93.3191	93.666	93.48	97.2885	96.6435	97.186	96.765	99.2246	98.8505	98.973	98.951	100	99.738	99.564	99.837
Avg	19399	11242				3538.5				2756				6241.25			

240	1/4--	1/3--	1/3GC	1/3C	1/3G	1/2--	1/2GC	1/2C	1/2G	3/4--	3/4GC	3/4C	3/4G	T--	TGC	TC	TG
Suppressed	0	1241	324	293	977	5027	660	465	2795	10196	1096	647	4412	15800	1534	912	5815
Added	15490	10526	11117	10874	10897	4847	5709	4937	5535	1401	1716	1395	1685				
Total	255490	249285	250793	250581	249920	239820	245049	244472	242740	231205	240620	240748	237273	224200	238466	239088	234185
Original	240000	238759	238779	239228	238759	234973	235110	235119	234973	229804	230059	230101	229804	224200	224660	224721	224200
Difference	15490	9285	10793	10581	9920	180	5049	4472	2740	8795	620	748	2727	15800	1534	912	5815
Precision	93.9371	95.7775	95.552	95.821	95.637	97.9789	97.6067	97.89	97.697	99.394	99.1631	99.228	99.229	100	99.815	99.7	99.893
Avg	15490	10144.75				3110.25				3222.5				6015.25			

250	1/4--	1/3--	1/3GC	1/3C	1/3G	1/2--	1/2GC	1/2C	1/2G	3/4--	3/4GC	3/4C	3/4G	T--	TGC	TC	TG
Suppressed	0	1148	325	298	913	4792	680	490	2761	9979	1072	641	4402	15567	1490	899	5875
Added	14913	10321	10840	10634	10651	4855	5679	4937	5500	1397	1710	1393	1682				
Total	264913	259173	260515	260336	259738	250063	254999	254473	252739	241418	250638	250752	247280	234433	248510	249101	244125
Original	250000	248852	248869	248869	244852	245208	245339	245349	245208	240021	240247	240292	240021	234433	234894	234936	234433
Difference	14913	9173	10515	10336	9738	63	4999	4473	2739	8582	638	752	2720	15567	1490	899	5875
Precision	94.3706	96.0177	95.8258	95.899	95.897	98.0584	97.7148	97.968	97.804	99.4213	99.2002	99.264	99.263	100	99.824	99.716	99.901
Avg	14913	9940.5				3068.5				3173				5957.75			

Table 7 Experiment 3 Dataset Sizes and Fractions Changes

Experiment 4: Different Number of Quasi-identifiers

Once the datasets reviewed the records at various quasi-identifier amounts, it became clear there was a significant connection between the quasi-identifier pairs that met *xk*-anonymity. Illustrated in Table 8, as *xk*-anonymity and quasi-identifiers increased

Meet at Least						
30000	1/4k	1/3k	1/2k	3/4k	1k	Total
Q2	62	51	42	38	34	62
Q3	127	104	85	75	68	127
Q4	1534	1188	995	880	785	1534
Q5	8584	4575	2847	1977	1438	8584

Suppressing QPs Percentage						
30000	1/4k	1/3k	1/2k	3/4k	1k	Total
Q2		17.74%	32.26%	38.71%	45.16%	62
Q3		18.11%	33.07%	40.94%	46.46%	127
Q4		22.56%	35.14%	42.63%	48.83%	1534
Q5		46.70%	66.83%	76.97%	83.25%	8584

Exact Number						
30000	1/4k	1/3k	1/2k	3/4k	1k	Total
Q2	11	9	4	4		62
Q3	23	19	10	7		127
Q4	346	193	115	95		1534
Q5	4009	1728	870	539		8584

Suppressing QPs						
30000	1/4k	1/3k	1/2k	3/4k	1k	Total
Q2		11	20	24	28	62
Q3		23	42	52	59	127
Q4		346	539	654	749	1534
Q5		4009	5737	6607	7146	8584

Record Amount				
30000	1/4k	1/3k	1/2k	3/4k
Q2	11	18	12	16
Q3	23	38	30	28
Q4	346	386	345	380
Q5	4009	3456	2610	2156

Suppressing Records					
30000	1/4k	1/3k	1/2k	3/4k	1k
Q2		11	29	41	57
Q3		23	61	91	119
Q4		346	732	1077	1457
Q5		4009	7465	10075	12231

Record Amount at Least					
30000	1/4k	1/3k	1/2k	3/4k	>=1k
Q2	30000	29989	29971	29959	29943
Q3	30000	29977	29939	29909	29881
Q4	30000	29654	29268	28923	28543
Q5	30000	25991	22535	19925	17769

Suppressing QPs Percentage					
30000	1/4k	1/3k	1/2k	3/4k	1k
Q2		0.04%	0.10%	0.14%	0.19%
Q3		0.08%	0.20%	0.30%	0.40%
Q4		1.15%	2.44%	3.59%	4.86%
Q5		13.36%	24.88%	33.58%	40.77%

Table 8 Experiment 4 30k QPs & Records Comparison

the suppression of records increased. Another observation found as the quasi-identifier pairs and *xk*-anonymity increased the amount of records that met the fraction increased. Logistically, this transpired because there was a decreased in the amount of records that fulfilled *k*-anonymity. For example, when the quasi-identifier value equaled two there were 11 records that equaled to *1/4k*, but once the quasi-identifier value increased to five there were 4009 records equaled to *1/4k*. As the quasi-identifier pairs increased, the number of records that met *k*-anonymity decreased resulting in more records meeting *xk*-anonymity. Experiment 4 illustrated the importance of understanding the effects of publishing a dataset with higher quasi-identifiers and the influence potential of processing at different fractions.

By introducing different quasi-identifiers, the research started to develop the knowledge necessary to process datasets with diverse attribute amounts. This experiment

showcased there was no clear answer to balancing noise and data loss. Each dataset individually required a review of the number of records that met the quasi-identifier pairs at each fraction to see which *xk*-anonymity holds the best balance.

Experiment 5: Inclusion of l-diversity

The fifth experiment highlighted the effects of a dataset containing sensitive values. In this experiment, each dataset processed three separate times with different diversity requirements, l = 2, 3, or 4. First, during this experiment, the study discovered when processing a dataset that included a sensitive value the fractions under the diversity requirement produced the same outcome as the fraction that equals the diversity requirement. For example, when the diversity requirement was three, *1/3k* and *1/2k* yielded the same outcome because *1/3k* would only require two records in the quasi-identifier pair set; however, that would not fulfill the diversity requirement, thus resulting in suppression. Table 9 outlined the difference in results when the 250,000 record dataset processed with a diversity requirement of two and three. From these results, the research revealed as the diversity requirement increased the number of total records published decreased.

250k – 2085 – D2	1/3--	1/2--	1/2GC	1/2C	1/2G	3/4--	3/4GC	3/4C	3/4G	T--	TGC	TC	TG
Total	250634	249939	250177	250164	250045	249324	249910	249912	249613	248799	249740	249758	249317
End QP	1965	1826	1828	1828	1826	1703	1707	1707	1703	1598	1603	1603	1598
QP Precision %	100	100	98.2796	99.1323	98.8630	100	96.0608	98.1599	97.0924	100	94.6840	97.9829	95.6886
QP Remain %	94.2446	87.5779	87.6739	87.6839	87.5779	81.6787	81.8705	81.8705	81.6787	76.6427	76.8825	76.8825	76.6427
Suppressed	134	412	200	196	326	781	216	198	508	1201	260	242	683
Added	768	351	377	360	371	105	126	110	121	0	0	0	0
Precision %	99.6936	99.8596	99.8460	99.8518	99.8508	99.9579	99.9415	99.9444	99.9492	100	99.9287	99.9810	99.9948
Original	249866	249588	249590	249590	249588	249219	249265	249224	249260	248799	248810	248810	248799
OwM	249866	249588	249800	249804	249674		249784	249802	249492	248799	249740	242683	249317

250k – 2085 – D3	1/2--	3/4--	3/4GC	3/4C	3/4G	T--	TGC	TC	TG
Total	249724	249274	249478	249487	249340	248794	249356	249362	249048
End QP	1783	1693	1694	1695	1693	1597	1600	1601	1597
QP Precision %	100	100	98.9486	99.2970	99.3545	100	97.2644	98.7662	97.9755
QP Remain %	85.5156	81.1990	81.2470	81.2950	81.1990	76.5947	76.7386	76.7866	76.5947
Suppressed	552	822	618	609	7856	1206	644	638	952
Added	276	96	96	96	96	0	0	0	0
Precision %	99.8895	99.9615	99.9582	99.9573	99.9608	100	99.9914	99.9888	99.9975
Original	249448	249178	249181	249181	249178	248794	248801	248801	248794
OwM	249448	249178	249382	249391	249244	248794	249356	249362	249048

Table 9 Experiment 5 250k, QP =4, Comparison at Diversity of 2 & 3

Experiment 6: Suppression of Sensitive Value

To counter the effects of Experiment 5, Experiment 6 suppressed just the sensitive value in the records where the quasi-identifier pair did not meet the diversity requirement. When using this anonymization technique a post-processed dataset kept more original records and more quasi-identifier pairs, but there was an increase in additional false records. Table 10 detailed the 29 different approaches in processing the 30,000 record dataset when the quasi-identifier equaled five and diversity requirement was two. This

	1/3---	1/3GCS	1/3CS	1/3GS	1/3S	1/2---	1/2GCS	1/2CS	1/2GS	1/2GC	1/2C	1/2G	1/2S
Total	32499	40394	37699	38479	33454	24764	32998	31843	30780	27464	27352	26267	24814
TE QP	4384	5941	4666	4575	4575	2837	2947	2963	2847	2898	2919	2837	2847
QP Precision %	100	78.3033	89.7480	81.9892	100	100	67.5762	85.2662	71.3713	87.7650	93.0210	90.8421	100
QP Remain %	51.0718	54.1938	54.3569	53.2968	85.2968	33.0499	34.3313	34.5177	33.1664	33.7605	34.0051	33.0499	33.1664
Suppress	4401	449	287	1589	4009	7495	882	593	2760	5125	4995	6225	7465
Add	6900	10843	7986	10068	7463	2259	3880	2436	3540	2589	2347	2492	2279
Precision %	78.7686	72.3864	77.7892	73.3125	77.6332	90.8779	86.9028	90.6647	87.7369	90.2638	90.9699	90.3919	99.8096
OwM	25599	29551	29713	28411	25991	22505	29118	29407	27240	24875	25005	23775	22535

	3/4---	3/4GCS	3/4CS	3/4GS	3/4GC	3/4C	3/4G	3/4S	T---	TGCS	TCS	TGS	TGC	TC	TG	TS
Total	20464	29678	29667	26993	25439	25467	23937	20464	17769	28230	28655	25036	24125	24400	22336	17769
TE QP	1977	2088	2104	1977	2054	2549	1977	1977	1438	2777	1572	1438	1515	1438	1438	1438
QP Precision %	100	60.7860	81.9634	64.9901	74.3126	88.3623	77.5598	100	100	55.9237	79.1142	60.0418	67.9372	86.9395	71.0826	100
QP Remain %	23.0312	24.3243	24.5107	23.0312	23.9282	24.3243	23.0312	23.0312	16.7521	18.0918	18.3131	16.7521	17.6491	18.0685	16.7521	16.7521
Suppress	10075	1269	918	3851	5337	5130	6770	10075	12231	1770	1345	4964	5875	5600	7664	12231
Add	539	947	585	844	776	5897	707	539	0	0	0	0	0	0	0	0
Precision %	97.3661	95.0057	95.8580	95.9467	96.3563	96.6997	96.7012	97.3661	100	97.9528	97.5163	98.9666	99.1759	98.6631	99.4888	100
OwM	19925	28731	29082	26149	24663	24870	23230	19925	17769	28230	28655	25036	24125	24400	22336	17769

Table 10 Experiment 6, 30k, Q5, D2 Comparison Chart

example highlighted the best anonymization approach to produce the highest original records included the anonymization technique sensitive suppression. Interestingly, in some of the dataset sizes, once the approach reached *3/4k*, both traditional and sensitive suppression methods suppressed and added the same amount of records. This showed in

some cases that even with sensitive suppression the dataset does not show improvement in post-processed results because the records with the anonymization technique applied does not meet xk-anonymity.

Experiment 5 and 6 brought new knowledge to the research that assisted in understanding how to best process datasets with a sensitive value. Experiment 5 found that it was not plausible to consider xk-anonymity where x is less than the diversity criteria. Furthermore, the lower the diversity the higher the end quasi-identifier pairs and original records but the lower the precision. From Experiment 6, sensitive suppression can aid in maintaining more records, but alone this anonymization technique may not be enough to keep more records. In most cases however sensitive suppression in combination with additional anonymization techniques could result in more published records.

Research Question Answers

After completing the six experiments, there was enough knowledge learned to answer all five research questions. These answers guided the study to create an adaptable algorithm dependent on the diverse composition of the dataset.

Can a single quasi-identifier affect the precision rate of the overall publishable dataset by more than two percent?

Experiment 1 focused on examining one quasi-identifier's impact on the overall dataset's ability to publish more records. In both *1/2k*-anonymity and traditional *k*-anonymity, having one record modified and/or suppressed did not affect the precision rate by more than two percent when comparing the highest percentages of each approach,

displayed in Table 2. However, Table 11 when comparing between the two anonymity

approaches there was a difference greater than two percent. When completing cell-based

	1/2--	1/2GC	1/2C	1/2G
OwM		97.7148	97.9679	97.8038
O	98.0584	96.2117	96.4146	97.0202

	T--	TGC	TC	TG
OwM		99.8235	99.7157	99.9007
O	100	94.5209	94.3136	96.0299

Table 11 Experiment 1 Original & Original with Modification Comparison

suppression in the *1/2k*-anonymity approach the precision for original records is

96.4146% while *k*-anonymity only comes to 94.3136%. In return, when reviewing the

precision including modified records *1/2k*-anonymity is 97.9679% and *k*-anonymity

99.7157%. Experiment 1 illustrated when combining a dataset's anonymization approach

with a single cell value suppression the overall publishable dataset has the potential to

affect the publishable amount by more than two percent.

Alternatively, in Experiment 2, there was a difference of more than two percent

precision on the *1/2k*-anonoymity approach for a dataset size of 30,000, illustrated in

Table 4.When processing this specific dataset, the end precision fluctuated significantly.

Without using any anonymization techniques, the end precision was 90.8157%, but when

using both generalization and cell-based suppression anonymization techniques the

precision changed to 87.6195%, cell-based suppression alone was 91.2931%, and

generalization was 88.1169%. These results directly confirmed, by more than two percent

precision difference, a single cell value could affect the publication quantity of a dataset.

What impact does different size datasets have on precision rate?

The primary goal of Experiment 2 was to analyze the difference in results of different size datasets. Exhibited in Table 12, averaging the eight different methods for each dataset size, the precision increased the larger the dataset: 30k – 94.2599%, 60*k*-95.9465%, 120k – 97.5599%, 180k – 98.3778%, 240k – 98.8226%, and 250k – 98.8731%. Based on the findings it was rational to state the larger the dataset size, the

30k - 8584	1/2--	1/2GC	1/2C	1/2G	T--	TGC	TC	TG	1/2k Avg	1k Avg	Both Avg
Precision %	90.8157	87.6195	91.2931	88.1169	100	98.8234	98.1361	99.2743	89.4613	99.0585	94.2599

60k - 11335	1/2--	1/2GC	1/2C	1/2G	T--	TGC	TC	TG	1/2k Avg	1k Avg	Both Avg
Precision %	93.5459	91.2700	93.5090	91.7127	100	99.2651	98.7651	99.5039	92.5094	99.3835	95.9465

120k - 13703	1/2--	1/2GC	1/2C	1/2G	T--	TGC	TC	TG	1/2k Avg	1k Avg	Both Avg
Precision %	96.0089	94.8685	95.9124	95.0355	100	99.6002	99.3191	99.7347	95.4563	99.6635	97.5599

180k - 14742	1/2--	1/2GC	1/2C	1/2G	T--	TGC	TC	TG	1/2k Avg	1k Avg	Both Avg
Precision %	97.2885	96.6435	97.1861	96.7652	100	99.7383	99.5644	99.8367	96.9708	99.7849	98.3778

240k - 15215	1/2--	1/2GC	1/2C	1/2G	T--	TGC	TC	TG	1/2k Avg	1k Avg	Both Avg
Precision %	97.9789	97.6067	97.8902	97.6968	100	99.8152	99.6995	99.8934	97.7932	99.8520	98.8226

250k - 15268	1/2--	1/2GC	1/2C	1/2G	T--	TGC	TC	TG	1/2k Avg	1k Avg	Both Avg
Precision %	98.0584	97.7148	97.9679	97.8038	100	99.8235	99.7157	99.9007	97.8862	99.8600	98.8731

Table 12 Experiment 2 Precision Comparison (30k - 250k)

higher the precision, but a deeper dive into each dataset provided more insight on this trend. As the dataset sizes increased, so did the quasi-identifier pairs, shown in Table 5. Additionally, the percentage of quasi-identifier pairs that met *k*-anonymity increased. This information changed the conclusion that a dataset size alone does not affect the precision rate; more importantly, it was the number of quasi-identifier pairs that met *k*-anonymity requirement that influenced the precision rate.

Does processing a dataset with a lower fraction level of k improve the precision rate?

The precision rate decreased when a dataset was processed at a lower fraction level of *k*. Experiments 1 and 2 used two different fractions, *1/2k* and *k*. Experiment 1's average precision had about a two percent difference, where *1/2k* approach averaged

97.8862% and *k* approach averaged 99.8600%, displayed under 250k in Table 12. This tendency continued in Experiment 2. Experiment 3 introduced *xk*-anonymity with an additional three fractions. As the fractions increased, the precision rates increased. For example, Table 7 outlined all the datasets processed at each fraction level. The 120,000 record dataset, the lowest fraction, 1/4, only had an 83.7258% precision, once processed with traditional anonymity the precision averaged 99.6635%. Based on the first three experiments, the research concluded the lower the fraction, the lower the precision. *What affects can different quantities of quasi-identifiers have on a dataset of the same size?*

On the surface, the different size quasi-identifiers affected the end publication amount of records differently in the same size dataset. In Experiment 4, the study processed datasets with four different quasi-identifier sizes (2, 3, 4, and 5). As the number of quasi-identifiers increased, the end total records decreased, as well as the precision. These findings defended Yang's et al. (2013) results when handling more quasi-identifiers the data loss increased; however, when analyzing the data further, there appeared to be a bigger reason why there was a decrease in published records. After reviewing the 60,000 record dataset, there was a trend between the quasi-identifier values that directly affected the end publication quantity, highlighted in Table 13. When there was two quasi-identifiers, 42 out of the 69 different quasi-identifier pairs met *k*-anonymity, 60.8696% of the pairs; however, once the quasi-identifier pairs reached five only 2963 of the 11335 pairs met *k*-anonymity, 26.1403%. As the total ratio of pairs decreased in meeting *k*-anonymity, the amount of publishable records also decreased. Interestingly, when the dataset had three quasi-identifiers 56.5517% met *k*-anonymity, at

QP Meet at Least						
60000	1/4k	1/3k	1/2k	3/4k	>=1k	Total
Q2	69	59	49	44	42	69
Q3	145	117	99	85	82	145
Q4	1742	1441	1265	1121	1011	1742
Q5	11335	7525	5186	3855	2963	11335

Records Meet at Least						
60000	1/4k	1/3k	1/2k	3/4k	>=1k	Total
Q2	60000	59990	59970	59955	59947	60000
Q3	60000	59972	59936	59894	59882	60000
Q4	60000	59699	59347	58915	58475	60000
Q5	60000	56190	51512	47519	43951	60000

QP Exact Number						
60000	1/4k	1/3k	1/2k	3/4k	>=1k	Total
Q2	10	10	5	2	42	69
Q3	28	18	14	3	82	145
Q4	301	176	144	110	1011	1742
Q5	3810	2339	1331	892	2963	11335

Records Exact Number						
60000	1/4k	1/3k	1/2k	3/4k	>=1k	Total
Q2	10	20	15	8	59947	60000
Q3	28	36	42	12	59882	60000
Q4	301	352	432	440	58475	60000
Q5	3810	4678	3993	3568	43951	60000

QP Percentage Kept					
60000	1/4k	1/3k	1/2k	3/4k	>=1k
Q2	100%	85.5072%	71.0145%	63.7681%	60.8696%
Q3	100%	80.6897%	68.2759%	58.6207%	56.5517%
Q4	100%	82.7210%	72.6177%	64.3513%	58.0367%
Q5	100%	66.3873%	45.7521%	34.0097%	26.1403%

Records Percentage Kept					
60000	1/4k	1/3k	1/2k	3/4k	>=1k
Q2	100%	99.9833%	99.9500%	99.9250%	99.9117%
Q3	100%	99.9533%	99.8933%	99.8233%	99.8033%
Q4	100%	99.4983%	98.9117%	98.1917%	97.4583%
Q5	100%	93.6500%	85.8533%	79.1983%	73.2517%

Table 13 Experiment 4 60k QP & Records Comparison

four quasi-identifiers there were 58.0367%, meaning proportionally more pairs met *k*-anonymity at four than three, but overall, the number of records associated with the pairs still decreased. This was why the trend in records continued to decrease as the value of quasi-identifiers increased. Considering this example and all the experiments' results, different quantities of quasi-identifiers can have a large impact on a dataset of the same size.

How does diversity value affect the overall record quantity in a single dataset?

When datasets have sensitive values, in order to begin anonymization, the quasi-identifier pairs must meet not only *xk*-anonymity, but also the diversity value set. Experiment 5 and 6 focused on *l*-diversity requirements. Experiment 5 discovered as the diversity requirements increased the number fractions available to process the dataset decreased. In Experiment 6, by suppressing the sensitive value when a quasi-identifier pair does not meet the diversity criteria, more records met eligibility for publication, but the precision decreased. Depending on the significance of the sensitive value to all other details in a record, it may or may not be worth utilizing sensitive suppression. When the research included a sensitive value, the lower the diversity requirement the higher the

total records, quasi-identifier pairs, and original records. In contrast, the higher the

diversity requirement the higher the precision. The diversity criteria greatly influenced

the anonymization process.

Algorithm

The goal of this dissertation research was to create an adaptable algorithm that

only required a data processor to provide three pieces of information. It also enabled the

ability of the processor to provide more information than necessary to customize the

results. This allowed the research to outline the process of determining the best fraction

as followed (* notes optional to the processor):

1. Dataset
2. Amount of Quasi-identifiers
3. Sensitive value (if applicable)
   - Set Diversity Criteria
   - Open to Sensitive Suppression*
4. Minimum Precision Percent*
5. Preference*
   - Original Records
   - Post-processed Precision

Figure 1 User Requirements and Options

Best Fraction

When reviewing the data collected from the six experiments many factors

highlighted as potential key elements in determining the best fraction to process a dataset:

data loss, quasi-identifier pair diversity, precision, difference in suppression and

falsification, original records, amount of records saved due to xk-anonymity, and false

data. From those candidates, many elements unfairly favored the lowest fraction available

or the traditional approach. The one key element that provided some balance was the

difference. The absolute difference between added and suppressed records allowed for

neither the lowest fraction nor traditional method to automatically be classified as the best approach; it did however, tend to favor one of the fractions in the middle because the datasets did not meet k-anonymity. The biggest influence in including the difference was it did not automatically mark one fraction the best. This element relied on the quasi-identifier pairs' records count that met the fraction conditions to align the fraction ranking against the other fraction approaches.

In the end, three elements factored in the calculation of the algorithm's recommended xk-anonymity: precision, original records, and absolute difference of addition and suppression. The first element factored into the algorithm, precision, favored a more traditional approach in anonymizing a dataset, as proven continuously throughout all six experiments. To equally counter precision, the original records published required inclusion. These two elements cancelled each other out when a data processor does not have a preference on how to anonymize their dataset. For example, Table 14 charts Experiment 3's 180,000 record dataset; as the fractions increased the precision increased and the original records including approaches with modifications decreased. The balanced element between the other two factors was the absolute difference of data falsification and suppression.

	1/4--	1/3--	1/3GC	1/3C	1/3G	1/2--	1/2GC	1/2C	1/2G
Total	199399	189619	192431	191784	191134	178389	185314	184455	182774
End QP	14742	12786	13305	13144	13089	10540	11749	11134	11393
QP Remain %	100	86.7318	90.2523	89.1602	88.7871	71.4964	79.6975	75.5257	77.2826
Suppressed	0	1956	364	282	1312	6448	724	445	3065
Added	19399	11575	12795	12066	12446	4837	6038	4900	5839
Precision %	90.2713	93.8957	93.3191	93.6662	93.4799	97.2885	96.6435	97.1861	96.7652
OwM	180000	178044	179636	179718	178688	173552	179276	179555	176935

	3/4--	3/4GC	3/4C	3/4G	T--	TGC	TC	TG
Total	169584	180608	180639	177202	163009	178486	179142	174398
End QP	8779	10437	9476	10030	7464	9350	8200	8939
QP Remain %	59.5509	70.7977	64.2789	68.0369	50.6301	63.4242	55.6234	60.6363
Suppressed	11731	1136	1321	1678	16991	1514	858	5602
Added	1315	1744	1321	1678				
Precision %	99.2246	98.8505	98.9732	98.9507	100	99.7383	99.5644	99.8367
OwM	168269	178864	179318	175524	163009	178486	179142	174398

Table 14 Experiment 3 - 180k Chart

Experiment 4 introduced processing a dataset of the same size with different quasi-identifiers. As concluded in the research questions, quasi-identifier pairs was essential when developing an adaptable algorithm. Table 15 illustrated the 30,000 record dataset processed with the quasi-identifier set to two, three, four and five. In this table, there were three factors outlined and ranked. Once the dataset reached five quasi-identifiers the approach with the least difference between suppression and addition recommended *1/3k*-anonymity. After reviewing the quasi-identifier pairs and records at each requirement, there was substantial difference in quasi-identifier pairs and records suppression when there are at least five quasi-identifiers, documented in Table 8. For example, when there were four quasi-identifiers 28,543 records met *k*-anonymity, but when there were five quasi-identifiers there were only 17,769. This was a drastic decrease in records meeting *k*-anonymity compared to four. Quasi-identifiers two, three, and four were all within 2,000 records of *k*-anonymity, whereas there was a difference larger than 10,000 once the where a total of five quasi-identifiers. This resulted in a

Q2	1/4k	1/3k	1/2k	3/4k	1k
Total	30083	30028	29983	29963	29943
Suppressed	0	11	29	41	57
Added	83	39	12	4	0
Precision %	99.7241	99.8701	99.96	99.9867	100
Original	30000	29989	29971	29959	29943
Abs Difference	83	28	17	37	57

Ranking	1/4k	1/3k	1/2k	3/4k	1k
Abs	1	4	5	3	2
Precision	1	2	3	4	5
Original	5	4	3	2	1
Total	7	10	11	9	8
Pref P	8	12	14	13	13
Pref O	12	14	14	11	9
If Equals		8	8		
Equals O		4	3		

Q3	1/4k	1/3k	1/2k	3/4k	1k
Total	30176	30061	29966	29916	29881
Suppressed	0	23	61	91	119
Added	176	84	27	7	0
Precision %	99.417	99.721	99.91	99.977	100
Original	30000	29977	29939	29909	29881
Abs Difference	176	61	34	84	119

Ranking	1/4k	1/3k	1/2k	3/4k	1k
Abs	1	4	5	3	2
Precision	1	2	3	4	5
Original	5	4	3	2	1
Total	7	10	11	9	8
Pref P	8	12	14	13	13
Pref O	12	14	14	11	9
If Equals		8	8		
Equals O		4	3		

Q4	1/4k	1/3k	1/2k	3/4k	1k
Total	32288	305588	29593	20918	28543
Suppressed	0	346	732	1077	1457
Added	2288	904	325	95	0
Precision %	92.9138	97.0417	98.9018	99.5458	100
Original	30000	29654	29268	29018	28543
Abs Difference	2288	558	407	982	1457

Ranking	1/4k	1/3k	1/2k	3/4k	1k
Abs	1	4	5	3	2
Precision	1	2	3	4	5
Original	5	4	3	2	1
Total	7	10	11	9	8
Pref P	8	12	14	13	13
Pref O	12	14	14	11	9
If Equals		8	8		
Equals O		4	3		

Q5	1/4k	1/3k	1/2k	3/4k	1k
Total	53499	33454	24814	20464	26993
Suppressed	0	4009	7465	10075	12231
Added	23499	7463	2279	539	
Precision %	56.0758	77.6918	90.8157	97.3661	162.9057
Original	30000	25991	22535	19925	19925
Abs Difference	23499	3454	5186	9536	12231

Ranking	1/4k	1/3k	1/2k	3/4k	1k
Abs	1	5	4	3	2
Precision	1	2	3	4	5
Original	5	4	3	1	1
Total	7	11	10	8	8
Pref P	8	13	13	12	13
Pref O	12	15	13	9	9
If Equals		7	7		7
Equals P		2	3		5

Table 15 Experiment 4 30k Different QPs Requirements

higher data loss. These results demonstrated depending on the quasi-identifier requirement, a dataset of the same size can have a difference balance fraction.

Besides the elements that calculated the best *xk*-anonymity, *l*-diversity requirement directly influenced the overall recommendation. The process used throughout Experiments 5 and 6 first reviewed the quasi-identifier pairs that did not meet *l*-diversity requirement. In Experiment 5, records that did not meet the diversity criteria needed suppression. Results of Experiment 5 showed if the fraction was less than the diversity requirement, those equivalent pairs' records applied suppression anyway. This

influenced the algorithm when processing a dataset that needs *l*-diversity, and excluded

sensitive suppression. If a dataset does not have a sensitive value or the data processor

elects to include sensitive suppression anonymization technique, the algorithm can

review all fraction levels. If the dataset has a sensitive value and does not opt-in for

sensitive

	No Sensitive Values Sensitives Values with Sensitive Suppression		Sensitive Values	
1	$xk = 1$ loops until $xk = k$		$xk = d$ loops until $xk = k$	
2	if $QP\# >=k$ else if $QP\# >=xk$ else	then $krecords = krecords + QP\#$ then $xkrecords = xkrecords + QP\#$ & $addrecords = addrecords++$... until $QP\# = k$ $suppressrecords = suppressrecords + QP\#$	if $l\# <d$ else if $QP\# >=k$ else if $QP\# >=xk$ else	then $suppressrecords = suppressrecords + QP\#$ then $krecords = krecords + QP\#$ then $xkrecords = xkrecords + QP\#$ & $addrecords = addrecords++$... until $QP\# = k$ $suppressrecords = suppressrecords + QP\#$
3		$originalrecords =$ $krecords + xkrecords$ $totalrecords =$ $addrecords + originalrecords$ $difference =$ $\|addrecords - suppressrecords\|$ $Precision =$ $\dfrac{((xkrecords + krecords)(Q\#))}{(totalrecords(Q\#))}$		
4		Preference = Original $((OriginalRecordsRank*2) + PrecisionRank + DifferenceRank)$ Preference = Precision $(OriginalRecordsRank + (PrecisionRank*2) + DifferenceRank)$ No Preference $(OriginalRecordsRank + PrecisionRank + (DifferenceRank))$		
5		Equals with Preference (first) $(PreferenaceRank + DifferenceRank)$ Equals with Preference (second) $PreferenceRank$ No Preference $DifferenceRank$ Equals after second preference or no preference Higher fraction of the $PreferenceRank$ or $DifferenceRank$ equal fractions		

Figure 2 Fraction Algorithm

suppression then there is a limited fraction opportunity.

Ultimately, for an adaptable algorithm to recommend the best *xk*-anonymity

approach, the dataset may go through five levels of processing. First, if there are no

sensitive values or the processor elects into sensitive value suppression the *xk* fraction

would begin reviewing the dataset when *xk* = 1. If the dataset has a sensitive value, then

xk is set to the diversity criteria value. Second, a series of if-else statements matches the

quasi-identifier pair to the equivalent statement. The sensitive value process initially

reviews the quasi-identifier pair's diversity value because if the pair does not meet the

diversity requirement it did matter if the pair meets *xk*-anonymity it requires suppression.

If the quasi-identifier pair total records (QP#) is greater than or equal to k then the QP# is added to the total records fulfilling k-anonymity (krecords). There are two actions when QP# is less than k, but greater than or equal to xk-anonymity. The difference between k and QP# is the total additional records (addrecords) that needs to be added to the dataset. Furthermore, QP# is added to the total records meeting xk-anonymity (xkrecords). When the QP# is less than xk-anonymity the records value is added to the suppression group (suppressrecords). At the third stage, once all the dataset's quasi-identifier pairs went through the second level for one fraction, there is enough collected information to calculate the original records, total records, absolute difference of added and suppressed records, and precision of the dataset. The xk fraction then increases by one and the dataset is re-processed with that criteria; steps two and three loop until $xk=k$. Step four determines the rank of each fraction level. If more than one xk-anonymity approach ranks the highest then additional processing determines the recommended fraction.

Step five provides three additional points to determine the recommended dataset fraction. As seen in Table 15, when processing the 30,000 record dataset with a preference all four quasi-identifiers requirements originally had at least two fractions equal: quasi-identifiers two, three and four occurred when they preferred original records and quasi-identifier five when it preferred precision. The first additional layer narrows down the ranking to sum of the preference and absolute difference. If still more than one fraction equals, then the algorithm selects the higher ranked fraction in the preference category. If the preference has more than one equal, the algorithm recommends the higher fraction. When there are no preferences, the absolute difference determines the recommended fraction.

By creating the three additional points in step five, the algorithm is more versatile to finding a recommended *xk*-anonymity approach. In addition, the inclusion of preferences aids in adapting to the data processors desires. Table 15 showed the changes in the recommended *xk*-anonymity approach depending on the data processor's preference and the end rankings. At two, three, and four quasi-identifier requirements with no preference or a preference of precision, the recommended approach is *1/2k*-anoymity; however, when preferring original records it decreased to *1/3k*-anonymity. Preferences aid in the adaptability of the algorithm.

Some combination of requirements may find the best approach of processing a dataset with a traditional *k*-anonymity. In these situations, two approaches are recommended, the traditional approach and the next ranked fraction. The suggestion of a second approach provides the data processor with a comparison between the traditional method and an *xk*-anonymity approach. One part of their criteria could directly impact the recommended fraction and by mentioning another approach the data processor may be more open to processing the dataset differently, or at the very least allow them to compare the difference in end publication, to see which post-processed dataset gives their audience the most value.

Precision Range

Precision range was particularly difficult to discover. Between all experiments, there were no patterns between fractions, dataset sizes, or quasi-identifiers. In order to formulate how to set a precision range the algorithm had to take into account known facts. At least two of the three anonymization techniques did not factor into the recommended *xk*-anonymity approach. In every scenario, processing a dataset without

anonymization techniques resulted in the highest amount of suppressed records. The precision range can considerably vary depending on the amount of records that meet *xk*-anonymity once applying anonymization techniques. The lowest value an anonymization technique can apply to the quasi-identifiers without removing the entire record was to suppress one of quasi-identifier's cell value. With these four details in mind, the research created a formula to define the precision range.

$$lpv = \frac{((Suppressed\#*(Quasi\text{-}identifier\#\text{-}1)) + (Original\#*Quasi\text{-}identifier\#))}{(total\# + (smaller\# \; between \; suppressed \; \& \; addition) + Difference)*Quasi\text{-}identifier\#)}$$

$$dv = \text{diversity criteria or } 1$$

$$Precision \; Range = ((lpv*100)\text{-}dv) - Precision \; Value \; Rounded \; Up$$

Equation 1 Precision Range

In order to define a precision range multiple values must be captured: the number of suppressed records, the quasi-identifier value, the original records remaining, the total records, the number of added records, the absolute difference between added and suppressed records, the precision value, and the diversity criteria (if applicable). The highest precision potential in an *xk*-anonymity approach occurs when there is zero anonymization techniques, so the highest value in the precision range equals the calculated precision value at the recommended *xk*-anonymity rate, rounded up. To define the lowest value in the precision range, facts previously discovered aided in the calculation. With the highest amount of suppression occurring in non-anonymization technique methods, the formula takes the suppressed value and multiples the records by the quasi-identifier value minus one. The deduction of one is to factor in cell-based suppression, the lowest potential value in anonymization techniques. That result is added to the result of the original records kept in *xk*-anonymity multiplied by the number of

quasi-identifiers to create the dividend part of the formula. The formula then combined three values: the lower value of records suppressed or added records, the absolute difference, and the total records. The sum of these values multiplied by the quasi-identifier value set the divisor. The lower value between addition and subtraction applied because the algorithm used an absolute difference to find the best fraction.

After finding the lowest possible value (lpv), the result is multiplied by 100 to set it to a percentage format. The last step in setting the lowest precision value depends on the dataset's sensitivity. If the dataset has a sensitive value and the processor considers sensitive suppression then the minimum diversity value must be deducted from the minimum precision. When processing a dataset with sensitive value suppression, the approach is open to all fractions; however, the precision changes depending on the diversity required, shown in Figure 3. As the diversity criteria increased the precision decreased about a half percent, so to compensate this decline the minimum precision subtracts the diversity set value. If the dataset does not have sensitive value or the processor is not open for sensitive suppression then the minimum precision is reduced by one.

60k With Sensitive Value & Sensitive Suppression

```
Best Method to process this dataset is 1/2k-anonymity
The precion range is between 91% and 94%
With a maximum data loss of 15%
```

D = 2

	GCS
Total	64561
Suppressed	840
Added	5401
Precision %	90.9797
Original	51576
Abs Difference	4561

```
Best Method to process this dataset is 1/2k-anonymity
The precion range is between 90% and 94%
With a maximum data loss of 15%
```

D = 3

	GCS
Total	64561
Suppressed	840
Added	5401
Precision %	90.5731
Original	50162
Abs Difference	4561

```
Best Method to process this dataset is 1/2k-anonymity
The precion range is between 89% and 94%
With a maximum data loss of 15%
```

D = 4

	GCS
Total	64561
Suppressed	840
Added	5401
Precision %	90.1865
Original	45170
Abs Difference	4561

```
Best Method to process this dataset is 1/2k-anonymity
The precion range is between 88% and 94%
With a maximum data loss of 15%
```

60k Without Sensitive Value

	GC
Total	64561
Suppressed	840
Added	5401
Precision %	91.2700
Original	51728
Abs Difference	4561

```
Best Method to process this dataset is 1/2k-anonymity
The precion range is between 91% and 94%
With a maximum data loss of 15%
```

Figure 3 Precision Range with Sensitive Suppression

Data Loss

$$dl = \frac{Suppress records}{(total\ ingested\ records\ (minus\ any\ title\ rows))}$$

$$data\ loss = ((dl)100, rounded\ up)$$

Equation 2 Maximum Data Loss

The last element in the algorithm measured the maximum data loss. The data loss is always highest when there is no anonymization techniques applied, shown in Table 7. To calculate the data loss with the information learned in the fraction algorithm, the data loss is the result of the recommended *xk*-anonymity's suppressed records count divided by the original dataset record total (minus a title row, if included). The result multiplied

68

by a hundred, rounded up, provided the maximum data loss percentage in the recommended approach.

Preferences

For the algorithm to be adaptable, it must provide a data processor with options. When there is a sensitive value in the dataset, the processor has an option to suppress the sensitive value if the quasi-identifier pair does not meet the diversity requirement, otherwise all records associated with the quasi-identifier pair is suppressed. Beside the sensitive value option, setting a minimum precision can influence the end recommended precision. For example, in Table 15, with a 30,000 record dataset and a quasi-identifier value of five, if the minimum precision was 98%, without a preference set, none of the *xk*-anonymity approaches would meet criteria. The algorithm goes down the ranking approaches until it found a fraction that met the minimum requirement, which would be *k*-anonymity.

A processor can also set a preference of precision or original records to help determine the result. Table 15 showed how a data processor's preference can significantly influence the recommended approach. When original records was preferred, the quasi-identifier five recommend the same fraction as it did with no preferences; however, quasi-identifiers two, three, and four needed additional processing. Quasi-identifiers two and four ended up decreasing the recommended *xk*-anonymity approach from *1/2k* to *1/3k*.

Opposite of a data processor's preference of original records, when the preference equaled precision, the quasi-identifies five had additional layers of processing required. The example in Table 15 required two additional layers of processing. The largest

transition occurred when the quasi-identifier equaled five. With the preference set to precision three fractions totaled the best, as well as the next layer of processing. In the end, when preferring precision, the recommended fraction was the traditional *k*-anonymity. These examples attests that preferences can affect the end recommended approach, tailoring this to be an adaptable algorithm.

Algorithm Validation

The research validated the algorithm with multiple datasets including the experiment sample datasets and the test sample datasets. The experiment datasets were the original six datasets used to conduct all experiments: 30,000; 60,000; 120,000; 180,000; 240,000; and 250,000. The second set of datasets first extracted a new dataset from Dataverse, *HarvardX-MITx Person-Course Academic year 2013 De-Identified* (*Harvard,* 2014). This sample had 500,000 records; based on this sample the research created two subset datasets: 20,000 and 300,000. Combining the experiment and testing sample datasets, the study reviewed a range of dataset sizes in order to validate the algorithms' recommended fraction, precision range, and maximum data loss.

Experiment Datasets

Figure 4 validated the algorithm's recommended approach, precision range, and maximum data loss precision based on Table 15's predictions. The tests conducted on the 30,000 record dataset exemplifies the algorithm's ability to predict dynamically the precision range. For example, when the quasi-identifier value was five the precision range was 9%, with a value of four the range was 3%, and a value of three or two only 1%. Quasi-identifiers two and three only displayed one approach as the experiments did not conduct any generalization or cell-based suppression at those values. These tests also

30 - Q2	1/2--	1/2GC	1/2C	1/2G
Suppressed	29			
Added	12			
Total	29983			
Original	29971			
OwM	-			
M	-			
Difference	17			
Precision	99.96			

How many Quasi-identifiers are in your dataset: 2

Do you have a Sensitive Value? (y for yes or n for no) n

Do you have a minimal precision percentage?n

Do you have a Preference on Precision or Original Records?
(p for Precision, o for Original Records, & n for No Preferences) n

Best Method to process this dataset is 1/2k-anonymity
The precion range is between 99% and 100%
With a maximum data loss of 1%

30 - Q3	1/2--	1/2GC	1/2C	1/2G
Suppressed	61			
Added	27			
Total	29966			
Original	29939			
OwM	-			
M	-			
Difference	34			
Precision	99.9099			

How many Quasi-identifiers are in your dataset: 3

Do you have a Sensitive Value? (y for yes or n for no) n

Do you have a minimal precision percentage?n

Do you have a Preference on Precision or Original Records?
(p for Precision, o for Original Records, & n for No Preferences) n

Best Method to process this dataset is 1/2k-anonymity
The precion range is between 99% and 100%
With a maximum data loss of 1%

30 - Q4	1/2--	1/2GC	1/2C	1/2G
Suppressed	732	73	63	278
Added	325	450	338	435
Total	29593	30377	30275	30157
Original	29268	29272	29272	29268
OwM	407	377	275	157
M	98.9018	98.4482	98.7737	98.5199
Difference	732	73	63	278
Precision	325	450	338	435

How many Quasi-identifiers are in your dataset: 4

Do you have a Sensitive Value? (y for yes or n for no) n

Do you have a minimal precision percentage?n

Do you have a Preference on Precision or Original Records?
(p for Precision, o for Original Records, & n for No Preferences) n

Best Method to process this dataset is 1/2k-anonymity
The precion range is between 97% and 99%
With a maximum data loss of 3%

30 - Q5	1/3--	1/3GC	1/3C	1/3G
Suppressed	4009	449	287	1589
Added	7463	10843	7986	10068
Total	33454	40394	37699	38479
Original	25991	26068	20682	25991
OwM	3454	10394	7699	8479
M	77.6918	72.8756	70.712	73.6779
Difference	3454	449	287	1589
Precision	7463	10843	7986	10068

How many Quasi-identifiers are in your dataset: 5

Do you have a Sensitive Value? (y for yes or n for no) n

Do you have a minimal precision percentage?n

Do you have a Preference on Precision or Original Records?
(p for Precision, o for Original Records, & n for No Preferences) n

Best Method to process this dataset is 1/3k-anonymity
The precion range is between 70% and 78%
With a maximum data loss of 14%

Figure 4 Validation of 30k with Different Quasi-identifier Amounts

confirmed the algorithm's ability to calculate the most data loss when processing the

dataset with the recommended *xk*-anonymity approach. When the quasi-identifier value

equaled five there were 4,009 records suppressed. By dividing the suppressed records

from the original record count of 30,000, the data loss is approximately 13.3633%.

Rounding the data loss value up to the nearest whole number, the method confirmed the

algorithm's prediction of a maximum data loss of 14%.

To validate preferences, based on Table 15, when the dataset had four quasi-identifiers with a preference of original records and when the dataset had five quasi-identifiers with a preference of precision the recommended approaches changed. Figure 5 charted Experiment 4's results when processing the dataset with quasi-identifiers four and five, as well as the algorithm's recommendations. The algorithm's approach matched Table 15's prediction, fit in the precision range, and accurately concluded the maximum data loss. When quasi-identifier five preferred precision the algorithm displayed both traditional *k*-anonymity and *3/4k*-anonymity for a data processor to consider an alternative.

Preference = Original				
30 - Q4	1/3--	1/3GC	1/3C	1/3G
Suppressed	346	34	30	161
Added	904	1147	932	1109
Total	30558	31113	30902	30948
Original	29654	29656	29656	29654
Difference	558	1113	902	948
Precision	97.0417	96.2785	96.9332	96.4016

Preference = Precision				
30 - Q5	--	GC	C	G
Suppressed	12231	1770	1345	4964
Added				
Total	17769	28230	28655	25036
Original	12231	17952	17973	17769
Difference	12231	1770	1345	4964
Precision	100	98.8234	98.1361	99.2743

Preference = Precision				
30 - Q5	3/4--	3/4GC	3/4C	3/4G
Suppressed	10075	1269	918	3851
Added	539	947	585	844
Total	20464	29678	29667	26993
Original	19925	20082	20099	19925
Difference	9536	322	333	3007
Precision	97.3661	95.8762	96.5141	96.2968

```
How many Quasi-identifiers are in your dataset: 4

Do you have a Sensitive Value? (y for yes or n for no) n

Do you have a minimal precision percentage?n

Do you have a Preference on Precision or Original Records?
(p for Precision, o for Original Records, & n for No Preferences) o

Best Method to process this dataset is 1/3k-anonymity
The precion range is between 94% and 98%
With a maximum data loss of 2%
```

```
How many Quasi-identifiers are in your dataset: 5

Do you have a Sensitive Value? (y for yes or n for no) n

Do you have a minimal precision percentage?n

Do you have a Preference on Precision or Original Records?
(p for Precision, o for Original Records, & n for No Preferences) p

Based on your outlined criteria only traditional k-anonymity would process your dataset
The precion range is between 91% and 100%
With a maximum data loss of 41%

If you are open to xk-anonymity, by allowing 3/4k-anonymity
The precion range is between 91% and 98%
With a maximum data loss of 34%
```

Figure 5 Valuation of Preference Influence

Test Datasets

After multiple tests using the experiment datasets, independent tests certified the generalizability with different datasets at different diversity levels and quasi-identifier sizes. The first test conducted on the three new datasets only included a quasi-identifier

value of four, without any sensitive values. In this test, each dataset presented a different

combination of results, highlighted in Figure 6. Then, Table 16 certified the algorithm's
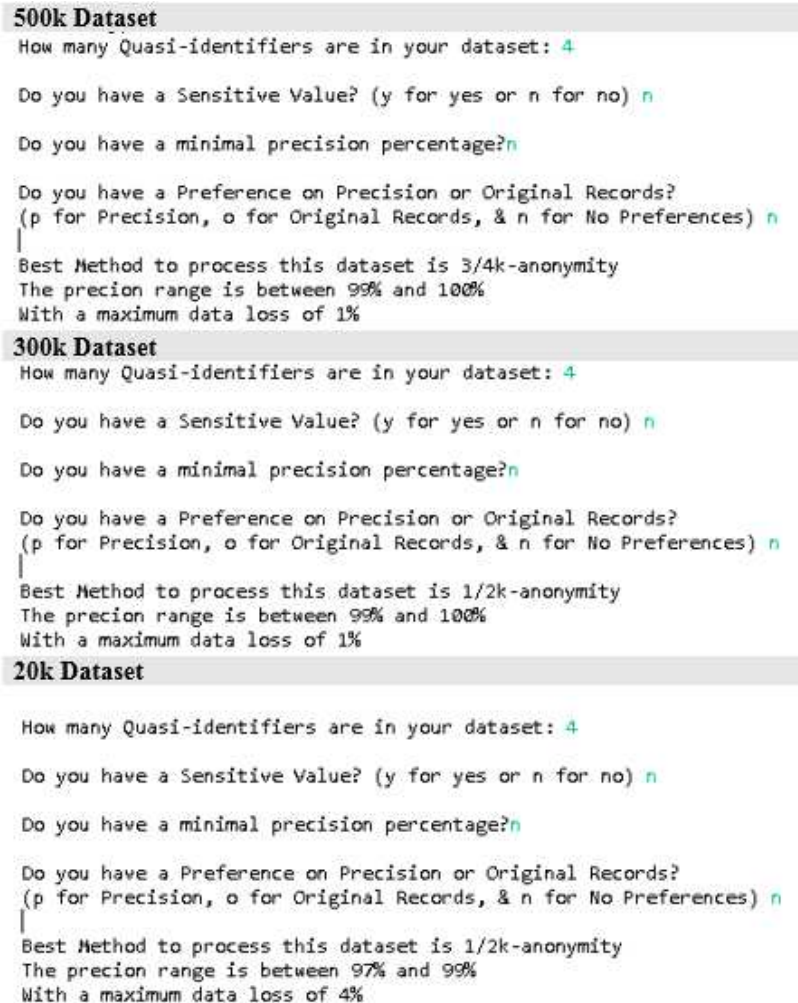
predictions.

500k Dataset
```
How many Quasi-identifiers are in your dataset: 4

Do you have a Sensitive Value? (y for yes or n for no) n

Do you have a minimal precision percentage?n

Do you have a Preference on Precision or Original Records?
(p for Precision, o for Original Records, & n for No Preferences) n

Best Method to process this dataset is 3/4k-anonymity
The precion range is between 99% and 100%
With a maximum data loss of 1%
```
300k Dataset
```
How many Quasi-identifiers are in your dataset: 4

Do you have a Sensitive Value? (y for yes or n for no) n

Do you have a minimal precision percentage?n

Do you have a Preference on Precision or Original Records?
(p for Precision, o for Original Records, & n for No Preferences) n

Best Method to process this dataset is 1/2k-anonymity
The precion range is between 99% and 100%
With a maximum data loss of 1%
```
20k Dataset
```
How many Quasi-identifiers are in your dataset: 4

Do you have a Sensitive Value? (y for yes or n for no) n

Do you have a minimal precision percentage?n

Do you have a Preference on Precision or Original Records?
(p for Precision, o for Original Records, & n for No Preferences) n

Best Method to process this dataset is 1/2k-anonymity
The precion range is between 97% and 99%
With a maximum data loss of 4%
```

Figure 6 Test Datasets with QP=4

Table 16 highlighted each of the three test datasets' recommendations based on

the end results at each fraction level, when the datasets have four quasi-identifiers. When

preferring original records in the 500,000 record dataset, the recommended

anonymization method decreased to *1/4k*. By doing this approach, zero records were

suppressed. There were only 313 records added, which meant overall many records met

k-anonymity, and there was minimal impact on the overall dataset when adding limited

500	1/4k	1/3k	1/2k	3/4k	1k
Total	500313	500298	500208	49903	499278
Suppressed	0	3	39	222	722
Added	313	301	247	125	0
Precision %	99.9374	99.9398	99.9506	99.975	100
Original	500000	499997	499961	499778	499278
Abs Difference	313	298	208	97	722

Ranking	1/4k	1/3k	1/2k	3/4k	1k
Abs	2	3	4	5	1
Precision	1	2	3	4	5
Original	5	4	3	2	1
Total	8	9	10	11	7
Pref P	9	11	13	15	12
Pref O	13	13	13	13	8
If Equals	7	7	7	7	
Equals O	5	4	3	2	

300	1/4k	1/3k	1/2k	3/4k	1k
Total	300994	300709	300074	299354	298869
Suppressed	0	57	311	743	1131
Added	994	766	385	97	0
Precision %	99.6698	99.7453	99.8717	99.9676	100
Original	300000	299943	299689	299257	298869
Abs Difference	994	709	74	646	1131

Ranking	1/4k	1/3k	1/2k	3/4k	1k
Abs	2	3	5	4	1
Precision	1	2	3	4	5
Original	5	4	3	2	1
Total	8	9	11	10	7
Pref P	9	11	14	14	12
Pref O	13	13	14	12	8
If Equals			8	8	
Equals P			3	4	

20	1/4k	1/3k	1/2k	3/4k	1k
Total	22380	20535	19530	18990	18565
Suppressed	0	369	771	1095	1435
Added	2380	904	301	85	0
Precision %	89.3655	95.5978	98.4588	99.5524	100
Original	20000	19631	19229	18905	18565
Abs Difference	2380	535	470	1010	1435

Ranking	1/4k	1/3k	1/2k	3/4k	1k
Abs	1	4	5	3	2
Precision	1	2	3	4	5
Original	5	4	3	2	1
Total	7	10	11	9	8
Pref P	8	12	14	13	13
Pref O	12	14	14	11	9
If Equals		8	8		
Equals O		4	3		

Table 16 Test Dataset Approach Comparison, QP = 4

amounts of records. Next, the 300,000 record dataset increased slightly when preferring precision, from *1/2k* to *3/4k.* Lastly, the 20,000 record dataset showed the largest different depending on the *xk*-anonymity approach used. All in all, these three datasets provided a larger range than the original experiment datasets.

The next series of tests, inspired by the dissertation's goal example, had the criteria for five quasi-identifiers and a single sensitive value with a diversity of three. These tests reviewed the datasets with and without sensitive suppression, charted in Figure 7. Each dataset had different effects when including sensitive suppression. Both the 500,000 and 20,000 record datasets fluctuated the recommended approach; however, the 300,000 record dataset remained the same. The one difference in the 300,000 record

dataset was the precision range. The anonymization technique sensitive suppression

changed the precision range.

500k Dataset		
Without Sensitive Suppression		With Sensitive Suppression
Best Method to process this dataset is 1/2k-anonymity The precion range is between 98% and 100% With a maximum data loss of 1%		Best Method to process this dataset is 3/4k-anonymity The precion range is between 97% and 100% With a maximum data loss of 1%
300k Dataset		
Without Sensitive Suppression		With Sensitive Suppression
Best Method to process this dataset is 1/2k-anonymity The precion range is between 97% and 99% With a maximum data loss of 2%		Best Method to process this dataset is 1/2k-anonymity The precion range is between 95% and 99% With a maximum data loss of 2%
20k Dataset		
Without Sensitive Suppression		With Sensitive Suppression
Best Method to process this dataset is 1/2k-anonymity The precion range is between 86% and 92% With a maximum data loss of 37%		Best Method to process this dataset is 1/3k-anonymity The precion range is between 66% and 74% With a maximum data loss of 20%

Figure 7 Test Datasets QP=5, Diversity = 3, with and without Sensitive Suppression

To confirm the results, the datasets processed the recommended approaches to

ensure accuracy of precision range and data loss. Table 17 outlined the recommended

fraction given the criteria in Figure 7. All three datasets only processed one method when

excluding sensitive suppression. This occurred because the research's data process states

any quasi-identifiers that meets xk-anonymity has $k-xk$ amount of records added. Since

the diversity criteria is three, the dataset suppresses records under three and any records

over three created data noise. All records processed at $1/2k$-anonymity satisfied the xk-

anonymity requirement, which removed any additional anonymization techniques. When

including sensitive suppression all four methods were eligible approaches. The 20,000

record dataset illustrated the influence when a dataset is or is not processed with sensitive

suppression. This dataset changed the recommended approach; additionally, there were

major changes in precision range and data loss.

Without Sensitive Suppression		With Sensitive Suppression			
500 - Q5	**1/2---**	**3/4GCS**	**3/4GS**	**3/4CS**	**3/4S**
Total	500823	500607	499421	500620	499042
Suppressed	1367	728	1929	706	2301
Added	2190	1335	1350	1326	1343
Precision %	99.5627	99.77	99.7216	99.7087	99.7262
OwM	498633	499272	498071	499294	497699
Data Loss	0.2734	0.1717	0.3958	0.1676	0.4649

Without Sensitive Suppression		With Sensitive Suppression			
300 - Q5	**1/2---**	**1/2GCS**	**1/2GS**	**1/2CS**	**1/2S**
Total	294446	304928	303107	304652	301617
Suppressed	5554	630	2307	506	3416
Added	3852	5558	5414	5158	3416
Precision %	98.7087	98.0612	98.142	98.1778	98.2959
OwM	294446	299370	297693	297768	296584
Data Loss	1.8513	0.2100	0.7690	0.1687	1.1387

Without Sensitive Suppression		With Sensitive Suppression			
20 - Q5	**1/2---**	**1/3GCS**	**1/3GS**	**1/3CS**	**1/3S**
Total	13965	28625	26805	26070	22050
Suppressed	7282	482	1563	317	3817
Added	1247	9107	8368	6387	5867
Precision %	91.0705	66.6213	67.5049	73.5061	72.6066
OwM	12718	19518	18437	19683	16183
Data Loss	36.4100	4.9300	9.8083	4.1848	19.9513

Table 17 Test Validation QP=5, Diversity=3

The 20,000 record dataset provided the largest precision range difference between the three test datasets, which made it the prime candidate to test the minimum precision percent criteria. If the dataset required a minimum precision of 90%, the results from Figure 7 would require the recommended approach to change when including sensitive suppression. Figure 8 illustrated the change from *1/2k* to *3/*4k when using sensitive suppression and a minimum precision rate of 90%. Likewise, Figure 8 confirmed there

were no changes when excluding sensitive suppress and when processing the dataset

without a diversity criterion.

20k Dataset, QP=5, Minimum 90%
Without Sensitive Value
```
Best Method to process this dataset is 3/4k-anonymity
The precion range is between 89% and 97%
With a maximum data loss of 43%
```
With Diversity = 3, No Suppression
```
Best Method to process this dataset is 1/2k-anonymity
The precion range is between 86% and 92%
With a maximum data loss of 37%
```
With Diversity = 3, Sensitive Suppression
```
Best Method to process this dataset is 3/4k-anonymity
The precion range is between 87% and 97%
With a maximum data loss of 43%
```

Figure 8 Test Dataset 20k, Minimum Precision

The study confirmed Figure 8's results by processing the dataset at the different

criteria. Figure 9 confirmed when processing the 20,000 dataset within the set criteria, the

minimum precision is at least 90%. By validating this test, it also certified the precision

range and maximum data loss accuracy.

20k Dataset, QP=5, Minimum 90%
Without Sensitive Value

Without Sensitive Value	3/4k--	3/4kGC	3/4kC	3/4kG
Total	11820	19455	19455	16940
Suppressed	8543	1272	966	3680
Added	363	727	421	620
Precision %	96.9289	95.0492	95.9224	95.6223
OwM	11457	18728	19034	16320
Data Loss	42.7150	8.1488	6.6915	19.6158

With Diversity = 3, No Suppression

D=3, No Suppression	1/2---
Total	13965
Suppressed	7282
Added	1247
Precision %	91.0705
OwM	12718
Data Loss	36.4100

With Diversity = 3, Sensitive Suppression

D=3, Sensitive Suppression	3/4kS	3/4kGCS	3/4kCS	3/4kGS
Total	11820	19455	19455	16940
Suppressed	8543	1272	966	3680
Added	363	727	421	620
Precision %	96.8875	93.4830	94.2776	94.4937
OwM	11457	18728	19034	16320
Data Loss	42.7395	9.6723	8.2915	19.6403

Figure 9 20k Minimum Precision Validation

The final series of tests verified the preferences' influence. Figure 10 tested both precision and original records preferences. Crosschecking Figure 10's results with Table 16, the research substantiated the approach recommendations based on preferences. When processing the 300,000 record dataset with the quasi-identifier value of four and with a preference of precision, the results recommended *3/4k*-anonymity. This result increased the approach from the original non-preference process. Oppositely, when processing 500,000 record dataset with the quasi-identifier value of four and a preference of original records, the result decreased to *1/4k*-anonymity.
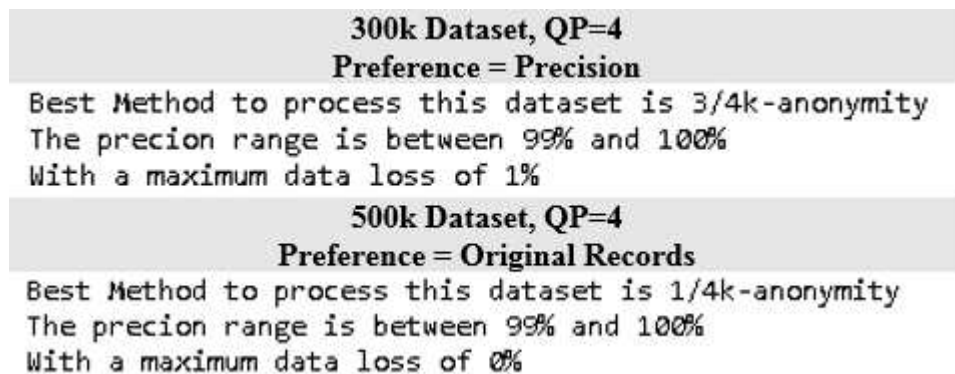
```
                        300k Dataset, QP=4
                      Preference = Precision
Best Method to process this dataset is 3/4k-anonymity
The precion range is between 99% and 100%
With a maximum data loss of 1%
                        500k Dataset, QP=4
                    Preference = Original Records
Best Method to process this dataset is 1/4k-anonymity
The precion range is between 99% and 100%
With a maximum data loss of 0%
```

Figure 10 Validation of Test Datasets Preferences

Utility

Since there is no standard for utility loss (Garfinkel, 2015), the researched used a correlation matrix to compare pre- and post-processed datasets. In comparing these matrixes, the study was able to visualize how anonymizing a dataset with the recommended *xk*-anonymity model affects the attribute relationships. The research tested the utility on the three test datasets: 20,000, 300,000, and 500,000. Three different correlation comparison occurred depending on the approach's recommendation, displayed in Table 18. In that table, the datasets had four quasi-identifiers: course ID, form post, gender, and YOB. Two of the quasi-identifiers were non-numeric values,

20k Correlation			
	Course	Post	Gender	YoB
Course	1			
Post	0.039845	1		
Gender	-0.00246	-0.0463	1	
YoB	-0.09429	-0.05568	0.782326	1

20k Correlation at *1/2k*-anonymity			
	Course	Post	Gender	YoB
Course	1			
Post	0.015629	1		
Gender	0.000172	-0.08866	1	
YoB	-0.09591	-0.10595	0.786025	1

300k Correlation			
	Coures	Post	Gender	YoB
Coures	1			
Post	0.045922	1		
Gender	-0.02056	-0.05928	1	
YoB	-0.10787	-0.06766	0.783079	1

300k Correlation at *3/4k*-anonymity			
	Course	Post	Gender	YoB
Course	1			
Post	0.045125	1		
Gender	-0.02068	-0.06398	1	
YoB	-0.10811	-0.07458	0.783399	1

500k Correlation			
	Course	Post	Gender	YoB
Course	1			
Post	0.044827	1		
Gender	-0.02046	-0.05892	1	
YoB	-0.10798	-0.06725	0.782469	1

500k Correlation at *1/4k*-anonymity			
	Course	Post	Gender	YoB
Course	1			
Post	0.0448	1		
Gender	-0.02043	-0.05784	1	
YoB	-0.10789	-0.06563	0.78238	1

Table 18 Utility Test Datasets QP = 4

which required transformation. Course ID and gender changed to numeric values by setting courses range to 0-15 and gender 0-3. After making those two adjustments the datasets had correlation matrix built comparing the four quasi-identifier relations. The 200,000 record dataset resulted in the largest impact on attribute relationship. When analyzing the difference against the other two datasets, this specific dataset had the largest data loss, highest absolute difference, and most added noise. The 300,000 record dataset compared the results when preferring precision. This dataset had less difference than the 20,000 record datasets, but more than the 500,000. The 300,000 record dataset did have data loss, because of the recommended approach, which influenced the end correlation between attributes. Finally, the 500,000 record dataset compared the dataset processed with the preference of original records. This dataset had the lowest relationship difference. These tests showed that processing datasets with these anonymization methods can impact the end attribute relationship.

Summary of Results

The result's chapter documented the significance of each element of this research. It began with a brief discussion of each experiment conducted and its worth to the general production of an adaptable algorithm. It ensured, based on the experiment's analyzed data, the study answered each research question. From those findings, the project had enough core information to design the algorithm. The third part proposed the algorithm, certifying its design with the discoveries learned. It then verified the effectiveness of the adaptability in the algorithm through testing experiment and test datasets. The last part to the chapter compared the utility difference in pre- and post-datasets. This chapter gave a thorough explanation of the all the data aspects to this project.

Chapter 5

Conclusion, Implications, Recommendations, and Summary

Foundational research provided the building blocks for this study's creation of an adaptable algorithm. The *k*-anonymity and *l*-diversity models set precedence on key attributes that potentially exposed personal information of an individual and how to protect them against such breaches (Sweeney, 2002a; Machanavajjhala et al., 2007). The *1/2k* theory found benefits compromising *k*-anonymity with false records to keep more original records (Brown, 2017). Those studies enabled this research to create a core set of experiments that assisted in the development of the algorithm.

The analysis of the six experiments conducted in the study established the basis needed to create the research's adaptable algorithm. First, discovered in Experiment 1, there was value and influence in suppressing a single quasi-identifier attribute to those pairs that did not fulfill requirements. Secondly, the balance between falsification and precision cannot be a unified approach across all datasets. All datasets are composed of a unique group of attributes and qualities that require an investigation. This examination must review the quasi-identifier pairs and the amount of records that are in each pair. In addition, if there is a sensitive value in the dataset, the pair must review the satisfaction of the diversity requirement. Each experiment held value and brought new knowledge to the study's work.

In order for this algorithm to be a success, it had to overcome two main obstacles. The algorithm had to overcome the primary obstacle of the contrasting goals of privacy and utility (Sedayao et al., 2014). It also tackled Yang's et al., (2013) findings, as the quantity of quasi-identifiers grew, the data loss grew. To handle both obstacles, the

fraction element of the algorithm ranked approaches by the lowest absolute difference of falsification and suppression. Then to ensure adaptability, the data processor could set preference to precision or original record and has the option to set a minimum precision level. By setting this balance, the research minimized the data loss at different quantities of quasi-identifiers. It also maintained privacy by using foundational *k*-anonymity and *l*-diversity methods, and it balanced the dataset's utility.

Conclusion

This algorithm maintained the objective of any privacy model. It is effective by publishing higher quantities of data without decreasing security (Aggarwal & Yu, 2008). This research created an algorithm that has the capability of processing a dataset with different quantities of quasi-identifiers. It has the ability to include a single sensitive value and determine the influence it has on the overall dataset's publication ability. The algorithm recommends the appropriate fraction of *k* with the option to customize the results with a set minimum precision percentage. It also provides the processor with a precision range of the recommended *xk*-anonymity. Overall, this study created a more versatile privacy model that can process datasets of different sizes and attributes.

Three components designed the end algorithm: finding the best fraction, providing the precision range, and stating the maximum data loss. Figure 2 outlined the algorithm's process to collect and recommend the *xk*-anonymity model. The study verified the reliability of the procedure by first predicting the recommendations displayed in Table 15 and Table 16, and it highlighted the results in Figure 4, Figure 5, and Figure 10. Equation 1 and Equation 2 stated the process to determine the precision range and maximum data loss. These equations verified in Figure 3, Figure 4, Figure 5, and 9 and

Table 17 validated their accuracy. Altogether, these three features created an adaptable

algorithm that accomplished the dissertation's goal.

The limitations set in this research may bound the usability for all data criterions.

This study limited the number quasi-identifiers and sensitive values processed in a

dataset. Additionally, the k value constantly remained five. Based on the structure of the

algorithm, the research is confident in processing a dataset with quasi-identifiers larger

than five and the diversity value larger than four; however, it did not test these

assumptions. The confidence comes from the foundation of the way the algorithm

recommends the fraction, precision range, and maximum data loss. The fraction can

process datasets with and without a sensitive value. It reviews and recommends a fraction

from the balance of the absolute difference, precision, and original record count, unless

otherwise preferred. It also knows the highest precision value and maximum data loss

based on numerous tests of processing datasets with and without anonymization

techniques. The uncertainty comes when including more than one sensitive value. The

research did not test nor did it include a process to handle datasets with more than one

sensitive value. From the results, this study would recommend processing the datasets

with the recommended algorithm as many times as it has sensitive values, each time

processing the dataset with one of the sensitive values and its corresponding diversity

requirement. Even though the research set these limitations based on the adaptability of

the components in the algorithm, it is satisfied necessary modifications can easily be

included.

Implications & Recommendations

When considering publishing datasets with personally identifying information, this research provided a new set of considerations. This study impacted the field of the privacy. Organizations have legal obligations to protect people's information (Angiuli & Waldo, 2016), but it also is necessary for organizations to publish research to improve society (Armer, 1981). With these findings, data processors have more options to consider when processing a dataset with privacy in mind. It showed each dataset is unique, and to maximize the return value on the dataset's information, it should balance the falsification and suppression. This way society can grow, and people remain protected.

Future research should dive more into the uniqueness of a dataset's attributes. It should study the ability of the algorithm to process datasets with more quasi-identifiers and sensitive values. It also could change the k value to see the effects on the recommended fractions. Finally, more research should consider an investigation on the association of quasi-identifier pairs and the records linked to them. This core development of knowledge can assist in growing the privacy field's way of processing a dataset.

Another area future studies should investigate is how to standardize the anonymization utility measurement. This research measured the dataset's utility using the correlation matrix outlined in Table 18. The correlation matrixes demonstrated that processing datasets with multiple different anonymization approaches and with preferences influence the end attribute relationship. Future research should build a tool to

assist in truly measuring the impact on quasi-identifiers and sensitive values, if applicable, against the pre- and post-processed dataset.

Privacy processed datasets that include anonymization techniques affect precision and change the relationship between its attributes (Angiuli & Waldo, 2016). This research's core goal was to aid in the development of processing diverse datasets. The study investigated four aspects in processing a dataset. It found the amount of quasi-identifier pairs and records fulfilling the xk-anonymity criteria influenced the end publishable quantity and the need to include anonymization techniques. It discovered when including sensitive values without the use of sensitive suppression, the dataset had a limitation of which xk-anonymity recommendations were eligible. The diversity of quasi-identifiers impact the overall publication potential when it is under xk-anonymity. Finally, the precision of a dataset has the potential of decreasing the lower the fraction when the ratio of pairs do not meet xk-anonymity. Each dataset is exceptional and needs to be uniquely processed.

Summary

Introduction

Chapter one presented the dissertation research. It explained how the completion of the study could benefit society. The problem this study anticipated fixing was to create an algorithm adaptable for diverse datasets. The introduction explained what questions needed answering to accomplish the dissertation's goal successfully. It also set limitations and assumptions to guide the research boundaries. The conclusion of the chapter outlined key study terms and acronyms used throughout the report.

Review of Literature

Chapter two summarized past works on the research's topic. It included a narration of foundational anonymization models with additional proposed models to build on privacy-preservation. This section also detailed different anonymization techniques used to complete de-identification and the importance of attribute classification. It provided insight on security weaknesses exposed by past models. Some models fixed vulnerabilities presented in others' models. Finally, this chapter illustrated the legal obligations organizations have when processing datasets for anonymization. This review of literature set the baseline on what the algorithm required and where it can grow.

Methodology

Chapter three outlined the study's plan of action. It broke down each experiments' process to address different aspects of the research. This section described how it approached extracting the experiment and test datasets. It detailed the Java programs built to assist in the process of the study. It also planned how to measure utility and precision for this project. To close the chapter, the section graphed the progress of each phases' completion date and what resources the research required to complete.

Result

Chapter four provided a complete summary of the experiment and algorithm results. It analyzed all experiments and how it influenced the approach in designing the adaptable algorithm. Afterwards, it answered each of the research questions. Based on the answers and findings in the experiments, this section proposed the algorithm's elements: best fraction, precision range, and data loss. It then validated the algorithm by testing the

experiment and test datasets. The chapter ended with the comparison of the pre- and post-dataset correlation matrices for measuring utility.

Conclusion

Chapter five presented the final thoughts on the overall outcome of this study. It recommended future directions for privacy-preservation research. In addition, it illustrated key takeaways, including the uniqueness of each dataset, and the significance of processing a dataset for anonymization depending on the best balance, set criteria, and overall preference of the data processor. The final chapter summarized the entire dissertation project and emphasized the prominence of data's individuality.

Reference

Aggarwal, C. and Yu, P. (2008). A general survey of privacy-preserving data-mining models and algorithms. In C. Aggarwal & P. Yu (Eds.). *Privacy-Preserving Data-mining Advances in Database Systems* (11-52). New York, NY: Springer.

Agrawal, R. and Srikant, R. (2000). Privacy-preserving data mining. *ACM Sigmond Record, 29*(2), 439-450.

Apricorn. (2016). Data encryption in education. *Apricorn,* 1-8.

Arend, C. (2017). Five essential steps for GDPR Compliance (IDC No. US42340117). *International Data Corporation*, 1-5.

Armer, P. (1981). Privacy: A survey. *ACM SIGCA Computer and Society,* 11(3), 16-18.

Angiuli, O., and Waldo, J. (2016). Statistical Tradeoffs between Generalization and Suppression in the De-identification of Large-Scale Datasets. *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC),* 589-593.

Bertino, E. (2016). Data Security and Privacy: Concepts, Approaches, and Research Directions. *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC),* 400-407. Breaux, T. and Gordon, D. (2013). What engineers should know about US security and privacy law. *IEEE Security & Privacy, 11*(3), 72-76.

Bindahman, S., Arshad, M. R. H. M., and Zakaria, N. (2017, May). Attribute based diversity model for privacy-preservation. *2017 8th International Conference on Information Technology (ICIT)*. 524-531.

Brown, E. (2017). Improving privacy-preserving methods to enhance data-mining for correlation research. *2017 SoutheastCon,* 1-4.

Canbay, P., and Sever, H. (2015). The effect of clustering on data privacy. *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA),* 277-282.

Chidambaram, S., and Srinivasagan, K. G. (2014). A combined random noise perturbation approach for multi level privacy-preservation in data mining. *2014 International Conference on Recent Trends in Information Technology,* 1-6.

Ciriani, V., Vimercati, S., Foresti, S., and Samarati, P. (2007). Microdata protection. In T. Yu and S. Jajodia (Eds.). *Secure Data Management in Decentralized Systems* (291-321)*.* New York, NY: Springer.

Dataverse. (2014) Person-course de-identification process. *MITx and HarvadX. Harvard Dataverse*, 1-12.

Dreschler, J., Bender, S., and Rässler, S. (2008). Comparing fully and partially synthetic datasets for statistical disclosure control in German IAB establishment panel. *Transactions on Data Privacy*, *1*(3), 1002-1027.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography Conference,* 265-284.

Dwork, C. (2009). The differential privacy frontier. *Theory of Cryptography Conference,* 496-502.

Dwork, C. and Roth, A. (2014). The algorithm foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science, 9*(3-4), 211-407.

Dwork, C. (2016). The definition of differential privacy. *Differential Privacy Symposium: Four Facets of Differential Privacy.*

Fung, B., Wang, K., Chen, R., & Yu, P. (2010). Privacy-preserving data publishing: A surveying of recent development. *ACM Computing Survey, 42*(4), 1-53.

Garfinkel, S. L. (2015). De-identification of personal information. *National Institute of Standards and Technology, Gaithersburg, MD, Tech. Rep. IR-8053,* 1-46.

Gkoulalas-Divanis, A., and Verykios, V. S. (2009). An overview of privacy-preserving data mining. *Crossroads, 15*(4), 23-26.

Goldman, J. (2017). Equifax breach highlights potential impact of GDPR. *eSecurity Planet*, 1-2.

Goswami, P., and Madan, S. (2017, May). Privacy-preserving data publishing and data anonymization approaches: A review. *2017 International Conference on International Computing, Communication and Automation (ICCCA)*, 139-142.

Harvard. (2014). HarvardX-MITx person-course academic year 2013 de-identified dataset, version 2.0. *Harvard Dataverse.* Retrieved from https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/26147

HHS. (2012). Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability (HIPAA) privacy rule, *US Department of Health & Human Services,* 1-32.

InfoLawGroup. (2013, Feb 12). The privacy legal implications of big data: A primer. *InfoLawGroup LLP,* 1-7.

Leichty, R. and Leong, B. (2015). De-identification & student data. *Future of Privacy Forum,*1-10.

Leonard, P. (2016, Aug). The internet of things (aka the internet of everything): What is it about + who should care. *Gilbert + Tobin,* 1-5.

Li, N. Li, T. and Venkatasubramanian, S. (2007). *T*-closeness: Privacy beyond *k*-anonymity and *l*-diversity. *IEEE 23rd International Conference on Data Engineering,* 106-115.

Li, N., Li, T., and Venkatasubramanian, S. (2010). Closeness: A new privacy measure for data publishing. *IEEE Transactions on Knowledge and Data Engineering, 22*(7), 943-956.

Li, T., Li, N., Zhang, J., and Molloy, I. (2012). Slicing: A new approach for privacy-preserving data publishing. *IEEE Transactions on Knowledge and Data Engineering, 24*(3), 561-574.

Listokin, S. (2017). Does industry self-regulation of consumer data privacy work?. *IEEE Security & Privacy, 15*(2), 92-95.

Liu, J., Luo, J., and Huang, J. (2011). Rating: Privacy-preservation for multiple attributes with different sensitivity requirements. *2011 IEEE 11th International Conference on Data-mining Workshops,* 666-673.

Lu, X., Li, Q., Qu, Z. and Hui, P. (2014) Privacy information security classification study in Internet of Things. *2014 International Conference on Identification, Information and Knowledge in the Internet of Things,* 162-165.

Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkatasubramanian, M. (2007). *l*-Diversity: Privacy beyond *k*-anonymity. *ACM Transactions on Knowledge Discovery from Data, 1*(1), 1-52.

Machanavajjhala, A. and Kifer, D. (2015). Designing statistical privacy for your data. *Communications of the ACM, 58*(3), 58-67.

Maldoff, G. (2016). Top 10 operational impacts of GDPR: Part 8 – pseudonymization. *International Association of Privacy Professionals,* 1-6.

Munir, A. B., Yasin, S. H. M., & Muhammad-Sukki, F. (2015). Big data: Big challenges to privacy and data protection. *International Journal of Social, Economics and Management Engineering, 9*(1), 355-363.

Nagendrakumar, S., Aparna, R., and Ramesh, S. (2014). A non-grouping anonymity model for preserving privacy in health data publishing. *2014 International Conference on Science Engineering and Management Research (ICSEMR),* 1-6.

OAIC. (2014). De-identification of data and information. *Office of the Australian Information Commissioner,* 1-7.

Polonetsky, J. and Tene O. (2013). Privacy and big data: Making ends meet. *Stanford Law Review Online, 66,* 25.

Rahmani, A., Amine, A., and Hamou, M. R. (2015). De-identification of textual data using immune system for privacy-preserving in big data. 2015 *IEEE International Conference on Computational Intelligence & Communication Technology*, 112-116.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics, 9*(2), 461-468.

Schwartz, P. (2013). Information privacy in the cloud. *University of Pennsylvania Law Review, 161*(6), 1623-1662.

Sedayao, J., Bhardwaj, R., and Gorade, N. (2014). Making big data, privacy and anonymization work together in the enterprise: Experiences and issues. *2014 IEEE International Congress on Big Data,* 601-607.

Sei, Y., Okumura, H., Takenouchi, T., and Ohsuga, A. (2017). Anonymization of
      sensitive quasi-identifiers for *l*-diversity and *t*-closeness. *IEEE Transactions on
      Dependable and Secure Computing,* 1-14.

Sharma, S. and Rajawat, A. (2016). A review of privacy-preserving models for multi-
      party data release framework. *ACM Symposium on Women in Research*, 165-168.

Shi, P., Xiong, L., and Fung, B. (2010). Anonymizing data with quasi-sensitive attribute
      values. *Nineteenth ACM International Conference on Information and Knowledge
      Management,* 1389-1392.

Smith, H., Dinev, T. and Xu, H. (2011). Information privacy research: An
      interdisciplinary review. *MIS Quarterly, 35*(4), 989-1016.

Sweeney, L. (2000). Simple demographics often identify people uniquely. *Data privacy
      working paper 3*, Carnegie Mellon University, 1-34.

Sweeney, L. (2002a). *K*-anonymity: A model for protecting privacy. *International
      Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* 10(5). 557-570.

Sweeney, L. (2002b). Achieving k-anonymity privacy protection using generalization and
      suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-
      Based Systems,* 10(5). 571-588.

Tene, O. & Polonetsky, J. (2013). Big data for all: Privacy and user control in the age of
      analytics. *Northwestern Journal of Technology and Intellectual Property, 11*(5),
      239-273.

Thuraisingham, B. (2015). Database security: Past, present, and future. *2015 IEEE
      International Congress on Big Data*, 772-774.

van der Aalst, W. M. P., Blichler, M., and Heinzl, A. (2017). Responsible data science. *Business & Information Systems Engineering, 59*(5), 311-313.

van der Sloot, B. and van Schendel, S. (2016) Ten questions for future regulation of big data: A comparative and empirical legal study. *Journal of Intellectual Property Information Technology and E-Commerce Law, 7*, 110.

Wong, R., Fu, A., Wang, K., and Pei, J. (2007). Minimality attack in privacy-preserving data publishing. *33rd International Conference on Very Large Data Bases*, 543-554.

Yang, G., Li, J., Zhang, S., and Yu, L. (2013). An enhanced *l*-diversity privacy-preservation. *2013 10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD),* 1115-1120.

Yaseen, S., Abbas, S. M. A., Anjum, A., Saba, T., Khan, A., Malik, S. U., ... and Bashir, A. K. (2018). Improved generalization for secure data publishing. *IEEE Access,* 27156-27165.