2016

De-Anonymization Attack Anatomy and Analysis of Ohio Nursing Workforce Data Anonymization

Jacob M. Miracle
Wright State University

DE-ANONYMIZATION ATTACK ANATOMY AND ANALYSIS OF OHIO NURSING WORKFORCE DATA ANONYMIZATION

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Cyber Security

by

JACOB M. MIRACLE
B.S.C.E., Wright State University, 2014

2016
WRIGHT STATE UNIVERSITY

WRIGHT STATE UNIVERSITY
GRADUATE SCHOOL

January 19, 2017

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY JACOB M. MIRACLE ENTITLED DE-ANONYMIZATION ATTACK ANATOMY AND ANALYSIS OF OHIO NURSING WORKFORCE DATA ANONYMIZATION BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science in Cyber Security.

Michelle Cheatham, Ph.D.
Thesis Director

Mateen Rizki, Ph.D.
Chair, Department of Computer Science and Engineering

Committee on
Final Examination

Michelle Cheatham, Ph.D.

John Gallagher, Ph.D.

Thomas Wischgoll, Ph.D.

Robert E.W. Fyffe, Ph.D.
Vice President for Research and
Dean of the Graduate School

ABSTRACT

Miracle, Jacob M. M.S.C.S., Department of Computer Science and Engineering, Wright State University, 2016. *De-Anomyization Attack Anatomy and Analysis of Ohio Nursing Workforce Data Anonymization.*

Data generalization (anonymization) is a widely misunderstood technique for preserving individual privacy in non-interactive data publishing. Easily avoidable anonymization failures are still occurring 14 years after the discovery of basic techniques to protect against them. Identities of individuals in anonymized datasets are at risk of being disclosed by cyber attackers who exploit these failures. To demonstrate the importance of proper data anonymization we present three perspectives on data anonymization. First, we examine several de-anonymization attacks to formalize the anatomy used to conduct attacks on anonymous data. Second, we examine the vulnerabilities of an anonymous nursing workforce survey to convey how this attack anatomy can still be applied to recently published anonymous datasets. We then analyze the impact proper generalization techniques have on the nursing workforce data utility. Finally, we propose the impact emerging technologies will have on de-anonymization attack sophistication and feasibility in the future.

Contents

List of Figures

List of Tables

Acknowledgment

I would like to express my thanks to my thesis advisor, Dr. Michelle Cheatham, for her dedication to her students and valuable intuition, and my mentors from the Air Force Life Cycle Management Center Engineering home office, Royce Few, Francis Erdman, and Gregory Boughton for their support throughout my graduate work. I would also like to thank Dr. John Gallagher and Dr. Thomas Wischgoll for serving on my thesis committee.

Dedicated to

the living memories of James Ned & Ola Mae Campbell.

Introduction

Publication of data containing personally identifiable information (PII) is core to academic and industry research. This data demand has resulted in legislative mandates requiring organizations to collect and publish data containing PII. Adversely there exists the need to protect individuals' identities within data to preserve individual privacy. Basic privacy preserving techniques have existed since early 2002, but adoption of these techniques by data publishers is still practically non-existent in 2016. This work addresses the importance of adopting privacy aware computing practices. The importance of privacy aware computing practices are reinforced in three ways: by understanding how de-anonymization attacks are conducted, by exploring how supposedly anonymous datasets are still vulnerable to these attacks, and by considering how emerging technologies will increase the sophistication of these attacks.

1.1 Data Accessibility versus Data Privacy

The World Wide Web provides a medium for the sharing of information, specifically documents. Hypertext links have been the connecting link on the Web allowing search engines to index documents, analyze structure of links between documents, and infer relevance to user searches [5]. While this linking was once nearly exclusive to documents, the Web has since evolved into a hybrid information space linking both documents and data. Linked Data removes limitations inherited through the use of document file formats (for example,

PDF, DOC, TXT) and accessibility limitations of privately owned servers requiring remote connection (by ssh or similar methods) to gather data. Removal of accessibility limitations provides a medium which facilitates Big Data analytics.

Linked Data uses the Web to create typed links between data from different sources facilitating several features: machine-readability, explicitly defined meaning, and relationships to other external datasets [5]. These features enable the development of applications that are capable of leveraging Linked Data to fulfill societal needs across a myriad of domains. For example, IBM's Watson for Oncology can utilize linked data to provide clinicians with evidence-based treatment options based on a patient's medical record, clinical expertise, external research and data [19]. Watson for Oncology can recommend treatment options and provide the supporting evidence for the recommendations. Linked Data provides the infrastructure necessary for the evolution of the next generation of intelligent applications.

With the establishment of any technology comes implications that must be understood and mitigated to ensure its benefits. Linked Data provides a way to infer data relationships that has never previously existed. These inferences are overarching, including not just desirable research applications, but also the ability to gather data about people on a personal level, compromising individual privacy. Therefore a need exists to grow privacy technologies in tandem with Linked Data to preserve the integrity of data while protecting the anonymity of individuals.

Privacy as a protection mechanism shields individuals from aggressive actions which threaten personal finance, health, property, and dignity. People maintain privacy to protect themselves against threats such as identity theft, health based discrimination, and embarrassment from the release of intimate life details. Individuals have the right to compromise their own privacy by disclosing personal details. Privacy is often compromised within the context of obtaining a needed or desirable service.

Service providers are granted some limited rights to the private information of their

clientele and expected to protect that data under the law set by the organization's governance. Over time service providers accumulate troves of data capable of being analyzed to better business practices, academic research, and support technological breakthroughs. Often service providers do not have internal divisions capable of performing big data analysis required to further their own business practices or are required by law to release data troves. Therefore the respective organizations require a means to share these datasets while protecting the privacy of their clients.

Linked Data has provided a means for sharing data held by organizations, but is not designed to preserve the privacy of individuals represented in the data. Linked Data which is not privacy considerate can result in powerful malicious applications which autonomously gather and infer private information about individual persons on a world wide scale. Realizing dangers associated with publication of PII has given rise to research and development of anonymization techniques which seek to protect the privacy of persons in big data while preserving data integrity. However advances in data anonymization methods are subject to increasing complexity because dataset anonymization intrinsically degrades the usefulness of the data.

1.2 Anonymization Basics

1.2.1 Attribute Types

There are four key attribute types in datasets containing personal information: Explicit identifiers, quasi-identifiers, non-sensitive attributes, and sensitive attributes.

- Explicit Identifiers
  - Name
  - National Identification Number (SSN, INSEE Code, NIE Number)

Explicit identifiers include attributes which uniquely disclose the identity of an individual (see above for a non-exhaustive list of examples). Explicit identifiers directly reveal the identity of an individual in a dataset and should never be used in anonymous datasets.

- Quasi-Identifiers
  - Age
  - Gender
  - Postal code (Eircode, PIN, PLZ, ZIP)

Quasi-identifiers (QIDs) are attributes that do not uniquely identify individuals, but are closely related to an individual. Combination of a QID with other QIDs can create a uniquely identifying set. QIDs can be used in anonymous datasets, but must be generalized (see section 1.2.2) to protect identities.

- Non-Sensitive Attributes
  - Profession
  - Nationality
  - Title (Dr., Mr., Ms.)

Non-sensitive attributes represent information that is already public knowledge or widely shared within a population. Non-sensitive attributes are represented in public records such as criminal records and voter registrations. The use of non-sensitive attributes in an anonymized dataset is subjective, under the correct circumstances non-sensitive attributes might qualify as QIDs.

- Sensitive Attributes
  - Genetic Data
  - Medical (Symptoms, Diagnoses, Medications)
  - Salary
  - Survey Choices

Sensitive attributes do not contain identifying information, but linking them to identified individuals compromises the personal privacy of an individual. Sensitive attributes are normally the attribute of research interest in a published anonymized dataset.

1.2.2 Anonymization Operations

Anonymization is implemented through anonymization operations: Generalization, Suppression, Anatomization, and Perturbation. Generalization reduces the granularity of attributes within a dataset where an attribute denotes a field or semantic category of information in a set [27]. For example, given an age attribute of a particular record with a value of 34, this attribute can be generalized to a predetermined range of 5 years providing a generalized value of [30-35). This generalization helps reduce the viability of identifying persons (protecting their privacy) of a specific age between 30 and 34, but reduces the granularity of the data. Complete generalization of an attribute is known as suppression, resulting in an attribute being omitted from a dataset.

Anatomization uses substitute values to conceal QID attributes. QIDs in a data set are mapped to substitute values. These substitute values are supplied when the data set is published instead of the original QID values. These anatomized records allow organizations external to the publisher operate on data without disclosure of the original values of the anatomized QID fields. Once external organizations complete research on the data set the publisher can recover the original QID values from the anatomized values allowing them

to recognize additional outcomes from the analyses.

Data perturbation replaces original data values with synthetic values which aim to preserve chosen statistical properties such as mean and standard deviation. A host of perturbation techniques exist which add noise, swap data, and substitute synthetic data. Data perturbation can be viewed as orthogonal to generalization techniques and therefore consideration of perturbation techniques is outside of the scope of this work.

The aforementioned operations provide methods for performing anonymization on datasets, but do not provide a quantification to assess the total anonymity of a dataset. There are properties which can be possessed by an anonymized dataset which detail its anonymity: k-anonymity, ℓ-diversity, and t-closeness. The achievement of such properties in an anonymous dataset is acquired by applying anonymization algorithms which perform anonymization operations on the data.

1.2.3 k-anonymity

An anonymized dataset has the property k-anonymity if the information for each person in the data cannot be distinguished from at least k-1 individuals in the dataset. k-anonymity provides protection against the linking of quasi-identifiable information to external data. Implementation of k-anonymity is achieved through the generalization of quasi-identifying attributes until the number of indistinguishable records matches the desired value of k. The goal of implementing k-anonymity is to achieve k-anonymity with minimal change to the chosen information metrics.

The above table details the denomination and GPA of various majors in a hypothetical Christian university. This dataset has a k=1 because we can distinguish individual records and therefore does not protect the privacy of some individuals. By generalizing the quasi-identifying attributes: Major, Sex, Age, and Denomination we can achieve a k-anonymous table for a k greater than 1.

Table 1.1: Example of an identified table detailing religious preferences and academic performance of students.

Major	Sex	Age	Denomination	GPA
Accounting	F	22	Southern Baptist	3.8
Management	M	24	Free Will Baptist	3.4
Management	F	21	Eastern Orthodox	3.6
Marketing	F	23	Eastern Orthodox	3.0
Biology	F	25	Oriental Orthodox	3.6
Chemistry	M	27	Oriental Orthodox	3.1
Physics	F	25	Roman Catholic	3.9
Physics	M	28	Angelican Catholic	2.8
English Edu.	F	34	Pentecostal	4.0
History Edu.	F	31	Pentecostal	3.4
Math Edu.	F	32	Eastern Orthodox	3.5
Special Edu.	M	30	Eastern Orthodox	2.9

This k-anonymous table has a k=2 because k-1 (1) individuals cannot be distinguished from each other. Complete generalization (suppression) of Sex is required to achieve a k=2 because of records such as the male Management major in Table 1. Without suppressing the Sex attribute the table would retain a k=1 because a single male Management major's record (among others) would be distinguishable from the other records. This protects the male Management major's GPA from being identified by attackers who know his record is in the dataset.

1.2.4 ℓ-diversity

The ℓ-diversity property is an extension of k-anonymity which deals with the weaknesses in k-anonymity, namely the threat of attacker background knowledge and homogeneity attacks. Homogeneity attacks leverage cases of k-anonymous data where all values for a sensitive attribute within a set of indistinguishable records are identical. Therefore even though the data is k-anonymous a sensitive attribute only consists of one value for a set of records making it possible to determine the sensitive value of all records in the set.

Table 1.2: Example of an anonymized table 1.1 with k=2.

Major	Sex	Age	Denomination	GPA
Business	*	[20-25)	Baptist	3.8
Business	*	[20-25)	Baptist	3.4
Business	*	[20-25)	Eastern Orthodox	3.6
Business	*	[20-25)	Eastern Orthodox	3.0
Science	*	[25-30)	Oriental Orthodox	3.6
Science	*	[25-30)	Oriental Orthodox	3.1
Science	*	[25-30)	Western Catholic	3.9
Science	*	[25-30)	Western Catholic	2.8
Education	*	[30-35)	Pentecostal	3.0
Education	*	[30-35)	Pentecostal	3.4
Education	*	[30-35)	Eastern Orthodox	3.5
Education	*	[30-35)	Eastern Orthodox	2.9

Background knowledge attacks leverage external knowledge to reduce the set of possible values for a sensitive attribute.

Achieving ℓ-diversity requires organizing anonymized data into blocks and increasing the diversity (ℓ) of sensitive attributes within each block. If a block contains too many instances of the same value for sensitive attributes it is possible to infer that a person within the given block probably has this sensitive value. $\mathcal{L}$-diversity addresses the weaknesses of k-anonymity and provides a way to protect data against attackers with background knowledge, but doesn't guarantee data utility. [18] demonstrates that background knowledge about low heart attack rates in Japanese patients could be leveraged against a dataset, narrowing the range of values for a patient's disease (a sensitive attribute).

Reorganizing the table of student denomination and GPAs into blocks sorted by the QID attribute denomination we can determine the ℓ-diversity of the dataset. In this case the dataset has an ℓ=2 because each block has at least 2 different QID value combinations for the set of sensitive attributes. While this dataset is ℓ-diverse, k-anonymous datasets are not ℓ-diverse by default.

Table 1.3: Example of an anonymized 1.1 with ℓ=2.

Major	Sex	Age	Denomination	GPA
Business	*	[20-25)	Baptist	3.8
Business	*	[20-25)	Baptist	3.4
Business	*	[20-25)	Eastern Orthodox	3.6
Business	*	[20-25)	Eastern Orthodox	3.0
Science	*	[25-30)	Oriental Orthodox	3.6
Science	*	[25-30)	Oriental Orthodox	3.1
Science	*	[25-30)	Western Catholic	3.9
Science	*	[25-30)	Western Catholic	2.8
Education	*	[30-35)	Pentecostal	3.0
Education	*	[30-35)	Pentecostal	3.4
Education	*	[30-35)	Eastern Orthodox	3.5
Education	*	[30-35)	Eastern Orthodox	2.9

1.2.5 t-closeness

The t-closeness property is an extension of ℓ-diversity which considers the distribution of values for sensitive attributes. As explained by [17], ℓ-diversity ensures "diversity" of sensitive values in each block, but it does not take into account the semantic closeness or statistical significance of those values. This makes ℓ-diverse datasets vulnerable to semantic similarity and skewness attacks.

Achieving t-closeness requires examination of each equivalence class in a dataset. Each equivalence class is considered t-close if the distance between the sensitive attribute's (GPA) distribution and the overall dataset's distribution for the same attribute are less than an established threshold (t). The value of t is subjective and up to data publishers to decide what distance from the overall data is acceptable.

The overall average GPA of the dataset presented in Table 1.4 is 3.3. Let's assume the publisher has established a t=0.10, this means it is acceptable for the average GPA of an equivalence class to be 10% different from the overall average of the dataset. By using the 10% threshold the publishers establish AVG GPA for an equivalence class must fall between [2.96-3.63] to be compliant with their anonymization requirements.

Achieving a meaningful t-closeness for small datasets is sometimes difficult because there may not be enough records to retain data use-ability after generalization of the set to meet t-closeness requirements.

Table 1.4: Example of statistically t-close table based on mean GPA=3.3 with threshold t=0.10. Distance allowed: 3.3±0.33, [2.96-3.63].

Major	Sex	Age	Denomination	GPA	AVG
Business	*	[20-25)	Baptist	3.8	3.6
Business	*	[20-25)	Baptist	3.4	
Business	*	[20-25)	Eastern Orthodox	3.6	3.3
Business	*	[20-25)	Eastern Orthodox	3.0	
Science	*	[25-30)	Oriental Orthodox	3.6	3.33
Science	*	[25-30)	Oriental Orthodox	3.1	
Science	*	[25-30)	Western Catholic	3.9	3.4
Science	*	[25-30)	Western Catholic	2.8	
Education	*	[30-35)	Pentecostal	3.0	3.2
Education	*	[30-35)	Pentecostal	3.4	
Education	*	[30-35)	Eastern Orthodox	3.5	3.2
Education	*	[30-35)	Eastern Orthodox	2.9	

1.3 Contributions and Scope

This chapter has introduced the importance of personal privacy, the need to publish sensitive data anonymously, and basic techniques used to perform anonymization. The remainder of this work will elaborate on de-anonymization attacks and dataset generalize-ability.

First, we conduct a survey that explores the past of anonymization failures. These anonymization failures share many common features. By examining these shared features we define a formalized anatomy of de-anonymization attacks.

Second, we conduct a comprehensive analysis of the limitations of the anonymization done to a dataset containing nursing workforce data for the state of Ohio. This analysis includes several components which highlight dataset anonymization failures, dataset

generalize-ability, and utility of properly anonymized sets.

Third, this work discusses the core features of Semantic Web technologies that could potentially be leveraged to conduct de-anonymization attacks on anonymous data more quickly and effectively in the future.

Survey of Attacks on Anonymous Data

This survey is focused on incidents in which a publishing organization claimed to have released an anonymized dataset that was later compromised, resulting in the revelation of at least some of the identities of individuals with records in the dataset. Three of these incidents were highly publicized and have been the subject of several news stories and academic papers: America Online (AOL) search data [4], Massachusetts Group Insurance Commission medical records [27], and the Netflix Prize movie rating dataset [21]. We also examine a fourth incident in which multiple individual actors provide the tools necessary to compromise an anonymous NYC Taxi Trip dataset, revealing religious practices of some New York cab drivers.

2.1 AOL Search Data

In August 2006, AOL released detailed search logs of AOL users intended for research purposes. The dataset included the following attribute set:

- User ID
- Query
- Query Time
- Clicked Rank
- Destination Domain URL

The dataset was ostensibly anonymized [21] through the use of anatomization. AOL replaced explicit identifiers (names) of its users with a numeric User ID. Shortly after release AOL removed the dataset after identifying the dataset was vulnerable. However, the dataset had already been acquired and mirrored across the web by many parties.

The New York Times specifically identified a user and published an article about her within 5 days of the data release [4]. As explained by [4] the Query attribute in the dataset often contained PII of users who performed searches of their own identifying information. User No. 4417749's search logs contained queries such as "landscapers in Lilburn, Ga,", searches for several individuals with the last name Arnold, and "homes sold in shadow lake subdivision gwinnet county georgia" [4]. The New York Times was able to use these values as QIDs to identify User No. 4417749 as Thelma Arnold, a 62-year-old widows living in Lilburn, GA [4] through the use of public records.

2.2 Massachusetts Group Insurance Commissions

In 1996 The National Association of Health Data Organizations (NAHDO) reported that 37 states in the United States required hospitals to publish anonymized information about patient conditions and outcomes. The intent of these laws is to allow patients to make informed decisions about their health care and to encourage hospitals to find and treat the

root causes of illnesses rather than only the symptoms. In [27] Sweeney describes how she was able to use this data, together with voter registration rolls, to identify the medical records of William Weld, a former governor of Massachusetts.

In Massachusetts, the Group Insurance Commission (GIC) collected patient data with over 100 attributes for approximately 135,000 state employees. GIC distributed ostensibly anonymous copies of this data to researchers and industry. The anonymization process removed explicit identifiers of individuals such as name and social security number, but left other information, including the ZIP code, birth date, and gender of patients, in place and unaltered.

To compromise the dataset Sweeney purchased the voter registration list for Cambridge Massachusetts. The voter registration list contained the same quasi-identifying attributes as the anonymous dataset (ZIP code, birth year, and gender) in addition to the name, address, date registered, party affiliation, and date last voted of registered voters. Through the shared fields present in two data sets, Sweeney was able to link individuals and thereby determine the diagnosis, procedures and medications of particular people. In particular, [27] explains that Governor Weld lived in Cambridge Massachusetts and according to the the voter registration list six people shared his birth date; only three of them were men; and he was the only one in his ZIP code. Sweeney used this example to help form the basis for presenting the k-anonymity protection model.

2.3 Netflix Prize

From 2006 to 2009 Netflix ran an annual open competition for the best collaborative filtering algorithm.[1] The goal of the Netflix Challenge was to improve the "you might also like to watch..." feature such that subscribers watched one of the suggested movies 10% more often than they did under the current recommendation algorithm, Cinematch. To facilitate

[1] http://netflixprize.com/rules.html

the competition Netflix released a dataset containing anonymized movie ratings of 500,000 Netflix subscribers. All customer identifying information was removed from the dataset, so that for each subcriber the dataset contained only the titles of movies they watched, the date and time they were viewed, and the ratings the subscriber gave them.

Almost as soon as the contest began, a large group of researchers took up the challenge to improve upon the Cinematch recommendation algorithm, while another, also fairly large, group of researchers set about finding the identities of the Netflix subscribers in the dataset. An example of this latter group is Narayanan and Shmatikov, who described their work in [21]. The pair noted that the Internet Movie Database (IMDb) contains attributes in common with the Netflix dataset, namely, movie rating and date of rating. In addition, IMDb contains a person's username and any associated profile information they have provided.

The researchers assumed that the same person would give a movie the same rating on both Netflix and IMDb. Additionally, they reasoned that these ratings would not be very far apart in time. Based on these assumptions, they created a metric that compared a Netflix record to an IMDb one based on movie title and rating (for which an exact match was required) and timestamp (which could be off by a specified amount, which was a parameter to the de-anonymization algorithm; in [21], the values 3, 14 and ∞ days were used). This method is particularly effective if the subscriber has watched any movies that are not frequently reviewed on IMDb. Their results show that they are able to uniquely identify 99% of subscribers who have reviewed at least eight movies, six of which are outside of the top 500 rated movies. In fact, just two ratings are enough to identify 68% of subscribers.

The de-anonymization of the Netflix Challenge data is notable in that it was still feasible despite containing *none* of the traditional information about people that is used to link them to another dataset, such as ZIP code or age. What makes this possible is the high-dimensionality of the data – an individual's viewing timeline and ratings are in effect almost like a fingerprint. The vulnerability of such high-dimensional data has been noted

by several researchers, including [1].

In 2009, a research group called "BellKors Pragmatic Chaos" was awarded the one million dollar price. The downside is that Netflix faced a class action lawsuit filed by its subscribers and was criticized by the Federal Trade Commission over privacy concerns.

2.4 FOIL NYC Taxi Trip Data

The Freedom of Information Law (FOIL) of New York requires state agencies to supply requested information unless otherwise specified. In 2014, Chris Whong received a dataset containing historical taxi trip and fare logs of New York City via a Freedom of Information request [23]. To facilitate examination of the data, Whong developed a visualization tool which graphically shows individual taxi activity over a 24 hour period. Whong published both the data he received and his tool on the Web. Noah Deneau discovered Whong's visualization tool in conjunction with a study showing the most common name for taxi drivers in New York City to be Mohammad. Deneau was able to use Whong's tool to graphically identify four examples of drivers who have low activity during designated Muslim prayer times, possibly inferring their religious practices and compromising their privacy.

The dataset released by the NYC Taxi and Limousine Commission was ostensibly anonymized, concealing the driver's license number and taxi number. However Vijay Pandurangan [23] identified that these fields were cryptographically hashed via MD5. MD5 is a one-way function that always generates the same hash value for a given input. Pandurangen examined the structure of NYC taxi medallion numbers, finding there are a possible 18,954,000 medallion numbers. With this knowledge Pandurangen calculated all possible MD5 hashes, therefore linking the hashed identifiers to their actual values.

New York City maintains public listings of named drivers and their corresponding license numbers. The combination of the above linking and observations directly links the religious preferences of Muslim drivers to named individuals. Additionally any other

sensitive information that can be estimated by examination of driver habits (gross salary, approximate residence) can be linked to named individuals in the datasets.

Analysis and Application to Ohio Nursing Workforce Data

3.1 Analysis

3.1.1 Anatomy of an De-anonymization Attack

Reflecting on these anonymization failures led us to identify five key steps in conducting a de-anonymization attack:

1. **Identify and acquire the target dataset.** Datasets may be targeted because they contain sensitive information desirable to an attacker such as salaries, employment history, or disease information. Attackers may seek to de-anonymize such sensitive information for future malicious or criminal ventures (for example, selling de-anonymized data). However other attackers may seek anonymous datasets to practice their de-anonymization skills, embarrass the publishing party, or simply for bragging rights.

2. **Examine the target dataset for potentially identifying attributes.** Attackers must scan datasets for attributes they can leverage against other datasets. Often these scans are performed by the attacker through a manual inspection of dataset attributes. An attacker may have a particular set of attributes in mind or may discover attributes as

they scan the dataset that might be common in other data available. Of particular interest are fields for which many individuals have unique or nearly-unique values.

3. **Identify and acquire auxiliary information (external knowledge or a secondary dataset).** This data should contain the potentially identifying attributes and a superset of the individuals in the target dataset. Often auxiliary data consists of public data such as voter registries, but can also consist of other data sources such as social media profiles and privileged private data sources the attacker has stolen.

4. **Construct method(s) for establishing links between the target dataset and the auxiliary information.** The method for linking datasets is often case specific. The method for linking in one attack may be as simple as joining records with a common attribute value. Attacks against large sparse datasets (such as the Netflix Prize data) may require construction of additional data items (profiles) and algorithmic comparisons of those profiles against multiple records in auxiliary sets to establish a link.

5. **Generate a de-anonymized dataset composed of the target data and auxiliary information.** After de-anonymizing a dataset an attacker can generate new datasets often containing many more attributes than the anonymous set originally contained. The attacker not only obtains names of individuals in the anonymous dataset, but also the additional data tied to those individuals in the auxiliary datasets. This gives attackers a more diverse picture of an individual, supplying supplemental personal information the attacker may not have initially sought.

3.1.2 Failure to Utilize Anonymization Basics

Another outcome in the survey of anonymization failures presented in the previous chapter is the noteworthy and somewhat alarming observation that in each of the anonymization failures examined, the data providers failed to publish their datasets such that they met even the basic criteria for anonymity. The Massachusetts GIC anonymization failure pre-dates

much of the existing body of anonymization research, so they at least have something of an excuse. However, each of the other organizations published data containing documented vulnerabilities. The AOL Search Data was published after [27] warned against the use of QID attributes and the introduction of k-anonymity (the idea that each individual in a dataset should be indistinguishable from at least k others), yet the company failed to screen the search queries for such attributes. The Netflix Prize datasets were published years after [1] identified the breakdown of traditional anonymization techniques in large datasets. And the NYC Taxi and Limousine Commission could have protected its data by performing a proper anatomization of license numbers rather than using a deterministic hash that provides no value.

Of course, while it can be argued these anonymization failures would not have occurred if publishers had adhered to the best available anonymization practices, the notion is somewhat tenuous. In the last decade we have witnessed the breakdown of established anonymization principles such as k-anonymity, ℓ-diversity, and t-closeness [27, 18, 17] within highly dimensional datasets [1, 21]. Beyond this, the proliferation of linked data and expansion of the Semantic Web will increase overall data dimensionality and thereby degrade the effectiveness of current anonymization techniques. In short, technological advances are driving the need to adapt anonymization methods for the age of the Semantic Web even as data providers are lackadaisical in adopting current best practices.

3.2 Application: Ohio Nurse Workforce Data

This section applies the five steps in a de-anonymiztion attack to an actual dataset. The data provider was notified of this vulnerability more than six months ago, and we will not reveal any particular individual's information.

3.2.1 Identify and Acquire the Target Dataset

The Ohio Board of Nursing conducts a mandatory annual workforce survey to obtain data on its workforce composition. As stated on the Ohio Board of Nursing's website, "The Board is proud that the data will assist with the workforce planning initiatives of government and private industry." The nursing datasets are available on the Ohio State Board of Nursing website: http://www.nursing.ohio.gov/Workforce.htm. The anonymized raw data obtained from these annual surveys are available in CSV and XSLX file formats since 2013.

3.2.2 Examine the Target Dataset for Potentially Identifying Attributes

The Ohio 2015 Nursing Workforce Data Package contains 43 fields (see Appendix A for a full list) which include quasi-identifiable information such as race, gender, year of birth, and postal ZIP code. The data set also includes very specific details about an individual such as any military affiliation (self or spouse), if they can speak a foreign language, and the title of their working position. The data set contains 183,188 records of nurses who took the survey in 2015. Select fields from this dataset are available in other explicitly identified public datasets. By further examining this dataset with the auxiliary data described in the next section 3.2.3 we are able to identify three QID fields that serve a record linkage attack: Year of Birth, Initial Year Licensed, and Postal Zip Code. While this attack did not rely on other fields commonly used in record linkage attacks to create record links, there were isolated instances within this attack where race or gender could have been used to reason about a person's identity when one anonymous record mapped to multiple potential explicitly identified records.

3.2.3 Identify and Acquire Auxiliary Information

There are many potential auxiliary data sources which share QID attributes with the target data set. A record linkage attack may require a combination of multiple auxiliary sources, and this increases the complexity of the attack. High complexity may deter attackers from attacking a target by providing a lower return on investment (ROI). However, compromising the 2015 nursing data has very low complexity. To identify individuals in the target dataset we only need to leverage one auxiliary dataset: the Ohio License Center Website nursing database. Ironically the state provides both the anonymous data set and the identified set needed to compromise its anonymity.

The relevant dataset is available through the Ohio License Center website (https://elicense.ohio.gov/oh_verifylicense). This website contains information about all types of medical license holders in Ohio, from acupuncturists, to physical therapists, to registered nurses. The website allows users to verify a nursing license by entering the nurse's license number or the name of the individual. The website also allows access to a list of all registered nurses in the state of Ohio. When submitting queries, the website will return a table of named individuals (nurses) with data containing several attributes of interest: Status, License Number, License Issue Date, License Expiration Date, City, and Board Action. Every nurse working in Ohio should be present in this dataset. This dataset is not seamlessly accessible because it must be accessed via a web interface, but this can be automated with a simple web crawling program or interception of the JavaScript Object Notation (JSON) response sent by the website. To overcome website access limitations due to a website programming error when requesting medical licenses by state, we instead made a Freedom of Information Act (FOIA) request to obtain a copy of the nursing data presented by the website. The FOIA data contains 205,698 records with 18 fields (see Appendix B for a full list) and shares three fields in common with the target data set:

- Year of Birth (as Date of Birth MM/DD/YYYY)

- Initial License Year (as Issue Date MM/DD/YYYY)
- Postal Zip Code

The FOIA data request was meant to overcome two constraints caused by database query errors in the Ohio License Center website. Foremost the entire set of nurses in the state of Ohio cannot currently be pulled (an error which is acknowledged on the website). The website's intended use case is to verify small numbers (presumably one at a time) of nursing licenses by their individual name or license number. Querying individuals by name seems to return the record for the individual even if they were not returned when querying for all nurses in Ohio. We checked this functionality by querying for a small number individual nurses known by name who did not appear in the state wide data query. Secondly the responses returned on the website do not include date of birth (a key QID field for linking to the anonymous set). By receiving the date of birth directly the attack complexity is significantly reduced.

If date of birth were not granted as a part of the FOIA request we would have needed a secondary auxiliary source to provide it. The Ohio voter's registry contains date of birth, ZIP code, and full names of individuals. Before linking the anonymous data set we would first have to link records between the Ohio License Center website and the Ohio voter's registry. Since not every nurse will be registered to vote, the success of the attack would be degraded (yielding less disclosed identities). This is a more complex starting point and provides a lower ROI for attackers trying to recover the most identities possible. These constraining factors are also a limitation of other existing work on this topic. In [27] Latanya Sweeney had to purchase a copy of the Cambridge Massachusetts voter registry to conduct a record linkage to Massachusetts Group Insurance Commission data which would suffer from the same caveat as our work.

While the FOIA data supplies us with date of birth directly, an important set of individuals were lost in the data supplied. The FOIA request asked for all nurses in the state of Ohio, but did not specify the data set should include both active and inactive nurses.

The Ohio License Center website provides records of both active and inactive nurses while the FOIA data only includes records of active nurses. This is an important note because the anonymous data set being targeted contains records of both active and inactive nurses. Therefore the anonymous data is not a perfect subset of the identified set. This fact changes the conclusions that can be made about record links between the two sets (described in the next section).

3.2.4 Link the Target Dataset and the Auxiliary Information

The FOIA data contains three fields in common with the 2015 nursing data: date of birth (year of birth in the nursing dataset), issue date (initial license year in the nursing dataset), and postal ZIP code. Records are linked between the FOIA data and the 2015 nursing data through exact matches of all three QID fields. Linked records were then classified into one of three groups: 1:1 matches, 1:Many matches, and 1:0 matches. When an anonymous record is explicitly identified after linking it is a 1:1 match. When an anonymous record links to multiple potential identifying records that shared identifying information it is a 1:Many match. When an anonymous record could not be linked to any potentially identifying records in the FOIA set it is considered 1:0. Those records marked 1:Many may still be identifiable by an attacker through further examination of QID fields not used in this attack as described in section 3.2.2.

Table 3.1: Number of records identified through de-anonymization.

Record Group	Records	Percent of Total Records
1:1	64,051	34.97%
1:Many	59,988	32.75%
1:0	59,139	32.28%

Linking between the anonymous data set and the FOIA data set yields the results presented in table 3.1. The linkage attack explicitly identifies up to 64,051 of the 183,188

anonymous records by linking with the FOIA data set. We cannot say 64,051 identity disclosures are guaranteed due to the inactive nurse constraint described in 3.2.3. To make such an assertion the FOIA data must be requested again to contain every active and inactive nurse in the state of Ohio. The anonymous set includes records from nurses who are not actively practicing RNs including retirees (10,505 records), medically disabled persons, and persons who have elected to maintain their license while choosing not to work as a RN. Furthermore many records in the anonymous set were erroneous, in the sense that they contained obviously invalid values (e.g. a license year of 6377) making linking of those records impossible. A total of 27,806 unique records were deemed impossible to link, due to 28,011 invalid fields (27,511 invalid ZIP codes, 184 invalid initial license year, and 316 questionable year of births compared to the record's initial license year). In the case of year of birth all fields contained valid values, but when comparing the delta between year of birth and initial license date some nurses would have received their license at an unlikely age. It is difficult to say if the year of birth or initial license year fields contain invalid data in these cases.

3.2.5 Generate a De-anonymized Dataset Composed of the Target and Auxiliary Data

Completing the linking process provides an attacker with a de-anonymized dataset that contains an incredible number of attributes about a nurse's life. To name a few: full name, nursing license number, nursing license status (including termination), full address, former military affiliations, hours worked per week, and many more. The two data sets combined provide 57 unique fields about an individual nurse.

3.2.6 Improving Data Set Privacy

While the Ohio State Board of Nursing has proposed that the published work force data is anonymous, we see through this attack that statement is false. The data is ostensibly anonymized, a common problem witnessed across the data sets presented in chapter 2. There seems to have been an attempt in anonymizing the data set. We see that not only have explicit identifiers been suppressed, but specific dates have been generalized to the year level. Unfortunately, even with the generalizations made the data set is not instilled with the most basic anonymous data property: k-anonymity. This section explores the steps necessary to achieve k-anonymity, ℓ-diversity, and t-closeness.

Data Cleaning

Before we can generalize the 2015 nursing data we first must address the invalid records described in section 3.2.4. Nonsensical values such as non-4-digit initial license date must be corrected or suppressed from the data in order to achieve k-anonymity with any respectable amount of precision. The option to correct data in fields was decided against because there is no discernible pattern that justified adjustment of QID values. For example, a year entered as 86 could reasonably be assumed to mean 1986, but what about a year 3147? It is not clear how to properly correct this value. As a result, the following procedure was followed. In the case that postal ZIP code was deemed invalid a value of "*****" (suppression) was assigned to the field. For invalid year of birth or initial license year the value was suppressed by assigning "0" to the field. No data modification was conducted on valid year of birth and initial license data when the range between these dates for a single record did not make sense. Those records were treated as valid when generalizing the data.

Generalizing the Data: k-anonymity

To instill the dataset with the k-anonymous property we use an exhaustive anonymization algorithm: Preferred Minimal Generalization Algorithm (MinGen). This algorithm was presented by Latanya Sweeney in [26] which describes the algorithm as "a theoretical algorithm that uses generalization and suppression to produce tables that adhere to k-anonymity with minimal distortion." To achieve minimal distortion in the generalized tables we use a data utility metric called precision on the QID fields as defined by Sweeney in the same work. Effectively the precision metric describes how close the QID fields in the generalized data set are to their values in the original data set. When we further generalize the work force data using MinGen, the generalized data set that achieves the desired value of k with the highest precision is the data set that is selected.

In an attempt to discover the most precise generalized sets for each value of k, QID fields year of birth and initial year licensed are generalized in increments of 2 years while postal ZIP code is suppressed by one rightmost digit per iteration. Generalizing the fields at a higher rate (ranges greater than 2 years) can result in a less precise generalized set.

Table 3.2: Example data not generalized.

Year of Birth (YOB)	Initial License Year (ILY)	ZIP Code
1967	1992	47025
1951	1974	43082
1979	2008	43701

Table 3.3: Example data for k=2.

Year of Birth (YOB)	Initial License Year (ILY)	ZIP Code
[1950-1976)	[1984-2000)	*****
[1950-1976)	[1968-1984)	*****
[1976-2002)	[2000-2016)	*****

Tables 3.2 and 3.3 provide an example of the generalized format for k=2. The Year of Birth field takes on a range of 26 years and Initial License Year takes on a range of 16 years. ZIP Code is completely suppressed or generalized to its highest level possible (5). The values presented in tables 3.2 and 3.3 are consistent with the rows 1 and 2 in table 3.4 respectively. Table 3.4 details the precision, generalized ranges for year of birth and initial license year along with the suppression level of postal ZIP code for increasing values of k.

Table 3.4: Impact of k-anonymity on dataset precision.

k-value	Precision	YOB Range	ILY Range	ZIP Level
1	100%	1	1	0
2	39%	26	16	5
3	33%	24	26	5
7	29%	24	32	5
8	25%	14	48	5

Figure 3.1 shows that achieving a minimal value of k=2 results in a large loss in precision (61% loss). This huge loss in precision to achieve a small value of k is commonplace in large data sets, often due to what we will call "outlying records". Once an initial k of 2 is established, subsequent increases in k-anonymity require less compromise to precision. To achieve a k=3 we suffer an additional 6% (67% total) loss in precision from the original dataset and for k=7 we suffer another 4% (71% total) loss. Losses in precision for each QID have specific drawbacks when analyzing survey data. Losses of precision for Year of Birth reduce the accuracy of analyses about age demographics for the Ohio nursing work force. Losses in Initial License Year degrade the ability to correlate information related to workforce experience. A complete suppression of postal or residential ZIP code removes the ability to conduct geographic based analyses, for example the geographic distribution of nurses across the state of Ohio. While these losses in precision are significant, obtaining a small value of k=2 degrades the ability to conduct this record linkage attack bringing the number of 1:1 matches from 64,051 (34.97%) to 0.

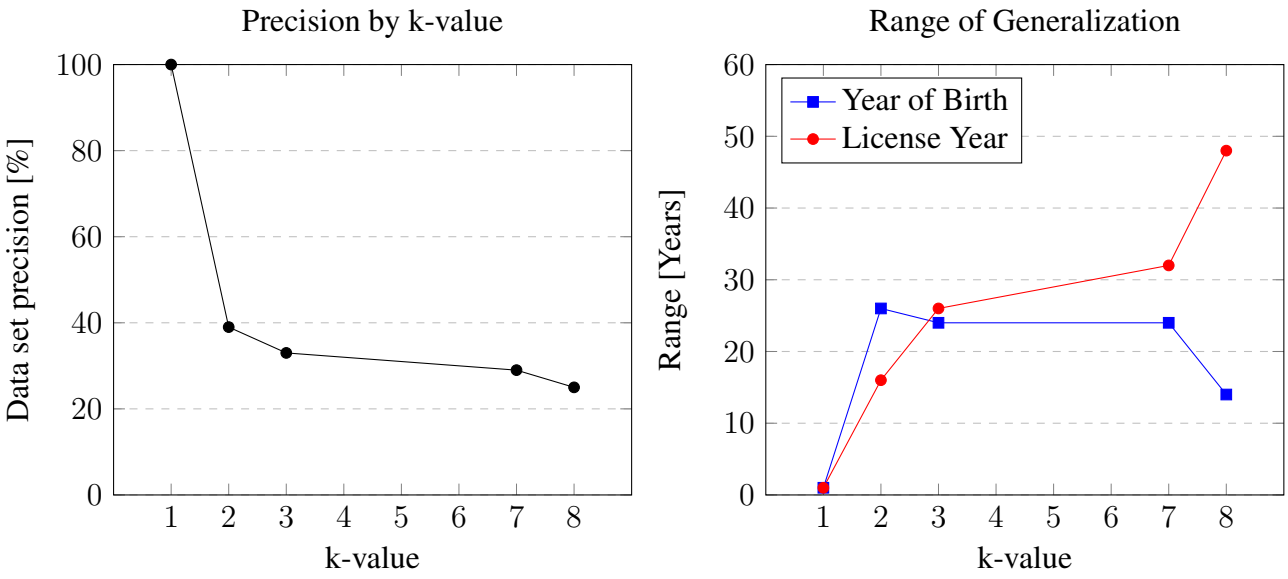

Figure 3.1: Loss in precision with increase in k-anonymity (left). Range of field generalization with increase in k-anonymity (right).

Figure 3.1 also shows some interesting compromises between the generalization ranges for Year of Birth and Initial License Year for increasing values of k. We see that a particular QID is constraining the data set from achieving different desired values of k. For example, to achieve k=2, Year of Birth is the constraining QID field, while for k=7 and k=8 License Year is the constraining QID and Year of Birth can remain more precise. The work force survey is quite expansive in terms of the age groups surveyed. The expansiveness of the survey results in several "outlying records" that increase the difficulty of achieving k-anonymity without major losses in data precision. In other words: the year of birth for several records is so far from the average year of birth (1970) in the data set that the QID groups required to achieve a small value of k-anonymity result in a significant loss in data utility.

There are some caveats to these results. This work weights all QIDs the same, but the QIDs can be weighted to favor precision of particular QID fields. Additionally, we use a very general data utility metrics in this work, whereas a more case-specific utility metric might better meet the needs of the Ohio Board of Nursing. It is also possible to

apply special purpose metrics such as data classifiers described in [12] when "the purpose of the data is known at the time of publication". In this case the data is general purpose demographic data. We therefore continue using the precision metric for the remainder of this analysis, with the exception of Section 3.2.6, which considers another general-purpose utility metric: the impact of the anonymization process on the accuracy of range queries made over the QID fields.

Generalizing the Data: ℓ-diversity

In order to assess ℓ-diversity on a dataset with high dimensionality we need to specify which attribute(s) are sensitive. In this case we will assume the sensitive attribute to be BSN_Plans. This attribute indicates whether or not an RN plans to pursue their Bachelor of Science in Nursing and, if so, the time frame they intend to obtain it. The possible survey choices for BSN_Plans are as follows:

- Do not plan to obtain a BSN
- Currently enrolled to obtain a BSN
- Plan to obtain in 1-5 years
- Plan to obtain in 6-10 years
- Plan to obtain in 11-15 years
- Plan to obtain but do not know when
- Not applicable (have BSN or higher nursing degree)
- Not Applicable

BSN_Plans is a good hypothetical sensitive value in the nursing workforce dataset because it could be important in a nurse's ability to be promoted. Consider a supervisor

who wants to promote an RN into a new position. The supervisor is biased against those who have plans to pursue their BSN because she thinks it will distract the RN from the new position and degrade their performance. The supervisor turns to the annual work force data to determine which candidates plan to obtain a BSN. Even if the work force data has been anonymized, the supervisor has enough information about her employees (background knowledge) to identify the QID groups of the employees in the anonymous data set.

Table 3.5: ℓ-diversity impact on dataset precision.

ℓ	k	Precision	YOB Range	ILY Range	ZIP Level
1	1	100%	1	1	0
2	7	29.3%	24	32	5
3	5	27.9%	32	26	5
4	71	9.3%	38	48	5

Table 3.5 shows that the value of k is several times that of ℓ when ℓ is small, and it increases dramatically at $\ell = 4$. The precision of the anonymous datasets is very low for even small values of ℓ.

Generalizing the Data: t-closeness

To assess t-closeness we can continue with our BSN_Plans example. First we must determine how often each of the 8 possible responses listed in 3.2.6 are represented. Counting the representation of each response will help us decide on a threshold value t. The threshold value will determine how large the distance between the representation of a response in an equivalence class can be from the representation of that response in the overall dataset.

By examining table 3.6 we can see representation of the responses varies significantly. The "Not Applicable (have BSN or higher nursing degree)" has extremely low representation (only 20 instances). This indicates that RNs who have obtained their BSN are particularly vulnerable to de-identification in the workforce dataset. The low representation also

Table 3.6: BSN_Plans Response Representation

Response	Count	Representation
Do not plan to obtain a BSN	46,672	25.48%
Currently enrolled to obtain a BSN	12,678	6.92%
Plan to obtain in 1-5 years	15,085	8.23%
Plan to obtain in 6-10 years	341	0.19%
Plan to obtain in 11-15 years	186	0.10%
Plan to obtain, but do not know when	14,522	7.93%
Not Applicable (have BSN)	20	0.011%
Not Applicable	93,684	51.14%

indicates picking a meaningful t-threshold will be difficult. A higher threshold value will allow more leniency in response representation at the equivalence class level. The threshold will need to be undesirably high to find acceptable anonymous datasets that do not fail the t-closeness criteria with respect to the under-represented responses.

The minimum t-threshold value that can be achieved without generalizing all of the QID fields to the greatest allowable range is t=1.0 (100%). A t-value this high may be acceptable for the fields with low overall representation such as "Not Applicable (have BSN)", but is unacceptable for more common fields like "Do not plan to obtain a BSN". For the most commonly represented field a t=1.0 allows a range of 0% to 50.96% representation for the field in any given equivalence class. To achieve t-closeness for a more meaningful value of t, it might be helpful to clean the data and consolidate fields. All of the "Plan to obtain" fields can be combined to "Plan to obtain in 1-15" years for an overall representation of 8.5% and "Not Applicable (have BSN)" can be made a part of "Not Applicable" raising its representation to 51.15%. However, even after performing this consolidation, the value of t required to obtain t-closeness was 0.99. Our conclusion is that data cleaning and field consolidation prove inadequate to instill this dataset with t-closeness.

Generalizing the Data: z-score record pruning

Working with the nursing work force data revealed interesting types of records that had significant impact on data set generalize-ability. The Ohio Board of Nursing is surveying RNs that are no longer a part of the core work force, such as retirees. We suggest that RNs with special life circumstances should be surveyed independently of the general work force survey. Including these individuals in the workforce survey significantly degrades the dataset's generalize-ability.

To illustrate this point we can score QID attributes of records using z-score (standard score) to identify and prune outlying records. This use of z-score allows us to show the potential gains in generalize-ability and precision if the scope of the workforce survey were reduced. An attribute value's z-score can be calculated with the following equation:

$$z = (X - \mu)/\sigma$$

If the z-score for a domain value is greater than or equal to a specified number of standard deviations away from the mean of the domain, that record is omitted from the workforce dataset prior to generalization. We have two QID attributes that serve as good candidates for z-score pruning: year of birth and initial license year. First we can examine the impact of z-score pruning on k-anonymous datasets based on year of birth.

Tables 3.7 shows the improvements in precision at various levels of k-anonymity. Table 3.7 shows that we can improve the dataset's precision for k=2 by 11.6% by pruning of records with a year of birth greater than or equal to 1.9 standard deviations from the mean. For k=3, pruning 2,167 records recovers 10.9% of the overall dataset's precision. Through z-score pruning we can now obtain finer grained values of k-anonymity that we could not achieve before z-score pruning. A k=4 at 38.7% precision was not previously possible with the unclean data, instead the next step in k would have been k=7 at 29% precision.

Table 3.7: z-score pruning impact on k-anonymous dataset precision.

z threshold	k	Precision	Gain	YOB Range	ILY Range	Records Pruned
2.0	2	45.3%	6.3%	14	18	2,167 (1.2%)
1.9	2	50.6%	11.6%	14	10	3,637 (2.0%)
1.8	2	50.6%	11.6%	14	10	5,821 (3.2%)
1.7	2	50.6%	11.6%	14	10	8,924 (4.8%)
1.6	2	45.3%	6.3%	14	18	15,270 (8.3%)
1.5	2	49.3%	10.3%	14	12	24,351 (13.3%)

z threshold	k	Precision	Gain	YOB Range	ILY Range	Records Pruned
2.0	3	43.9%	10.9%	8	26	2,167 (1.2%)
1.9	3	43.9%	10.9%	8	26	3,637 (2.0%)
1.8	3	43.9%	10.9%	8	26	5,821 (3.2%)
1.7	3	43.9%	10.9%	8	26	8,924 (4.8%)
1.6	3	43.9%	10.9%	8	26	15,270 (8.3%)
1.5	3	43.9%	10.9%	8	26	24,351 (13.3%)

z threshold	k	Precision	Gain	YOB Range	ILY Range	Records Pruned
2.0	4	38.7%	N/A	12	30	2,167 (1.2%)
1.9	4	38.7%	N/A	12	30	3,637 (2.0%)
1.8	4	38.7%	N/A	12	30	5,821 (3.2%)
1.7	4	38.7%	N/A	12	30	8,924 (4.8%)
1.6	4	36.0%	N/A	14	32	15,270 (8.3%)
1.5	4	36.0%	N/A	14	32	24,351 (13.3%)

Additionally we can bring in initial license year to see if additional records can be pruned to increase our precision. Table 3.8 shows that pruning records based on initial license year in addition to year of birth yields very few additional records to prune. This is likely due to the relationship between the two fields. We would expect these fields to have similar distributions because most nurses receive their initial license in early adulthood.

The results on ℓ-diversity of employing z-score pruning are shown in Table 3.9. It should be noted that in this particular dataset, every ℓ-diverse anonymized dataset has an associated k value greater than 2. Based on our results presented in Table 3.7, we therefore use a z-score threshold of 2.0 for all values of ℓ in the table. Using a threshold lower than 2.0 would only omit unnecessary records while providing no additional gain in precision.

Table 3.8: z-score year of birth and initial license year combined pruning impact on k-anonymous (k=2) dataset precision.

z threshold	k	Precision	Gain	YOB Range	ILY Range	Records Pruned
2.0	2	45.3%	6.3%	14	18	2,170 (1.2%)
1.9	2	50.7%	11.7%	14	10	3,640 (2.0%)
1.8	2	50.7%	11.7%	14	10	5,824 (3.2%)
1.7	2	50.7%	11.7%	14	10	8,927 (4.8%)
1.6	2	45.3%	6.3%	14	18	15,273 (8.3%)
1.5	2	49.3%	10.3%	14	12	24,355 (13.3%)

Employing z-score pruning has some interesting impacts on ℓ-diversity. In particular, we can see from comparing Table 3.9 with Table 3.5 that z-score pruning sometimes actually lowers precision. For instance, the two tables show that pruning lowers the k value associated with ℓ=2 from 7 to 4, but achieving ℓ=3 results in a large increase of k (from 5 to 12) and a corresponding loss in precision. This is because many of the outlying records that are pruned also have underrepresented values for the sensitive attribute, and removing them therefore makes it harder to achieve ℓ-diversity for some values of ℓ.

Finally, attempts were made to achieve an anonymous dataset instilled with t-closeness at a respectable value for t through the use of z-score record pruning, but these were unsuccessful. This was expected. The pruning procedure handles constraining cases of QID attributes that prevent merging records into a single equivalence class, which directly improves k-anonymity. While this procedure has some effects that are passed onto ℓ-diversity, t-closeness is isolated from these effects because t-closeness solely deals with distributions of sensitive field values.

Table 3.9: z-score pruning impact on ℓ-diversity dataset precision.

z threshold	ℓ	k	Precision	Gain	YOB Range	ILY Range	Records Pruned
2.0	2	4	38.7%	9.4%	12	30	2,170 (1.2%)
2.0	3	12	25.3%	-2.6%	20	42	2,170 (1.2%)
2.0	4	17	24.0%	14.7%	22	42	2,170 (1.2%)
2.0	5	68	13.3%	N/A	38	42	2,170 (1.2%)

Generalizing the Data: range query analysis

The precision metric provides QID utility information that encompasses all of the chosen QID fields. While this is beneficial, the precision metric does not provide direct accuracy information about a specific QID. A range query (RQ) analysis can provide more granular accuracy information for particular QID fields.

To conduct this analysis we need three components: the original dataset, the anonymous dataset, and two sets of randomly generated queries (one for YOB and ILY). In this case we randomly generated sets of 1,000 range queries. For each range query a comparison is made between the number of records returned in the original data to the number of records returned in the anonymous data. Then the average accuracy and standard deviation for the 1,000 range queries is calculated for each QID. Note that because the records are generalized rather than perturbed, all of the records returned are always correct, i.e. their true QID value is always within the desired range. The only inaccuracy comes from the return of some additional records whose true QID value is not in the desired range.

$$Accuracy = OriginalRecords/AnonymousRecords$$

The process of randomly generating and executing range queries over the QID fields is repeated for each anonymous version of the dataset to build tables 3.10 and 3.11. Note

Table 3.10: Range query analysis of k-anonymous datasets.

k	Precision	RQ YOB Precision	YOB Std Dev	YOB Range
1	100%	100%	N/A	1
2	39%	55.48%	29.58%	26
3	33%	53.59%	29.10%	24
7	29%	54.48%	29.34%	24
8	25%	62.06%	27.57%	14

k	Precision	RQ ILY Precision	ILY Std Dev	ILY Range
1	100%	100%	N/A	1
2	39%	55.88%	29.58%	16
3	33%	50.65%	29.71%	26
7	29%	43.10%	28.51%	32
8	25%	32.04%	27.23%	48

Table 3.11: Range query analysis of k-anonymous datasets with z-score pruning at z threshold 2.0.

k	Precision	RQ YOB Precision	YOB Std Dev	YOB Range
2	45.3%	67.66%	24.92%	14
3	43.9%	75.39%	22.32%	8
4	38.7%	69.23%	24.91%	12

k	Precision	RQ ILY Precision	ILY Std Dev	ILY Range
2	45.3%	49.70%	29.66%	18
3	43.9%	50.12%	29.93%	26
4	38.7%	47.15%	30.47%	30

that the standard deviation column in each of these tables is with respect to the precision column. For example, the RQ YOB precision for k = 2 gives us the range 25.9% to 85.06%. In these tables we see that the average range query precision of YOB and ILY is considerably higher than the overall precision of the dataset. This is expected because ZIP code has a large influence in overall dataset precision because it is always suppressed to the highest level to achieve k-anonymity. If range query analysis were conducted on ZIP code its RQ precision would be extremely low because every range query on the anonymous dataset would return the entire set of records (183,188). We also see that both fields have

large standard deviations that increase with higher ranges for the QID field (more obvious for ILY in table 3.11). At higher field generalization ranges, higher standard deviation is expected because there is a smaller number of generalization ranges (buckets). This generalization range trade-off feature of the QID fields was originally highlighted in figure 3.1. Finally in table 3.11 we see that z-score pruning has similar benefits in range query precision as overall data precision.

3.2.7 Outcomes

This analysis has provided several perspectives on the impact of generalization schemes on the Ohio nursing workforce dataset. In its raw form the dataset is vulnerable to record linkage and background knowledge based attacks. It is possible to reveal a significant number of nurses who participate in the annual survey from a general approach. More targeted individual attacks are likely to be significantly easier to conduct and far less sophisticated to construct. The methods employed in this work took basic programming skills with one auxiliary knowledge source. Adversaries targeting a specific individual in the dataset have an even lower threshold of difficulty to identify their target. De-anonymizing a single individual in the set can be as simple as filtering fields in the raw data data by column, based on an attacker's background knowledge about their target. This type of attack should be a significant concern of the Ohio Board of Nursing because attackers who are a part a nurse's workplace have a significant advantage in de-anonymizing an individual co-worker.

The use of k-anonymity proves to be enough to stop basic record linkage attacks against the dataset as a whole. However data utility of this dataset is significantly degraded due to the survey's broad scope, which has been demonstrated through z-score record pruning. We recommend that the Ohio State Board of Nursing re-evaluate the needs of their annual workforce survey from three positions. First, reconsider the population of nurses being surveyed to better represent the active work force to reduce outlying records. Second, reconsider the number of questions being surveyed since they provide a wide scope

of characteristics unique to individuals. Third, consider more thorough data cleaning and auditing of the workforce data to improve its generalize-ability. It would be most practical to re-design the survey with anonymization in mind. Finally, we recommend the Ohio Board of Nursing halt advertising the workforce data as anonymous to survey takers and data consumers until the privacy vulnerabilities have been resolved. Improving the survey design can likely mitigate the alarming losses in anonymous data utility and lead to a both useful and privacy preserving workforce dataset.

Analysis of the Impact of the Semantic Web on De-Anonymization Attacks

4.1 Data Availability

Record linkage attacks require a dataset against which to link the anonymized data. Historically, most data was inconvenient to use since it was available only as databases or on file servers as spreadsheets, CSV files, or tables in PDF documents. Retrieving the data could also be difficult. For instance, some repositories might be accessible via websites or structured query mechanisms while others required a login and use of secure file transfer protocols. Financial drawbacks also inhibited data integration. Some data might be stored using proprietary formats that required expensive software licenses to read. These obstacles made finding and retrieving data related to an attacker's target dataset difficult.

The rise of linked data has changed this situation drastically. Linked data is expressed as RDF and can be accessed using standard protocols. It is also prolific. The Linked Open Data (LOD) Cloud now contains over 31 trillion triples across 295 datasets, with more than 503 million links across datasets [3]. The most represented domains in the cloud are social network information and government data, composing over 51% and 18%, respectively of the total [3]. This is a huge amount of information about many different aspects of people's lives, and the potential for its misuse should not go unconsidered. For example,

Table 4.1 shows data from the Open University in the United Kingdom.[1] This information can be downloaded in various formats or accessed via a SPARQL endpoint. It includes information about a person's job title, groups they belong to, and publications they have co-authored. While this data is not generally considered sensitive, it could be used as a quasi-identifier for a target dataset. Additionally, some information included in this data, including usernames on social media platforms such as Twitter and LinkedIn and a list of other linked datasets in which this person appears, can be used to find more information about this person.

Table 4.1: A record from the Open University personal profile linked dataset.

Property	Object
Title	Dr
Given name	James
Family name	Rees
Job title	Anthony Nutt Sr Research Fellow
inDataset	Open Research Online
inDataset	OU People Profiles
Account	https://www.linkedin.com/nhome/...
Holder of	http://data.open.ac.uk/role/ResearchStaff
Mailbox SHA1 sum	4d1c4c2e8f5...34d55078a3f
Has membership	http://.../the-open-university-business-school

Another quickly growing type of data on the Semantic Web is that annotated with schema.org markup.[2] Schema.org is an initiative by major search engine companies to facilitate the description of entities and the relationships between them using a basic syntax expressed as RDFa, Microdata, or JSON-LD. As of 2014, more than 36% of websites in Google's crawl contained schema.org markup [20]. It is particularly commonly used to describe people, businesses, events, and reviews. Fields relevant to people include many likely quasi-identifying fields, such as birth date, birth place, gender, nationality, and affiliation.

[1] http://data.open.ac.uk/page/context/people/profiles
[2] https://schema.org

4.2 Data Relevance

While the rise of linked data and schema.org markup has made much more data available in an easily accessible manner, a record linkage attack relies on finding datasets that include relevant information about the individuals in the target dataset. Finding an appropriate dataset is often the most time-consuming aspect of a record linkage attack. However, some research currently underway can speed up this process, thereby lowering the barrier to deanonymization.

Many linked datasets have very complex or extremely simple schemas. It is often difficult for a potential user of a dataset to quickly identify whether or not the data is useful for their purpose, but numerous methods for summarizing a linked dataset speed up this process. For example, the linked data summarization approach described in [29] ranks the axioms within an ontology based on graph-based measures such as centrality, while Loupe is an online tool that provides statistics regarding the usage of properties to describe instances of particular types within a linked dataset.[3]

Visualization tools are another avenue for quickly determining the general content and structure of a dataset. Many visual interfaces for data exploration on the Semantic Web involve displaying the RDF data as a graph. Unfortunately, graph-based representations frequently place entities based on graph metrics such as centrality or density, rather than according to their semantic meaning. They also have difficulty scaling to large datasets without becoming unwieldy. Kow and his colleagues have attempted to move beyond this towards more semantic-based layout algorithms with their idea of an "information landscape," which places similar concepts near one another and labels clumps of entities with terms describing the group. Users can select areas of interest that seem likely to contain relevant entities, which automatically filters the mappings shown in the list. This method of filtering allows users to systematically explore an ontology at a high level of detail without losing track of the big picture [16]. Other approaches handle the problem of scalability by

[3] http://loupe.linkeddata.es/loupe/

providing an RDF triple browser interface rather than attempting to show the entire dataset at once [10].

Ontology Design Patterns (ODP) provide another means for quickly determining linked dataset relevance to attackers. ODPs are self-contained, reusable patterns that model concepts that commonly occur across different ontologies. A well designed ODP describes the key aspects, and only the key aspects, of the concept being modeled [6]. By considering only the core features of a concept, ODPs avoid making any unnecessary specific ontological commitments, which allow them to be applied in a wide range of situations. As mentioned previously, many linked datasets either have very complex schemas or extremely simple schemas. In either case, it is difficult for a potential user of the dataset (including an attacker seeking data to de-anonymize a target dataset) to quickly determine whether or not the data is useful for their purpose. An ODP is a precise domain-agnostic representation of a concept. If an attacker can quickly isolate the part of a complex schema most related to an ODP of interest, or quickly determine whether or not data with little schema information fits into the ODP model, they would have a better idea of whether or not the dataset in question was useful. For example, a person's communications often reveal much about them. Blomqvist posted an ODP on the website ontologydesignpatterns.org to model a "Communications Event." A simplified illustration of this ODP is shown in Figure 4.1.

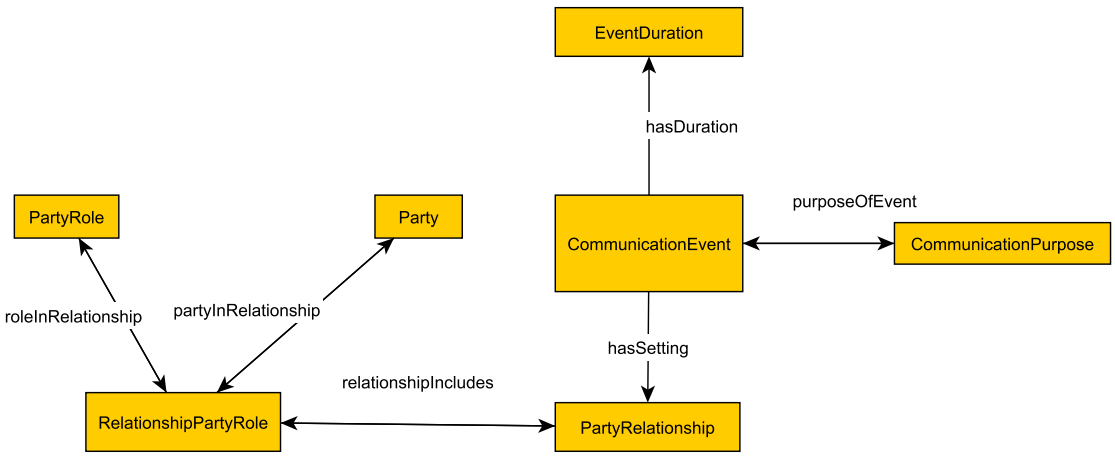

Figure 4.1: A subset of the Communications Event ODP

The dataset containing information about the 2012 European Semantic Web Conference available at http://data.semanticweb.org/dumps/conferences/ contains information about, among other things, the keynote talks given at the conference, including their start and end times, the speaker, the topic, the setting in which the talk occurred, and the title and subject matter. A method for detecting the presence of an ODP (such as Communications Events) in a linked dataset was proposed by Khan and Blomqvist in [15]. This method could be employed to learn that the ESWC data may be useful for de-anonymizing a dataset known to consist of Semantic Web researchers. For datasets with a significant schema, an ontology alignment system (described in Section 4.3) could also be used to recognize that this dataset contains Communications Events.

The Enron email corpus is another dataset that contains Communications Events. This is an example of a dataset with very little schema information (in fact, this data is only available as a relational database or PST files). Some researchers are exploring techniques to recognize when information in data sources such as this is related to a particular ODP. Since Personal Storage Tables (PSTs) are known to store information about e-mail, calendar events, and other items associated with Microsoft Software (such as Microsoft Outlook) it is reasonable to associate PST files with Communication Events.

By associating or identifying target datasets in alignment with a given ODP we can

immediately gain knowledge about the data contained in that set. Methods for associating ”flat data” such as CSV or text files with ODPs means that even data which is not represented in linked data format can be identified as if it were and leveraged to provide quasi identifiable information for identifying individuals in target datasets.

4.3 Data Linking

Once a relevant dataset has been identified, the target dataset must be joined with it based on the quasi-identifier values. This seems straightforward, but can actually be quite difficult in practice because the schemas for the two datasets were likely developed by different people, for different purposes. Because of this, even two ontologies that represent the same domain will generally not be the same. They may use synonyms for the same concept or the same word for different concepts, they may be at different levels of abstraction, they may not include all of the same concepts, and they may not even be in the same language. Furthermore, the classes and properties in the ontologies may not be used consistently when describing the entities within the dataset. The goal of ontology alignment is to determine when an entity in one ontology is semantically related to an entity in another ontology, despite these challenges.

Alignment systems typically use a combination of three different approaches to evaluate entity similarity: syntactic, semantic, and structural similarity metrics. Syntactic metrics compare entities from each of the ontologies to be aligned based on strings associated with the entities. The strings are generally the entity label, but can also include comments or other annotations of the entity. Semantic similarity metrics attempt to use the meanings of entity labels rather than their spellings. External resources such as thesauri, dictionaries, encyclopedias, and web search engines are often used to calculate semantic similarity [14, 28]. Structural techniques consider the neighborhoods of two entities when determining their similarity. For instance, two entities with the same superclass that share some

common instances are considered more similar than entities that do not have these things in common. Graph matching techniques are often used for this [13, 9]. For a comprehensive discussion of ontology alignment, including a formal definition, see [11].

The previous section detailed how ontology design patterns represent the core components of a concepts, as identified by domain experts. ODPs can facilitate ontology alignment by reducing the complexity inherent in dealing with two ontologies likely developed by different designers with different applications in mind. Rather than trying to align these two ontologies directly with one another, they can instead each be aligned to the application-neutral ODP. Furthermore, leveraging ODPs can enhance the scalability of an alignment algorithm by clustering entities into different ODPs. Individual entity relations would then need to be determined only within those in an ODP rather than across the whole of the ontologies.

The Ontology Alignment Evaluation Initiative (OAEI) is a set of benchmarks for evaluating the performance of alignment systems. The initiative has held evaluations annually since 2005. Over that time, the accuracy and the variety of problems handled by alignment systems have increased, while runtimes have decreased.[4] The top performing alignment systems include two that are available online: AgreementMakerLight[5] and LogMap.[6] These systems achieve an F-measure of .76 and .73, respectively, on an OAEI track based on aligning ontologies related to conference organization. These results are approaching the level of consensus that humans have when performing alignment tasks [8], implying that the dataset linking phase may be an aspect of record linkage attacks that could be automated in the near future. Additionally, alignment systems could be used to attempt to align datasets to ODPs representing key concepts, such as a Person, in order to refine a collection of possibly-relevant datasets for further analysis. Aligning a dataset against an ODP rather than another dataset can be easier, due to the limited scope and application-neutral nature

[4] http://oaei.ontologymatching.org
[5] https://github.com/AgreementMakerLight/AML-Jar
[6] http://csu6325.cs.ox.ac.uk

of an ODP.

Coreference resolution algorithms attempt to determine when the same instance (i.e. individual) is referred to in two in different ways. For instance, is John Q Public in one dataset the same person as J.C. Publick in another? This is the Semantic Web technology most closely related to deanonymization: determining whether a person whose name and social security number have been replaced with random strings, for example, is present within an external dataset such as voter registration records is precisely what coreference resolution algorithms attempt. Most current approaches use string similarity metrics to compare two instances based on their property values (e.g. zip code, age, height) or their property values together with the names of those properties. Top performing systems on the mainbox instance matching task include the aforementioned LogMap and RiMOM [25], with F-measures of .83 and .91, respectively.

The general decisions made by the designer of a coreference resolution system are: what instances to compare, how to compare them, and how to determine if the result of that comparison implies that the two instances are equivalent. The number of instances in a dataset can be extremely large. As a result, it is not considered feasible to compare every instance in one dataset to every instance in the other in order to determine if they are the same. Instead, some method of deciding whether two instances are "close enough" that they are worth comparing must be established. If the filtering step has decided that two instances are close enough to warrant further scrutiny, the algorithm will compare them based on a selection of features. In most current coreference resolution systems, these features are either property values alone or property values together with property names. Regardless of what features are compared, the most common method for comparing them is via string similarity metrics. This is because even when a property is a non-string data type, such as a date or URL, it is often expressed as a string in datasets. Different string metrics are employed, primarily depending on the length of the strings to be compared. A survey of string metrics commonly used by these systems is provided in [7]. A decision must also

be made on how much to weight each feature. Various methods have been proposed for this, including both supervised [24] and unsupervised [22] machine learning approaches. Finally, the coreference resolution system must take the outcome of a comparison of two instances and make a decision on whether or not those instances are equivalent. This is often done by specifying thresholds and other parameters of the algorithm.

4.4 Data Inferencing

Traditional record linkage attacks sometimes involve specific or general knowledge or assumptions an attacker has about a target. This can be made significantly easier using automated reasoners. For example, assume the attacker is working with a medical records dataset organized according to a schema which includes the statements below.

```
<exSchema:hasDisease> <rdfs:domain> <exSchema:Person>
<exSchema:hasDisease> <rdfs:range> <exSchema:Disease>
<exSchema:LungDisease> <rdfs:subClassOf> <exSchema:Disease>
<exSchema:HeartDisease> <rdfs:subClassOf> <exSchema:Disease>
<exSchema:hasEthnicity> <rdfs:domain> <exSchema:Person>
<exSchema:hasEthnicity> <rdfs:range> <xsd:string>
<http://data.ex.org/person/12345> a <exSchema:Person>
<http://data.ex.org/person/12345> <rdfs:nameFull> "Zhang Lu"
```

If the attacker knows that Zhang Lu has some disease and wants to determine what it is, he can add some additional statements to the knowledge base that reflect his assumptions and then use a reasoner to check whether or not it is possible to infer the disease Mr. Zhang has, based on those assumptions. For instance, the attacker may know that Mr. Zhang is of Asian ancestry. He could then add the following fact to the knowledge base:

```
<http://data.ex.org/person/12345> <exSchema:hasEthnicity> "Asian"
```

The attacker could further assume that people of Asian ancestry are unlikely to get heart disease (based on statistical knowledge). The attacker would add the following fact to the knowledge base:

```
(exSchema:Person and (exSchema:hasEthnicity Asian)) SubClassOf:
not (exSchema:hasDisease some exSchema:HeartDisease)
```

The attacker could then use an automated reasoner to determine whether or not Mr. Zhang's disease could be inferred. While space constraints force this example to be relatively simplistic, it shows that the attacker can use existing Semantic Web languages and tools to quickly explore the ramifications of any assumptions he would like to make.

Conclusion

This work has explored breaches of anonymized data from an attacks and defenses perspective. We surveyed four de-anonymization incidents to find common features of record linkage attacks. From this survey we conclude that a record linkage attack is composed of five major steps:

1. Identify and acquire the target dataset.
2. Examine the target dataset for potentially identifying attributes.
3. Identify and acquire auxiliary information (external knowledge or a secondary dataset).
4. Construct method(s) for establishing links between the target dataset and the auxiliary information.
5. Generate a de-anonymized dataset composed of the target data and auxiliary information.

We also learn from this survey that most anonymization breaches occur because data publishers fail to publish their datasets such that they meet even the basic criteria for anonymity. To further substantiate these findings we then apply the anatomy to conduct a de-anonymization (record linkage) attack. We successfully de-anonymize a dataset that is claimed to be anonymous by the data publisher. Through this attack we validate the attack anatomy and identify another case where a data publisher failed to instill their data set with basic criteria for anonymity. This work highlights that record linkage attacks are

still a threat despite the fact that techniques to defend against them have existed for over 14 years. We explore the trade space of the publisher's dataset and elaborate on how proper generalization impacts dataset utility and privacy. Based on these findings we recommend the data publisher re-design their survey to better suit their privacy needs.

Finally we address how the emergence of semantic web technologies may facilitate de-anonymization attacks in the future. Linked data and schema.org markup has increased the availability of data that can be used to conduct attacks. Tools for data summarization and visualization can assist an attacker in sifting through this data to find a relevant dataset with which to link the target dataset. Meanwhile, the emergence of automated techniques for ODP identification, ontology alignment, coreference resolution, and reasoners hold the potential to one day fully automate record linkage attacks. The ramifications of Semantic Web technologies on privacy are likely to be profound. Aggarwal showed that typical approaches to anonymize data break down in the face of high dimensionality [1], which is precisely what linked data provides. Dealing with this may require difficult decisions about how to publish sensitive data, potentially involving perturbing the sensitive values [2], which has corresponding impacts on its utility. These concerns are present whether the sensitive data is published as linked data, in a database, CSV file, or other traditional format, because as we have seen, even innocuous data about a person available on the Semantic Web can be used to de-anonymize a standalone dataset.

Several pieces of future work can stem from the contributions and predictions presented. One of the conclusions of the survey of anonymization breaches and record linkage attack on the nursing data was that data publishers often fail to understand how to properly protect their data from this type of attack. This seems to be an education issue more than a technical issue. Therefore, one line of future research might be to explore ways of expressing to data providers the weakness of their anonymization strategy and possible ways to address that, such as the positive impact on privacy from publishing fewer fields from each record or removing outliers from the dataset prior to anonymization.

Another obvious path for future work is to implement a proof of concept automated record linking system that utilizes the technologies described in Chapter 4. This would hopefully spur more research related to addressing the privacy implications of semantic web technologies.

Bibliography

[1] Charu C Aggarwal. On k-anonymity and the curse of dimensionality. In *Proceedings of the 31st international conference on Very large data bases*, pages 901–909. VLDB Endowment, 2005.

[2] Charu C Aggarwal and S Yu Philip. A general survey of privacy-preserving data mining models and algorithms. In *Privacy-preserving data mining*, pages 11–52. Springer, 2008.

[3] Chris Bizer Anja Jentzsch, Richard Cyganiak. State of the lod cloud, September 2011. http://lod-cloud.net/state/ [Online; posted 29-February-2016].

[4] Michael Barbaro and Tom Zeller Jr. A face is exposed for aol searcher no. 4417749, August 2006. http://www.nytimes.com/2006/08/09/technology/09aol.html?_r=0 [Online; posted 25-February-2016].

[5] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data - the story so far. pages 1–22, 2009.

[6] Michelle Cheatham. *The Properties of Property Alignment on the Semantic Web*. PhD thesis, Wright State University, 2014.

[7] Michelle Cheatham and Pascal Hitzler. String similarity metrics for ontology alignment. In *The Semantic Web–ISWC 2013*, pages 294–309. Springer, 2013.

[8] Michelle Cheatham and Pascal Hitzler. Conference v2.0: An uncertain version of the oaei conference benchmark. *International Semantic Web Conference*, pages 33–48, 2014.

[9] Beniamino Di Martino. Semantic web services discovery based on structural ontology matching. *International Journal of Web and Grid Services*, 5(1):46–65, 2009.

[10] Orri Erling and Ivan Mikhailov. Faceted views over large-scale linked data. *LDOW*, 2009.

[11] Jérôme Euzenat and Pavel Shvaiko. *Ontology matching*, volume 18. Springer Heidelberg, 2007.

[12] Benjamin Fung, Ke Wang, Rui Chen, and Philip S Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys (CSUR)*, 42(4):14, 2010.

[13] Brian Gallagher. Matching structure and semantics: A survey on graph-based pattern matching. *AAAI FS*, 6:45–53, 2006.

[14] Prateek Jain, Pascal Hitzler, Amit P Sheth, Kunal Verma, and Peter Z Yeh. Ontology alignment for linked open data. In *The Semantic Web–ISWC 2010*, pages 402–417. Springer, 2010.

[15] Muhammad Tahir Khan and Eva Blomqvist. Ontology design pattern detection-initial method and usage scenarios. In *SEMAPRO 2010, The Fourth International Conference on Advances in Semantic Processing*, pages 19–24, 2010.

[16] Weng Onn Kow, Vedran Sabol, Michael Granitzer, Wolfgang Kienrich, and Dickson Lukose. A visual soa-based ontology alignment tool. In *Proceedings of the 6th International Conference on Ontology Matching*, pages 242–243. CEUR-WS. org, 2011.

[17] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *ICDE (Purdue University)*, 2007.

[18] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. March 2007.

[19] Jennifer L Malin. Envisioning watson as a rapid-learning system for oncology. *Journal of Oncology Practice*, 9(3):155–157, 2013.

[20] Peter Mika and Tim Potter. Metadata statistics for a large web corpus. *LDOW*, 937, 2012.

[21] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pages 111–125, May 2008.

[22] Andriy Nikolov, Mathieu dAquin, and Enrico Motta. Unsupervised learning of link discovery configuration. In *The Semantic Web: Research and Applications*, pages 119–133. Springer, 2012.

[23] Vijay Pandurangan. On taxis and rainbows, June 2014. https://tech.vijayp.ca/of-taxis-and-rainbows-f6bc289679a1#.ymi2epx2x [Online; posted 25-February-2016].

[24] Shu Rong, Xing Niu, Evan Wei Xiang, Haofen Wang, Qiang Yang, and Yong Yu. A machine learning approach for instance matching based on similarity metrics. In *The Semantic Web–ISWC 2012*, pages 460–475. Springer, 2012.

[25] Chao Shao, Lin-Mei Hu, Juan-Zi Li, Zhi-Chun Wang, Tonglee Chung, and Jun-Bo Xia. Rimom-im: A novel iterative framework for instance matching. *Journal of Computer Science and Technology*, 31(1):185–197, 2016.

[26] Latanya Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):571–588, 2002.

[27] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10, May 2002.

[28] Julia M Taylor, Daniel Poliakov, and Lawrence J Mazlack. Domain-specific ontology merging for the semantic web. In *Fuzzy Information Processing Society, 2005. NAFIPS 2005. Annual Meeting of the North American*, pages 418–423. IEEE, 2005.

[29] Xiang Zhang, Gong Cheng, and Yuzhong Qu. Ontology summarization based on rdf sentence graph. In *Proceedings of the 16th International Conference on World Wide Web*, pages 707–716. ACM, 2007.

Appendix A

Ohio Nursing Workforce Survey Fields

License Type
License Sub-Category
Gender
Race
Year of Birth
Initial Nursing Credential
Highest Level of Education
Education Country
Initial Nursing Education State
Initial Year Licensed
Initial Country Licensed
Current Job Status
Unemployed Seeking Job
Unemployed Reason
Employed as Nurse
Nurse Work Type
Current Paid Positions Worked as Nurse
Weeks Worked in Last Year
Total Hours Worked as Nurse
Employment Zip Code
Residential Zip Code
Work Setting
Position Title
Certification 1
Certification 2
Certification 3
Primary Practice Area
Hours Worked in Primary Practice
Secondary Practice
Hours Worked in Secondary Practice Area
Secondary Work Setting
Secondary Position Title
Secondary Certification 1
Secondary Certification 2
Secondary Certification 3
Secondary Practice Area
Hours Worked in Secondary Practice
Plans to Obtain BSN
Why No Plans to Obtain BSN
Other Languages
Nursing Board Service
Changed Employer in last Year
In Armed Forces

Appendix B

Ohio FOIA Nursing Data Fields

License Type
License Sub-Category
License Status
License Sub-Status
Board Action
Initial License Issue Date
License Effective Date
License Expiration Date
License Number
Last Name
First Name
Birth Date
Street Address
City
State
ZIP Code
Country
County
E-mail Address