5-2015

Data Anonymization and its Effect on Personal Privacy

Nicole Ganz
University at Albany, State University of New York

Data Anonymization and its Effect on Personal Privacy

An honors thesis presented to the

School of Business,

University at Albany, State University Of New York

in partial fulfillment of the requirements

for graduation from the Honors College.

Nicole Ganz

Research Advisor: Yuan Hong

May, 2015

Abstract

Everyone's personal privacy is extremely vital and people go through great lengths to ensure that their information is secure. Corporations who collect or utilize data do realize this and strive to protect their customer's information. But can data be released to research companies without jeopardizing the privacy of the people? The more sensitive the released data is, the more companies have to mitigate the probability of being able to link the data together. Companies collect big data because it can be examined to predict patterns and be used for marketing, thus increasing their chances of becoming successful. To replicate the data companies and academic institutions collect, I will randomly simulate the information. The data I will use is not real; it cannot be linked to any actual persons. Additionally, I will discuss what methods are currently being used to anonymize data and why it is important to minimize the risk of linkage and leakage.

Acknowledgements

I would like to thank Dr. Hong for helping me and encouraging me throughout the entire thesis process.

Additionally, thank you to Professor Haugaard for his support throughout my Honor's college journey.

Table of Contents

1. Introduction

As technology is becoming more advanced, retaining personal privacy has proven to become severely more challenging. Every time any type of transaction takes place, sensitive data is exchanged and undergoes the possibility of being exposed. Attaining privacy can be described as "not having undocumented personal knowledge about one, possessed by others" [24]. A person's security and privacy significantly plummet as their information becomes more exposed. This could happen through newspapers, public voting lists, or any other type of legitimate document. For example, I ordered the Voter Registration Data from the New York State Board of Elections. I received the data set of everyone in Albany, who registered as a Democrat or a Republican. This information was mailed free to me within two days. The reason I was able to receive this information is because of the Freedom of Information Act (FOIA) [28]. This act says that any person has the right to access federal agency records, "except to the extent that such records are protected from public disclosure". This made it very simple for me to request the voting records. On the diskettes came the first and last names of each voter, his/her address, his/her town, and his/her area code. Depending on each state's rules, this could be possible across the United States. This sensitive data is open to the public, and many people do not realize that their information is accessible.

Furthermore, retaining complete privacy is debatably impossible to achieve; there are no limits to the kinds of information that can be acquired just from a person's physical appearance [24]. For example, if one was to meet a stranger in a local restaurant, many things could be derived from that interaction. Non-sensitive data such as his/her age, sex, hair color, or name can be noticed immediately. But afterwards, a simple internet search using the person's name and approximate

age could lead to finding sensitive information such as his/her address. Constantly, obtaining full privacy proves to be an ongoing challenge.

Section 1.1: Netflix Scandal

Ensuring privacy is extremely vital and when companies do not properly anonymize their data, lawsuits soon follow. Netflix, the famous on-demand movie streaming company, was sued and had to pay approximately $9 million for the exposure of their customers' data [1]. This began when they held a competition in 2010, to see if there could be an improved way to produce "better recommendations than the movie giant could serve up themselves". This led to two contestants and privacy researchers, Arvind Narayana and Vitaly Shmatikov, to realize that Netflix's data was not secure. They "showed that a second set of information such as comments on the popular Internet Movie Database could help third parties triangulate the identity of the 'anonymous' Netflix customers" [1]. They realized that by using both the information publicly available to them from Netflix, and another well-known public movie site, they were able to de-anonymize Netflix's customers. Millions of customers were affected by this because it was made possible to figure out their sexual preferences based on the movies they watched [23]. Additionally, AOL faced a similar privacy scandal. The AOL researchers released at least three months' worth of query logs to a website that is publically accessible [6, 7]. Even though personal identifying attributes were not exposed for the just under a quarter of a million users, there was still enough data to discern an individual's identity. AOL ended up withdrawing the information, but not before many people had a chance to download it. This raw data and even movie preferences turned out to be sensitive information, which is why there are various methods to protect such data [14, 19, 20].

Section 1.2: Sensitive and Non-Sensitive Information

Sensitive data is defined as "any data that compromises with respect to confidentiality, integrity, and/or availability that could adversely affect COV interests, the conduct of Agency programs, or the privacy to which individuals are entitled" [25]. In other words, this is information that one would not normally know from just looking at a person. This includes personally identifiable information such as: financial transactions, social security numbers, medical history, criminal records, server names, or IP addresses [25]. This information is constantly at risk during everyday activities, from purchasing a cup of coffee with a credit card, to attending a university. Credit card information must be exchanged and sensitive academic records are collected at universities. This information is highly confidential mainly because it can lead to identity theft. On the other hand, "data that are intentionally made public are classified as not sensitive" [21]. If information is voluntary given or if it is a physical characteristic, the data will not be considered dangerous. However, it is possible to decipher sensitive information about a specific person even if only non-sensitive data is given. This will be explained using simulated academic and health records.

Section 1.3: Data Anonymization

Data anonymization is the process of removing personally identifiable information (PII) from sets of data, so that the people who are described by the information can remain unknown [26]. Data is becoming more accessible over the internet, but most of the data has been limited and the personal identifiable information has been removed. This includes social security numbers or names paired with addresses. This anonymized data assurance protects the individuals' information and ensures their privacy. This also allows the government to share limited sets of data without acquiring legal permission. This data is useful for researchers who analyze data, particularly in health care. The major issue is that even though the data has been limited and the

PII's have been removed, is it still possible to de-anonymize the information and have individual's information exposed?

2. Privacy Laws

Section 2.1: FERPA

Currently, data privacy is being regulated by the United States government through the Family Educational Rights and Privacy Act [2]. FERPA is a Federal law that "prohibits a school from disclosing personally identifiable information about students' education records without the consent of a parent or eligible student, unless an exception to FERPA's general consent rules applies" [2]. If any educational institution receives funds under any program run by the Department of Education, they are subject to FERPA. However, since private schools do not receive this funding, the protection rules do not apply to them. Even if a student is over the age of 18, a parent or legal guardian can still have access to their academic records because the schools' data cannot be regulated by FERPA. For schools who do fall under that category, generally a student or parent would have to provide written consent to the school before the student's education records could be disclosed. Many students going into college are about 18 years old, which makes them a legal adult. Even though a student is at least 18 and paying for his/her own university, a guardian can still have access to his/her personal academic information if the guardian claims the student as a dependent [2]. Regardless of the student's want to keep his/her personal academic information private, FERPA allows such a loophole to exist. If the parent or guardian checks off the dependent box when they file their taxes, the student is extremely susceptible to their privacy being jeopardized.

Section 2.2: HIPAA

While FERPA protects student's data, the Health Insurance Portability and Accountability Act (HIPAA) protects people's medical data [5]. The rule "protects most individually identifiable health information held or transmitted by a covered entity or its business associate, in any form or medium, whether electronic, on paper, or oral" [5]. In other words, any information regarding an individual's health condition cannot be exposed to others. HIPAA includes a de-identification standard; health information that does not detect a specific person or if there is no rational basis to believe that the information can be used to identify that individual, then it is not individually identifiable health information.

There are two methods to achieve de-identification in accordance with the HIPAA Privacy Rule: expert determination and safe harbor [5]. The safe harbor method removes 18 types of identifiers such as names, all geographic subdivisions smaller than a state, telephones numbers, fax numbers, and biometric identifiers [5]. This prevents any actual knowledge residual information from identifying an individual. Unfortunately, de-identifying to this extreme leads to information loss and thus less useful information. It would not be able to be used for statistical analysis or for research purposes. On the other hand, the expert determination method is done by a hired professional. This expert generally has appropriate knowledge and experience with statistical data as well as how to apply such principles and methods. With this method, there is a very small risk that anticipated recipients could identify the individual. Experts determine, case by case, what information can be publicly displayed and what information should be suppressed or deleted.

Section 2.3: Example of HIPAA Expert Determination Method

Table 1: Identifiable and Sensitive Data

Age (Years)	Gender	Zip Code	Social Security	Diagnosis
45	Female	13453	738493728	Breast Cancer
23	Female	34573	516274857	HIV
14	Male	12345	732536468	Diabetes
56	Male	04564	019283746	Influenza

The table above shows five different categories: age, gender, zip code, social security, and diagnosis. The non-sensitive data is the age, gender, and zip code, while the sensitive data is the social security number and the diagnosis. I am going to show how experts determine what information to suppress or anonymize using the expert determination method. Only one category, social security number, can pinpoint a single individual, so it should be immediately taken off. Now there is an age in years, gender, zip code, and diagnosis left. "It has been estimated that the combination of a patient's date of birth, gender, and five-digit zip code is unique for over 50% of residents in the United States. This means that over half of U.S. residents could be uniquely described just with these three data elements" [4]. Because of this statistic, it is extremely vital for the expert to continue to mitigate the risk. The zip codes would be lessened to just three digits, and even with a year of birth and gender, risk plummets to 0.04%. With risk still being too high, the age in years could be more generalized. The ages can be 'below 21', 'between 21 and 34', and so on. The expert would then use his/her judgment and evaluate if even gender, with the other attributes, could still uniquely identify an individual. When an expert evaluates each person's information, there is always a chance that the expert could make a mistake simply due to human error. Their error has been classified as 'very small' and there is not an explicit numerical value to associate it with. The environment and context when identifying risk may not be appropriate for

the same data. Thus, it is up to an expert to deem the 'very small' risk; it is solely up to their discretion. After evaluating the specific situations, the end result could look like this:

Table 2: Anonymized Data

Age (Years)	**Zip Code**	**Diagnosis**
45 and over	134	Breast Cancer
Between 21 and 34	345	HIV
Under 21	123	Diabetes
45 and over	045	Influenza

3. Anonymization of Academic Records

Similar to determining personal information from health care information, it is also possible to determine such information from academic data. I created a simulation of data that is completely randomized. Using the random generator function, I associated each number with its corresponding letter grade. For example an 'A+' was a '1', an 'A' was a '2', an 'A-'was a '3' and so on. I paired the most common American names with each student and also randomly generated if they were from out of state or in state (note that the table is shown on the next page).

This hypothetical data set is similar to ones you would see as a professor or teacher. This information is extremely personal, as it shows who the person is and his/her grades in every single class. This is untouched and raw information, meaning that before it is released to the public or collected by the government, it needs to be anonymized. According to the Freedom and Information Act, information should be able to be accessible by all citizens, thus it needs to be publishable [28].

The raw data makes it readily available for personal identifiable information to be released. Similar to the expert determination method from the Health Insurance Portability and Accountability Act (HIPAA), making the information general and anonymous is key [5]. I took just the first 14 students to show some changes after I generalized the information. Instead of

having names of the students, I changed the data so that it showed an “M” for male, and an “F” for female.

Table 3: Hypothetical Student’s Grades

Students	In/Out of State	Grade 1 for Course 1	Grade 2 for Course 2	Grade 3 for Course 3	Grade 4 for Course 4	Grade 5 for Course 5
Noah	In	F	D	A	F	B
Sophia	In	A	B-	C+	D	D-
Liam	In	C-	C+	B-	A+	A
Emma	In	A-	C+	F	A	A-
Jacob	In	D	C+	B-	A+	C
Olivia	In	C	D-	C-	C+	C
Mason	Out	A+	B	A	B+	D+
Isabella	Out	D	B	D	D+	C
Ethan	In	C+	A	B-	B	A-
Ava	In	C	A	A-	C+	A-
Michael	Out	C+	A	C+	C+	D-
Mia	In	B-	D+	C	A-	A-
Alexandar	In	A-	B	C-	C+	A-
Emily	In	A	B+	F	B	D-
Jayden	In	D	A-	A-	B	A
Abigail	In	D-	C+	C+	D-	C-
Daniel	Out	A	D+	B	A+	C+
Madison	In	A+	B	A-	C+	F
Elijah	Out	C-	A+	B	B+	F
Elizabeth	Out	B+	D	A+	D+	C-
Aiden	In	B	B-	A-	C	B-
Charlotte	Out	A	B	B-	A+	B+
James	Out	D-	A-	B+	A	A+
Avery	In	A+	C+	A	C	D
Benjamin	Out	C	B-	D-	D-	D
Sofia	Out	B+	A+	C+	D-	C-
Matthew	In	A	A+	D+	A-	B-
Chloe	In	A-	B+	F	D	C
Jackson	Out	C	C+	B-	C+	A
Ella	In	C	C-	D	A	D+
Logan	Out	B-	A-	B+	C+	B+
Harper	In	A-	B+	B+	C	D-
David	Out	C	B+	B	A	B+
Amelia	In	B	B+	B-	C+	D-
Anthony	Out	A	F	A	B+	A-
Aubrey	In	A	C-	D+	B-	A
Joseph	In	C	B	B	B-	D+
Addison	Out	A	F	F	A-	B+
Joshua	Out	A-	A+	C+	B	B
Evelyn	In	B+	B-	A+	C+	C+
Andrew	In	A	B	D	F	D-
Natalie	In	A	A+	C	D-	B-
Lucas	In	D-	D+	A+	C+	B
Grace	In	A-	A	B	D+	F
Gabriel	Out	A	A-	B+	A+	C+
Hannach	Out	A-	C	A+	C-	A+
Samuel	In	A-	D-	D+	F	F
Zoey	In	B	B+	A+	C	D-
Christopher	In	D-	A-	A-	D	D-
Victoria	Out	D+	B	B+	A+	C+

Table 4: 1st Step in Anonymizing Data

Students	In/Out of State	Grade 1 for Course 1	Grade 2 for Course 2	Grade 3 for Course 3	Grade 4 for Course 4	Grade 5 for Course 5
M	In	F	D	A	F	B
F	In	A	B	C	D	D
M	In	C	C	B	A	A
F	In	A	C	F	A	A
M	In	D	C	B	A	C
F	In	C	D	C	C	C
M	Out	A	B	A	B	D
F	Out	D	B	D	D	C
M	In	C	A	B	B	A
F	In	C	A	A	C	A
M	Out	C	A	C	C	D
F	In	B	D	C	A	A
M	In	A	B	C	C	A
F	In	A	B	F	B	D

Additionally, I made the letter grades more general, without losing the integrity of the information. For example, all A-‘s and A+’s, were changed to A’s; all B-’s and B+’s were changed to B’s. I kept the in and out of state column, but it can be argued that it is too specific, even if it does not say what state/country specifically the person comes from.

If concerns are still present, I took the next step in anonymizing the data (note that the table is shown on the next page). Going off from the last step, I took the A and B grades and changed them to be higher than a B (B+), the C and D grades and changed them to D+, and kept the F’s. Making the information more general reduces the possibility of someone being able to link the information back to a specific person, however it greatly reduces the usefulness. Companies or governments that analyze data would get more use out of the raw data, because finding patterns would be significantly easier.

Table 5: Final Anonymization of Data

Grade 1 for Course 1	Grade 2 for Course 2	Grade 3 for Course 3	Grade 4 for Course 4	Grade 5 for Course 5
F	D+	B+	F	B+
B+	B+	D+	D+	D+
D+	D+	B+	B+	B+
B+	D+	F	B+	B+
D+	D+	B+	B+	D+
D+	D+	D+	D+	D+
B+	B+	B+	B+	D+
D+	B+	D+	D+	D+
D+	B+	B+	B+	B+
D+	B+	B+	D+	B+
D+	B+	D+	D+	D+
B+	D+	D+	B+	B+
B+	B+	D+	D+	B+

Section 3.1: Exceptions

The entire idea behind anonymizing data is to limit the probability of another person being able to link sensitive data to a specific person. Privacy needs to be upheld, but there are sometimes exceptions even with proper data anonymization. For instance in the previous example, Table 5, the data went through several steps of removing linkage and leakage. It seems reasonably safe to be published and the integrity of the data has not been altered to a significant extent. But looking more closely into it, there are only three F's in Table 5, making linking extremely easy. Hypothetically, if I was in course three with my friend Michelle, and I was very aware she did not study for any test, it is safe to assume she received an F. From this assumption, since Michelle is the only student with an F in course three, her other grades in all of her courses can be found. In other words, there is always going to be outliers in data anonymization. It might not always be this extreme and able to be narrowed down to one person, but it is still an exception to acknowledge. Overall, privacy has been improved but it is not guaranteed for every individual, and in this specific case, every student.

4. Attaining Full Data Anonymization

Section 4.1: How Privacy can be Protected

Privacy can be protected overall by reducing the probability of data linkage from something high, to something much lower. Optimizing the utility regarding privacy is important and can be defined as a "service to satisfy some human want" [27]. As previously stated, the combination of a patient's birthday, gender, and zip code is unique for at least half of the people in the United States [4]. From this statistic, there is an extremely high probability that someone can figure out sensitive information by knowing this data. The whole idea of privacy protection is to mitigate this risk. When the zip code given is three digits, the year of birth is presented, and the gender is shown, the risk of linkage drastically drops to 0.04% [4]. This significant drop is a great improvement, and the entire idea behind privacy protection. With any data relevant to a single person, the goal is to make the linkage statistic to as close to 0.00% as physically possible. Even though there are drawbacks, knowing this information is limited but overall utility can be attained.

Section 4.2: K-Anonymity

K-Anonymity has been commonly used to anonymize data [26]. It can be defined as where attributes are suppressed or generalized until each row is identical with at least k-1 other rows. The bigger the k is the better; that means there are more similar records. This makes it harder to differentiate between individual records. This prevents data linkages and in the worst case scenario, it can narrow down data to a group of k [26]. There are two common ways to achieve k-anonymity: suppression and generalization. Suppression is when certain values of the attributes are left purposely blank or only an asterisk is present. For example, if there is an attribute called "Name", the contents would be replaced with an * to prevent linkage. Instead of the data showing the name "Mary", the cell would simply show "*". Generalization is when individual values of attributes are replaced with a less specific category. For example, within an attribute of "Age", instead of an

age of “21” being shown, it would be changed from age “20 to 30”. Utility is sacrificed in both of these examples but it must be completed to retain privacy. With every anonymization method, there are various flaws that people can still utilize to attain personal information.

Section 4.3: Attacks on k-Anonymity

There are two major attacks on k-Anonymity protecting privacy: the homogeneity attack and the background knowledge attack [21] [3]. The homogeneity attack can take place if the sensitive values in an attribute lack diversity. Referencing the table below, with big data there could be instances where the information is the same. Three different people who live in the same area code and who are no more than ten years apart from each other all have malaria. If the diseases were different, then it would be impossible to link a person with a disease, but since the diseases are the same, the k-Anonymity attempt it flawed. If I knew someone in my town with the same first three letters of the zip code and who was 67, it would be safe to assume that that person has malaria. Usually data is significantly bigger than Table 6, however it is possible to narrow down this information by just knowing basic information about the specific individual.

Table 6: Healthcare Data

Zip Code	Age	Disease
110**	6*	Malaria
110**	6*	Malaria
110**	6*	Malaria
111**	7*	Heart Disease
111**	7*	Heart Disease
111**	5*	Cancer

Another attack on k-anonymity is the background knowledge attack [21] [3]. This occurs when the attacker knows some information about the person he/she is trying to gain more knowledge on. “K-anonymity does not prevent against other attacks in which an adversary equipped with public knowledge can associate fewer than k transactions with an individual” [10].

For instance, if you want to know what illness is preventing your boss from coming to work, you already have information on that person. If you have been to his/her house before, then you already know that the zip code is 11124. Additionally, through a previous casual conversation you find out that your boss is 71. With this background knowledge that did not take much effort to find out, the only option is that your boss has a heart disease. The ability to link already known information with publicly available information is a huge flaw, and it is a concern when using k-anonymity.

5. Advanced Privacy Notion for Statistical Information Disclosure

Section 5.1: Differential Privacy

"Differential privacy is a rigorous notion of privacy that allows statistical analysis of sensitive data while providing strong privacy guarantees even in the presence of an adversary armed with arbitrary auxiliary information" [3]. Overall, differential privacy is a statistical form of protecting data even when an attacker with outside knowledge tries to leak it. It can be explained by utilizing: $Pr[A(D_1) \in S] \leqslant e^{\in} \times Pr[A(D_2) \in S]$, which ensures that even two different records, when anonymized, have similar if not the same outcome. Laplacian noise is a major randomized algorithm that is used to satisfy differential privacy. There is a tradeoff between the accuracy of the information and the utility. Applying the noise to attain differential privacy is a sacrifice used to retain utility for personal privacy.

Section 5.2: Laplacian Noise

Another way to try to preserve the anonymity of data is to add 'noise' to the information [3]. The raw information is presented below. It shows the courses that each student took from the original data, and it shows how many times the students received a specific letter grade.

Table 7: Sum of Student's Grades

Courses	Class Equivilant	A+'s	A's	A-'s	B+'s	B's	B-'s	C+'s	C's	C-'s	D+'s	D's	D-'s	F's
Course 1	Digitial Forensics 100	11	14	12	4	14	4	4	11	5	2	6	6	7
Course 2	Nutrition 103	6	12	9	11	11	6	10	5	6	5	10	7	2
Course 3	Finance 111	9	9	8	8	6	13	13	5	6	8	4	4	7
Course 4	Marketing 101	14	9	7	5	7	8	17	9	5	5	3	4	7
Course 5	Computer Science 122	5	5	7	6	8	8	10	6	9	8	4	15	9

This information is not only important to students wanting to know where they stand in a class, but it is also important to academic officials. The data answers many questions such as are the professors too hard, is the class too easy, should we fire this specific professor, and so on. This raw data cannot be released because students can use background knowledge to link who received what grades in their classes. In order to anonymize the data, noise can be added to the numbers. This is done by using C + Lap (d/$\in$) [3]. The "d" represents the maximum sensitivity between sets; if the d is large, the amount of noise is large which ensures the same level of privacy guaranteed. $e^{\in}$ shows the probability of how close together the sets are to each other. To retain privacy, a smaller number would be more ideal because that proves that individuals have very similar attributes and cannot be distinguished from one another. Noise can be the same statistical number applied to each entry. This noise is random and can prevent attackers from linking the data, while still preserving its utility.

Table 8 below shows the anonymized data using the epsilon equal to 1.5. To add some noise to the data, I took each number and added it to Lap(d/Ɛ). For example, in course one in digital forensics 100, the amount of A+'s is 11; 11+ Lap(1/ Ɛ) is equivalent to 12 where Ɛ is equal to 1.5. This does not lose the integrity of the data as many other methods do so; the data is difficult to decipher once altered. Any random number that does not affect the integrity of the data can be

used. Even though the raw data would be more useful in predictions and for studying, it still allows for the significant information to be valid.

Table 8: Anonymized Data with Epsilon = 1.5

Courses	Class Equivilant	A+'s	A's	A-'s	B+'s	B's	B-'s	C+'s	C's	C-'s	D+'s	D's	D-'s	F's
Course 1	Digitial Forensics 100	12	14	12	4	13	5	3	11	6	3	5	6	8
Course 2	Nutrition 103	7	11	9	12	11	8	9	4	6	5	11	6	2
Course 3	Finance 111	9	10	9	8	6	13	12	5	6	9	5	5	8
Course 4	Marketing 101	12	8	7	6	6	8	15	9	4	4	4	5	8
Course 5	Computer Science 122	5	5	9	8	5	8	11	6	11	8	5	17	8

6. Future Work

It was just shown that in any anonymization case there can be outliers, and the best course of action is to mitigate the existing ones. There are some outliers that cannot be protected, thus new technology is constantly being created. Aside from k-anonymity, there have been other ways to combat the linkage of sensitive data. “To address this privacy threat, one solution would be to employ L-diversity: a well-established paradigm in relational data privacy, which prevents sensitive attribute (i.e. item) disclosure” [25]. Generalization and permutation-based L-diversity are the only two categories that exist. With the generalization category, which is similar to the generalizing information using k-anonymity, unique data would be analyzed. This unique data would either be combined with other persons within the same category, making the probability of linking limited, or removing the category if too unique. When the latter is used “it is likely that any generalization method would incur extremely high information loss, rendering the data useless” [25]. On the other hand, “a permutation method such as Anatomy would randomly pick groups of transactions with distinct sensitive items, and permute these items among transactions, to reduce the association probability between an individual transaction and a particular sensitive item” [25].

In other words, the sensitive information in each data set would be randomly distributed throughout the transaction. This permutation method unfortunately limits the amount of useful information that can be extracted.

Additionally, there are other privacy preserving techniques, used in other contexts such as social networks [9] and supply chain management [11-13] [15-18]. In the real world, collaboration of data between different parties for a business transaction is extremely necessary [25]. Due to their collaboration, proprietary information is shared, thus increasing the chances that the information could be exposed to the public. There are distributed optimization problems that can be mitigated for "privacy-preserving horizontally partitioned linear program with arbitrary number of equality and inequality constraints" [12].

7. Conclusion

Personal privacy cannot be completely guaranteed in any aspect in regards to online data or data collected from industries. Even HIPAA and FERPA admit to not being able to 100% ensure data anonymity [5]. The more ways data is manipulated to ensure privacy, the less useful it becomes to researchers; a common ground needs to be achieved. K-Anonymity and adding noise to data is effective, but unfortunately they both have flaws. The homogeneity attack and the background knowledge attack prevent k-Anonymity from being utilized to the full extent due to the possibility of linkage. Adding noise to the data is another useful way to attempt to achieve full data confidentiality, but then again data integrity is compromised.

8. References

[1] Taylor Buley. "Netflix Settles Privacy Lawsuit, Cancels Prize Sequel." *Forbes*. Forbes Magazine, 12 Mar. 2010. Web. 20 Feb. 2015.

[2] Campbell, Ellen. Family Educational Rights and Privacy Act (FERPA) and the Disclosure of Student Information Related to Emergencies and Disasters June 2010 (PDF) (n.d.): n.pag. Family Educational Rights and Privacy Act (FERPA) and the Disclosure of Student Information Related to Emergencies and Disasters. June 2010. Web. 17 Dec. 2014.

[3] Cynthia Dwork. "Differential Privacy", *ICALP (2)* 2006: 1-12.

[4] "FPCO Frequently Asked Questions." FPCO Frequently Asked Questions. N.p., 14 July 2005. Web. 07 Jan. 2015.

[5] "Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule." *Health Information Privacy*. N.p., n.d. Web. 07 Jan. 2015.

[6] Katie Hafner, "Researchers Yearn to Use AOL Logs, but They Hesitate", the *New York Times*, 22 Aug. 2006. Web. 19 Apr. 2015.

[7] Katie Hafner, "Tempting Data, Privacy Concerns; Researchers Yearn To Use AOL Logs, But They Hesitate", the *New York Times*, 23 Aug. 2006. Web. 19 Apr. 2015.

[8] Moritz Hardt and Kunal Talwar. "On the Geometry of Differential Privacy", (n.d.): n. pag. 9 Nov. 2009. Web. 2 Feb. 2015.

[9] Michael Hay, Chao Li, Gerome Miklau, David Jensen. "Accurate Estimation of the Degree Distribution of Private Networks". *ICDM 2009*: 169-178.

[10] Yeye He, and Jeffrey F. Naughton. "Anonymization of set-valued data via top-down, local generalization", *PVLDB*, 2(1):934–945, 2009.

[11] Yuan Hong. "Privacy-preserving Collaborative Optimization". *Ph.D. Dissertation*, Rutgers University, 2013.

[12] Yuan Hong, and Jaideep Vaidya. "An Inference-proof Approach to Privacy-preserving Horizontally Partitioned Linear Programs". *Optimization Letters*, Vol. 8(1), pp. 267-277, 2014.

[13] Yuan Hong, Jaideep Vaidya, Lu, and Lingyu Wang. "Collaboratively Solving the Traveling Salesman Problem with Limited Disclosure". In *DBSEC*, pages 179–194, 2014.

[14] Yuan Hong, Jaideep Vaidya, Haibing Lu, and Mingrui Wu. "Differentially private search log sanitization with optimal output utility". In *EDBT*, pages 50–61, 2012.

[15] Yuan Hong, Jaideep Vaidya, and Haibing Lu. “Secure and Efficient Distributed Linear Programming”. *Journal of Computer Security*, 20(5):583–634, 2012.

[16] Yuan Hong, Jaideep Vaidya, and Shengbin Wang. “A Survey of Privacy-Aware Supply Chain Collaboration: From Theory to Applications”. *Journal of Information Systems*, 28(1):243–268, American Accounting Association, 2014.

[17] Yuan Hong, Jaideep Vaidya, Haibing Lu, and Basit Shafiq. “Privacy-preserving Tabu Search for Distributed Graph Coloring”. In *PASSAT*, pages 951–958, 2011.

[18] Yuan Hong, Jaideep Vaidya, and Haibing Lu. “Efficient Distributed Linear Programming with Limited Disclosure”. In *DBSEC*, pages 179–194, 2014.

[19] Yuan Hong, Jaideep Vaidya, Haibing Lu, Panagiotis Karras, and Sanjay Goel. “Collaborative Search Log Sanitization: Toward Differential Privacy and Boosted Utility”. *IEEE Transactions on Dependable and Secure Computing*, to Appear.

[20] Yuan Hong, Xiaoyun He, Jaideep Vaidya, Nabil Adam, and Vijayalakshmi Atluri. “Effective anonymization of query logs”. In *CIKM*, pages 1465–1468, 2009.

[21] “Information Policy, U.Va". Administrative Data Usage Policy: Appendix A,. N.p., 17 Sept. 2013. Web. 15 Dec. 2014.

[22] Aleksandra Korolova, Krishnaram Kenthapadi, Nina Mishra, and Alexandros Ntoulas. “Releasing search queries and clicks privately”. In *WWW*, pages 171–180, 2009.

[23] Frank McSherry, and Ilya Mironov. “Differentially Private Recommender Systems: Building Privacy into the Netflix Prize Contenders”. In *KDD*, pages 627–636, 2009.

[24] Adam Moore. “Defining Privacy.” *Journal of Social Philosophy*, 39.3 (2008): 411-28.

[25] “Sensitive Data”. Types of Sensitive Data. N.p., n.d. Web. 01 Jan. 2015.

[26] Latanya Sweeney. “K-Anonymity: A Model For Protecting Privacy”, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5): 557-570 (2002)

[27] “Utility”. Dictionary.com. *Dictionary.com*, 2015. Web. 17 Feb. 2015.

[28] “What Is FOIA?” FOIA.gov. *United States Department of Justice*, 2015. Web. 19 Apr. 2015.