Giving to Get: Can Your Salary Stay Secret?

W233 Privacy Engineering
Final Project | Spring 2021 | April 17, 2021
Angela Gao | Alice Hua | Lesley Matheson | Dhyani Parekh

1 Introduction

The explosive growth in the amount of personally identifiable information (PII) available on individuals from different sources has wreaked havoc on personal privacy. The combination of data holders, publishing anonymized databases and data brokers publishing PII outright, has placed individuals in anonymized databases with sensitive attributes at risk of re-identification. This re-identification allows sensitive information to be imputed to a specific individual.

In this study, a database of employee compensation is examined to determine whether a specific individual in the database can be identified. Specifically, the study seeks to determine how and to what extent individuals can be identified when seemingly benign information such as location, company, job title, years of service, gender and race are provided along with salary and bonus compensation data. Using and joining additional databases, the study found that quasi-identifiers in the Levels dataset can be leveraged to perform somewhat successful identity disclosure attacks and suggests a promising protection measure.

2 Background

2.1 Job Compensation Information: A Privacy Utility Trade-Off

While the notion of compensation transparency has gained momentum in recent years [1], compensation packages are overwhelmingly still viewed as culturally sensitive and private. The vagaries of salary distributions and stock bonus manifestations infuse compensation information with a mystique and stigma that still runs deep in most modern societies. While individuals almost universally want compensation data about themselves to remain private, both companies and individuals seek important market research answers around competing compensation packages. For example, individuals seeking a new position or negotiating for a compensation package seek to better understand the market for their skills and experience. Understanding the market for a particular position, at a particular company in a specific geographic location is valuable. As such individuals have been willing to share their personal information in order to gain access to anonymized information from others. These data sharing sites such as Glassdoor and Levels FYI have gained popularity because of their utility in this and other use cases. But as a premise, they rely on maintaining participant privacy to obtain data. This dynamic sets up the privacy utility trade-off and inherent risks around re-identification. The research and history below reviews briefly the state of re-identification risks and specifically the growing risks and issues around re-identification exposing employment compensation.

2.2 Survey of Related Work

Sweeney [2] in her seminal work on k-Anonymity develops a privacy framework that provides a metric for the degree of anonymity inherent in a dataset with a given set of quasi-identifiers, as opposed to identifiers. In this paper a method of re-identifying an anonymized data set published by the Group Insurance Commission in order to obtain quotes for the purchase of insurance for the employees of the State of Massachusetts. This GIC dataset contained sensitive attributes such as medical procedures and diagnosis. Sweeney purchased the Cambridge, MA Voters Registration database. Using only three quasi-identifiers, Sweeney was able to identify are substantial proportions of the GIC entries, inclduing the Governer of MA at the time Willioam Weld, with only a five digits zipcode, birthdate and sex. Sweeney suggested that this form of attack, re-identification by linking would become a considerable threat to privacy and in the twenty years since the papers was published indeed such re-linking or join attacks have become widespread.

In a fairly recent survey published on re-identification attacks, Henriksen-Bulmer and Jeary [3] found that the majority of reidentification attacks use the popular search engines, (Google, Yahoo). In the majority of the attacks only a few quasi-identifiers, approximately three, were needed to re-identify. The study also suggested that GIS data is being used to re-identify in approximately one-third of the attacks [4].

While Sweeney's papers introduced a deterministic framework for privacy and re-identification, it is worth noting that in the medical sciences probabilistic linking and re-identification has been the subject of study for several decades. In a seminal paper published in 1995 Jaro[5] published an Expectation Maximization algorithm for linking large databases, not with specific deterministic re-identification but with probabilistic methods and bounds around a re-identification. Probabilistic methods for re-identification on very large datasets continue to make computational and accuracy progress with contributions such as Ferguson[6]. While these methods are outside the strict privacy models suggested by Sweeney, they demonstrate an alternative means of exposure for individuals and their sensitive attributes.

3 Methodologies

3.1 Levels FYI Database

The first database used in the study is the public Levels FYI (Levels) database https://www.levels.fyi/. This database ingests user-supplied compensation information, anonymized. The following fields are quasi-identifiers: Company, Title, Tag/Focus, Years Experience, Years at Company, Location, Gender, Race, and Education. The sensitive attributes are: Total Annual Salary, Base Salary, Stock Grant, and Average Annual Bonus. Initially, the study used only California information which resulted in 3,797 records out of the total 45,480 records.

3.2 Zoom Database

The second database, referred to here as Zoom, was constructed by scraping unstructured public information across over 28 million sites for verified business contacts. This database is a publicly available site for sales prospecting. While the Zoom database contains the following identifiers and quasi-identifiers: Employee Name, Employee Title, Work Email, Employee Phone Number, Company Name, Headquarter State, Employee Country. It contains approximately 635,000 individuals located in California. This subset was used in this study.

3.3 Levels and Zoom Join Re-identification Attack

An initial proof of concept join attack join on the California subsets of the Zoom and Levels database. This subset was chosen because the levels database was heavily populated with technical positions, many of which are located in California. If time had permitted smaller states with a more general distribution could have been explored.

The initial join was performed on as follows. In the Levels database location, area code was inferred from city and state (CA) with an external library. The databases were joined on Company, Title and Area Code. This join produced a one to one matching for 163 table entries. Once one-to-one matches were acquired the numbers were manually verified using the remaining quasi-identifiers, including Years at Company, Years of Experience, Gender, Race, and Level of Education through Google and Linked in searching. While most of the 163 matches were very close, variations in Years and Company and Years of Experience, were less precise. At least 2 of the 1:1 matches satisfied all quasi-identifiers. More results are presented in Section 4.

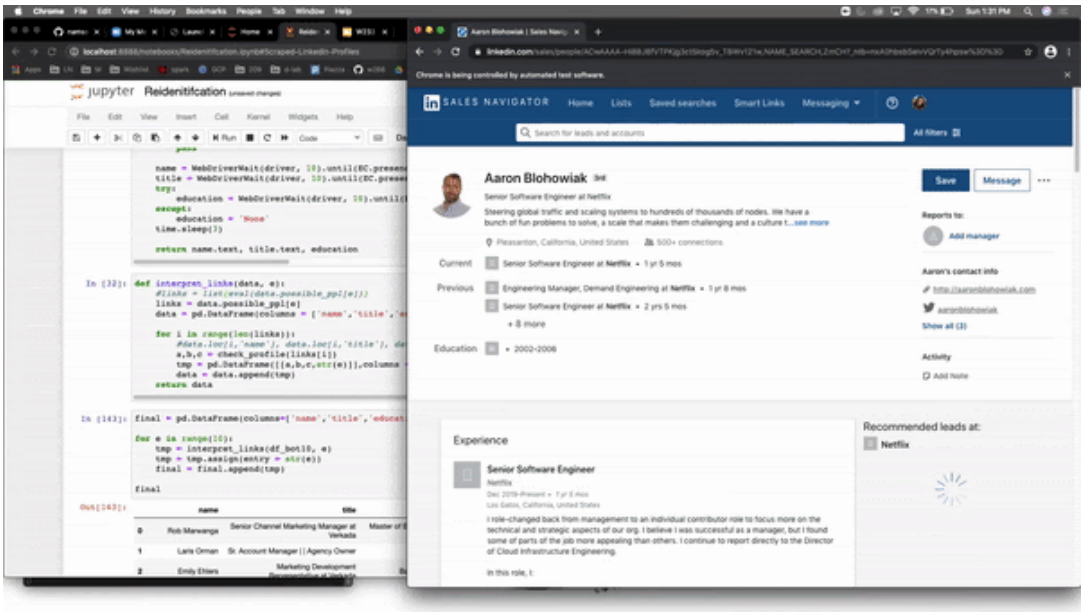
3.3 Re-identification of Levels FYI with LinkedIn Profiles: General Approach

The second approach taken to re-identify the Levels database participants was to use automated tools to crawl the LinkedIn website and automatically search for LinkedIn profiles that match the quasi-identifiers in Levels. The fields Company, Title, General Location and Tag from Levels were used to search LinkedIn profiles that matched the quasi-identifiers. The resulting output was a list of LinkedIn profiles for each of the Levels participants. In order to focus the search of potential candidates, only those rows (from Levels database) with a maximum 21 potential candidates were considered for further re-identification efforts. The list of potential candidates for each row now included full names. A classification package, Namsor, was used to classify race and gender for the given LinkedIn profile names. The resulting augmented Linked In profiles were re-joined on the Levels table using Company, Title, Gender, Race, Education, General Location, Tag.

3.4 Re-identification of Levels FYI with LinkedIn Profiles: Detailed Proof of Concept

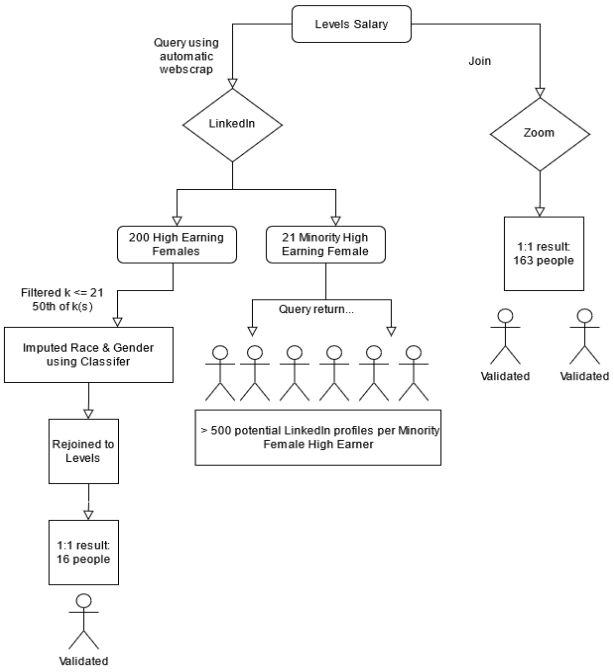
The Levels database was filtered for records within California and females only. Within the Levels database, only the companies with fewer reported entries were considered, the companies below the median in terms of number of Levels records in quantity reported. Of this Levels subset the top 200 earning females were chosen. An automated bot then queried LinkedIn on the following: Company, Title, Tag, and Location (California, US). The results are LinkedIn profile links of potential matches for each person in the Levels database. An animation of the bot developed is shown below in Figure 1 below.

Figure 1: Automated LinkedIn Search Bot



The motivation behind choosing top female earners at smaller companies (here treated as underreported companies) was to narrow the search to potentially smaller equivalence classes and perform the follow-on attacks on a more potentially vulnerable group. It is worth noting that a similar approach was taken with respect to minority females, not in smaller companies, but this was surprisingly not successful mainly because top earning minority females tended to work at very large companies which tended to protect their identity from such attacks. Figure 2 illustrates this workflow. A promising attack may have been to further subset Levels on these females on underreported companies.

Figure 2: Levels - Linked In Attack Flow Diagram



After collecting a series of potential matches for each Levels record, the next phase of the attack proceeded as follows. For each Levels entry with 21 or fewer potential Linked In profile matches which was the 50th percentile, a second round automated searching was employed. This second round used Full Name, Title, and Highest Degree of Education, Gender and Race (from Namsor). With augmented LinkedIn profiles including gender and race, they were joined again on the Levels leveraging the fuller set of Gender, Race, Highest Degree of Education, Company, Title, Tag, and General Location (California).

This join produced 16 one-to-one matches. These matches were then manually validated using auxiliary information, specifically, Years of Experience and Years at the Company. Many inconsistencies in these two pieces of information were found that would suggest that the accuracy of this auxiliary information is not high, both on the LinkedIn side and the Levels side. After manual

verification, as previously stated, we identified one individual on all of our quasi-identifiers.

3.5 Mondrian : A Proposed Protection Measure

In the two approaches taken for re-identification, the first a straight join attack using area code, which is basically general location, company and title, and the second an automated Linked In search using a broader set of quasi-identifiers in kind of a multi-step process, both relied heavily on location. One common result was that location was fairly significant in facilitating re-identification. To a lesser degree, gender, race and level of education. Location was also crucial to utility in this use case however. Thus generalizing location, and to a degree education level, had to strike the correct balance between utility and privacy. Generalization on the additional quasi-identifiers, race and gender do not impact utility in this use case.

The Mondrian generalization algorithm was chosen as a means to better understand the impact on the privacy utility trade off. The key aspects to Mondrian are choosing the appropriate quasi-identifiers to use for the generalization and the appropriate desired k to improve privacy but maintain utility. In this study multiple sets of quasi-identifiers were explored as well as a set of desired k values from k = 20 to k=200. The limited scope of the study led to focus primarily on generalizations of locations. These implementations are discussed below.

3.6 Mondrian Implemented on Levels FYI Database

The Mondrian implementation pursued in this study focused on foiling the Zoom join attack. The Levels table was prepared by assigning numerical values to the quasi-identifiers of race, gender, and education. Further, the Levels database was augmented with a zip code based on city (the state was restricted to California). Zip Code was chosen because its digits produce a natural hierarchy for location information.

The zipcodes were obtained for a city and state combination via uszipcode, an open source library that provides a search engine. This library is extremely slow so using it on a large table is prohibitive computationally. A cache was built for the cities in the Levels and Zoom databases to facilitate repetitive experimentation with privacy techniques. Multiple zip codes are associated with many cities. The lowest numerical zip code per city was chosen. This ordinal version of the Levels database was imported into a Mondrian implementation. The generalizations were obtained. The table below in Figure 3 is an example of the generalized zip codes created when Mondrian was run on zip codes as the only quasi-identifier and the desired k is 200.

Figure 3: Generalized Zip Codes

	zip_code		k
1	90001	91758	204
2	92008	92602	204
3	92626	94025	326
4	94040	94040	462
5	94061	94085	472
6	94101	94101	1127
7	94301	94582	306
8	94595	95014	292
9	95020	95050	260

3.7 Pre-Mondrian Baseline

A similar process of assigning zip codes to the Zoom database was followed. Using the area code of the “Employee Direct Phone” entry of the Zoom database the Zoom database was joined with a table listing all cities that are contained within that area code. There are 482 cities in California and 36 area codes or about 13 cities per area code. The result of this join produced a table with approximately 25M rows. These rows represent every possible city for a Zoom entry with a particular area code. Each record was then assigned a zip code, once again the lowest zip code per city.

To obtain a baseline measure to understand the efficacy of the Mondrian algorithm a join of the two tables prior to any generalizations was implemented. The Levels database appended with zip code was joined with the appended Zoom database on *Company*, *Title* and *Zip Code*.

3.8 Zoom Generalization: Preparation for Join of Generalized Tables

The Zoom database appended with zip codes was run through the generalizations produced on the Levels FYI database. These generalized zip codes were the only generalizations that time permitted. However, using the Namsor libraries to append race and gender to the Zoom database would enable generalizations produced by Mondrian on the Levels database to be appended. With generalized quasi-identifiers the approach would be to test a variety of generalizations in comparison to the baseline.

4 Results

4.1 Assumptions

Several assumptions are important to highlight. First, the verification of and true identity disclosure required reliance on self-reported data for *Years of Experience* and *Years at Company*. The study found a reasonably high degree of variability in this data when compared to Linked In profiles, the data most used to manually verify possible matches. Second, for both joins performed with the Zoom database, the study operated under the assumption that the verified business contacts' (data subjects') data was accurate at the time of the study. This assumption is probably reasonable given the frequency of the updates to that database. Third, in the initial Zoom join attack, pre-Mondrian, to infer the location of an individual in the Zoom database, the data subject's cell phone information was used. The cell phone numbers are personal numbers and not office numbers. As such, an assumption was made that the area code of the personal cell phone is the area code in which the person resides. However, people have a tendency to migrate with their cell phone number to areas other than then their current location. Thus the assumption in this join was that individuals reside in the location of their personal cell phone's area code. Fourth, in the Mondrian baseline and implementation the lowest zip code per city was chosen under the assumption that it is geographically close to all others within the city. An average or geocoding scheme could have been implemented as an alternative.

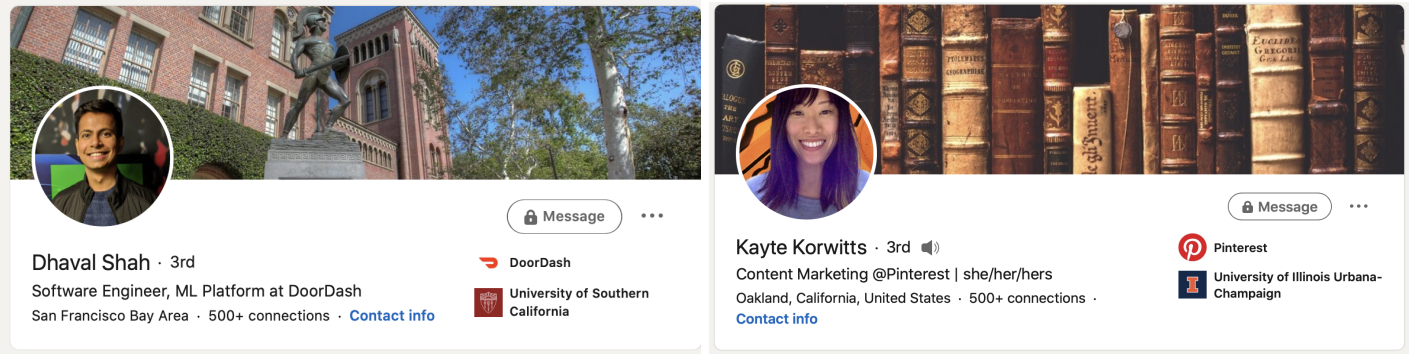
4.2 Zoom-Levels Join Attack Using Area Code for Location

The join of the Zoom and Levels databases (Methodology 3.3) yielded 163 1:1 potential matches. Each of these 1:1 matches were hand-validated for accuracy using public LinkedIn profiles. In order to be classified as a true match, every quasi-identifier had to match (first name, last name, company, level/title, years of experience, years at company, tag, race, degree, city, and gender). From the 163 potential matches, validation of each match found 2 true matches.

4.3 Levels FYI - Linked In Automated Search

Using the automated search tool on Linked In profiles (Methodology 3.4) and matching them with the Levels database yielded 15 potential 1:1 matches. Each of these 1:1 matches were hand validated for accuracy via public LinkedIn From the 15 potential matches. When these 15 matches were strictly cross-referenced with *Years at Company* and *Years of Experience* there was only one strict match. However, because of the degree of variability in these quasi-identifiers most, if not all of the matches could be verified if this strict matching requirement was relaxed by a year or two. Two examples of a manually verified match are presented below in Figure 4.

Figure 4: Levels - Linked In Attack Matches



The results of an automated attack on high earning, minority females was not successful. The motivation was to focus on a particular subset of the Levels database that would be a smaller class and be more vulnerable to re-identification. Most of the members of this class worked for large companies. When the initial link was performed with LinkedIn, the number of potential matches for all high earning minority females in California was in the range of 100-600. Unfortunately, the implementation of Namsor in the code base rejected the large batch and time did not permit an alternative implementation. Using non-Asian female minorities from the Levels database in California reduced the entries to six women., two of them are Native Americans. Namsor library does not contain classification for this *Race* category. However, these six women matched between 100-600 LinkedIn profiles without filtering them for *Race*.

4.4 Mondrian Protection for Levels Database

Mondrian was implemented on the Levels database with transformed gender, race, level of education and zip code quasi-identifiers (Methodology 3.5-3.8). In order to compare the efficacy of Mondrian a baseline table which was the result of a join between the Levels database and the Zoom database on *Company*, *Title,* and *Zip Code* . This table was used to create baseline statistics to which generalized tables could be compared.

One run of Mondrian using k = 200 (actual k = 204) on only zip code as a quasi identifier produced the 10 generalized zip code regions shown in Figure 2. When the generalized zip codes were used on both tables instead of the specific zip codes and the two tables were joined on the set of 10 generalized zip codes the joined table contained over 39,000 rows as opposed to the original join which produced approximately 3800 rows. The flow of this process is shown in Figure 5.

Figure 5: Join Attack With Generalization

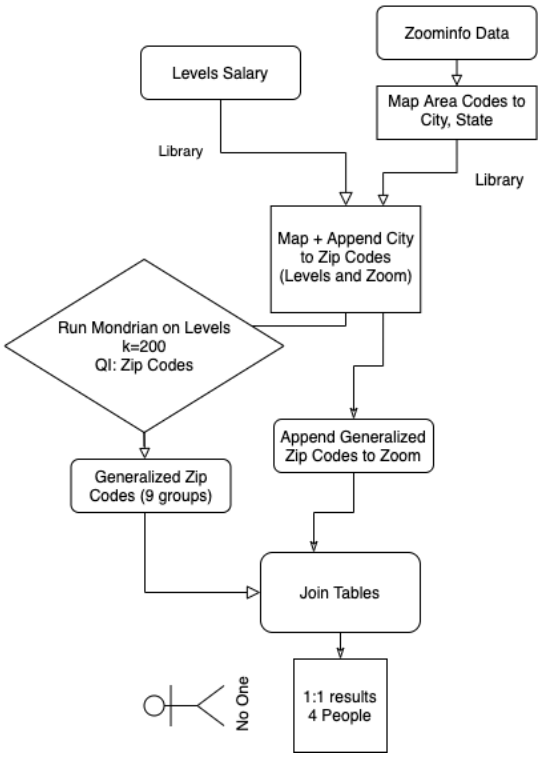
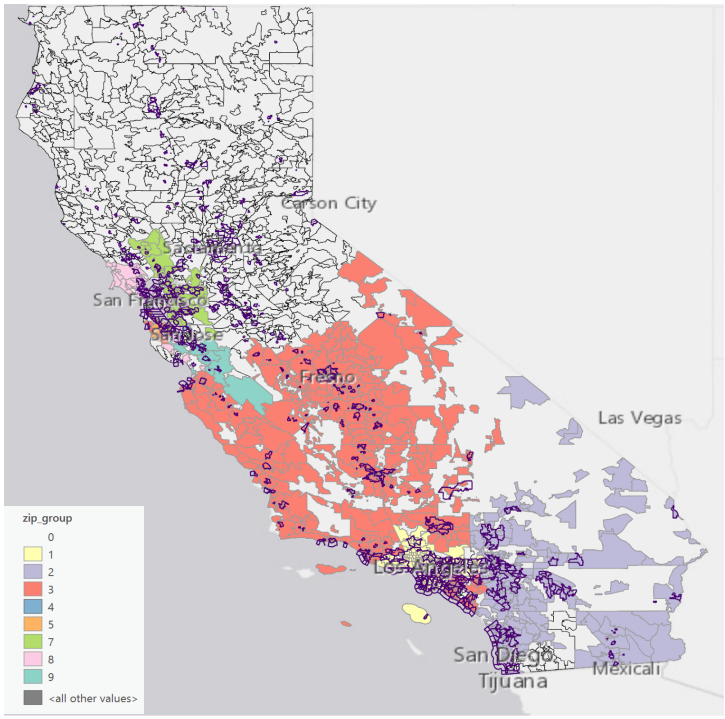
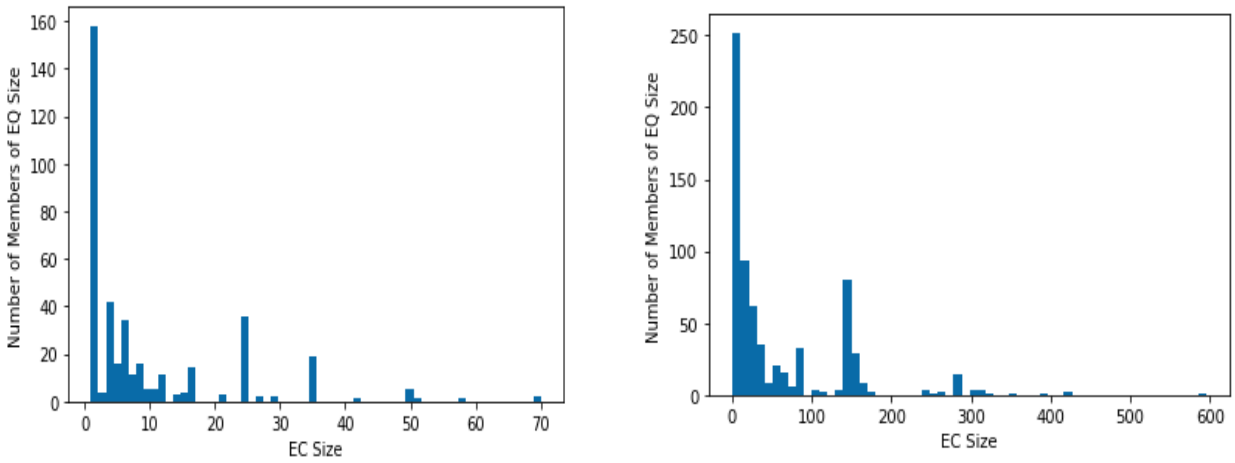


Figure 6: Generalized Zip Code Geographic Regions



The Levels FYI database was joined with the Zoom database on Company, Title and Zip Code. The k values of the equivalence classes resulting from the join before and after the generalization were distributed as follows in Figure 7 below.

Figure 7: k Values for Equivalence Classes Before and After Mondrian Generalization



A view of the impact of the generalization on privacy is shown in Figure 6. The generalization of the zip codes produces equivalence classes of an order of magnitude larger across the board. While in the strict k-anonymity sense k = 1 both before and after the generalization, viewing the impact in a broader sense yields some insights. For example, consider that before the generalization, there are 42 1:1 matches (one name per equivalence class) and 120 2:1 matches (two names per equivalence class). Thus, for 162 people in the Levels database there is a 50% chance or greater that their salary can be guessed or inferred. Note that this was without using *Race*, *Gender* or *Education Level*, which were used to manually spot check the 42 1:1 matches. After the generalization the number of 1:1 matches is 4 and the number of 1:2 matches is 5. Thus, the number of people for whom an adversary has a 50% chance of guessing or inferring a salary has fallen from 162 to 9 with the generalization to 10 zip code regions. The k-values suggest that the generalized Levels table protects privacy over the majority of the members of the table from a join attack leveraging location. The joined table is much larger, the sizes of the resulting equivalence classes are much larger and the subset who are highly exposed is much smaller. The privacy frameworks to analyze this distribution outside of strict k-anonymity is beyond the scope of this paper. In that framework k = 1 in both tables, but clearly the attack is less effective with the generalization in place for almost all of the members of the Levels database. With k = 50 the generalization produced 22, 1:1 and 32, 1:2 matches. These results are summarized below in Figure 8.

Figure 8: Results of Mondrian

Mondrian Generalization	1:1 Matches	1:2 Matches
None	42	120
k=50	22	32
k=200	4	5

An investigation of the generalization from the standpoint of geography suggests that not much utility was sacrificed for the increased privacy, though this is clearly a subjective evaluation. The colored regions in Figure 6 show the results of the zip code generalization for k = 200. The colored regions suggest that locality of information is still useful. Salary data can be compared in several regions of the San Francisco Bay Area, Central California, the Los Angeles area, San Diego and more remote locations in California. The regions are not contiguous because many of the zip codes did not appear in either the Zoom data or the Levels database.

4.5 Limitations

The results of this study must be interpreted within several important limitations. First, the re-identification was probabilistic in the sense that true identifiers such as a social security number or known address were not present. Without a unique identifier all the individuals on LinkedIn or Zoom the matches were an educated inference on the data subjects, not a truly definitive identification. Namsor used to classify gender and race is a classifer model, producing results with only a degree of probability. Second, the LinkedIn attack was restricted to only those in who had a LinkedIn profile, as opposed to the a more broad Internet presence. Third, LinkedIn does not have quasi-identifiers "race" and "gender". As such, Namsor had to be used to infer race and gender attributes for each possible LinkedIn profile matched to a Levels row. This isThis library has a free tier limit for its service. The limits to this free tier library prevent broader exploration and attacks. Fourth, the Namesor library only contains the races: Asian, White, Black and Hispanic. Some Levels records included Native Americans and due to the Namsor library's exclusion of Native American as a race, potentially Native American LinkedIn profiles could not be joined to the Levels database. Fifth, this study was limited to data in California.

5 Conclusions

The results of our re-identification study over the Levels database utilizing two attack mechanisms (scraping and joining) demonstrated that uploading compensation information along with even a few quasi-identifiers can expose a Levels users to re-identification and thus exposure of their salary and broader compensation information. Location information, even broad location information, can potentially expose salary information. While some small equivalence class participants, such as high earning minority females, may seem particularly vulnerable to re-identification, this study could not prove that to be true. Part of the reason for this appears to be that California (the scope of this study) and many such individuals tended to work for large companies. Also the limitations of the Namsor classifying library free tier prevented a more board approach based on gender and race.

Through a proof of concept use of the Mondrian algorithm, this study suggested that generalization of location information to a regional level produced protection for Levels users. Feeding a desired k of 200 to the Mondrian algorithm dramatically reduced the number of 1:1 matches in the Zoom table join attack. This generalization also preserved the utility of the use case, namely allowing Levels user is to obtain compensation comparisons for a company and job title within a geographic region in which costs of living are most likely fairly uniform.

Time, resources, and knowledge constraints limited the scope of this study. Future promising directions for further study could include training home grown classifiers for race and gender classification of names in order to ensure that the classifier has enough data points for all races including Native Americans, scraping across the web from a search engine such as Google instead of limiting scope of the search for auxiliary information to LinkedIn's search engine, exploring additional protection mechanisms derived from Principal Component Analysis (PCA) and K-means classifiers the Levels database.

6 Member Contributions

Angela Gao	Alice Hua	Lesley Matheson	Dhyani Parekh
Proposal Project Strategy Attacks Results Interpretation Report	Proposal Project Strategy Attacks Results Interpretation Report	Proposal Project Strategy Attack Protection Results Interpretation Report	Proposal Project Strategy Attack Results Verification Results Interpretation Report

Works Cited

[1] https://time.com/5353848/salary-pay-transparency-work

[2] Sweeney, L.. "k-Anonymity: A Model for Protecting Privacy." *Int. J. Uncertain. Fuzziness Knowl. Based Syst.* 10 (2002): 557-570.

[3] Henriksen-Bulmer, J., and Sheridan Jeary, S., "Re-identification Attacks—A Systematic Literature Review", International Journal of Information Management,Volume 36, Issue 6, Part B,2016, Pages 1184-1192..

[4] Yves-Alexandre de Montjoye, Radaelli, L, Vivek, Kumar Singh, and Pentland, A., "Unique in the shopping mall: On the the Identifiability of credit card meta data", Science 347 (6221):536-539.

[5] Jaro, M., Probabilistic linkage of large public health datafiles. Statistics in Medicine, 14(5-7) 491-498, 1995.

[6] Ferguson J, Hannigan A, Stack A. A new computationally efficient algorithm for record linkage with field dependency and missing data imputation. Int J Med Inform. 2018 Jan;109:70-75. doi: 10.1016/j.ijmedinf.2017.10.021. Epub 2017 Nov 6. PMID: 29195708.