Predictive Privacy:
Modeling Privacy Harms

Sarah Radway

Contents

Chapter 1

Introduction

The right to privacy is fundamental to a democratic society. It is the stronghold that maintains our freedom of speech, allowing us to express personal and political opinions without fear of judgement or retribution. It is the right to privacy that allows us to withhold our medical conditions from the public eye, to publish articles critical of the state, to conceal our private relationships and business.

As technology advances and becomes more accessible to the average person, privacy has become the unwilling victim of computational advancement. Thanks to Moore's Law, data processing capabilities once considered unthinkable have now become accessible to the average consumer. In the context of privacy violation, this has a significant impact on one's ability to deanonymize: data can be processed at an increasingly rapid pace, allowing databases to be matched, or social graphs to be aligned, at large scale.

This phenomenon becomes more detrimental to privacy when coupled with advancements in machine learning techniques and theory. Fundamentally, machine learning is used to draw conclusions from or about a dataset—this includes datasets of personal information. Machine learning is capable of comprehending the relationships between variables far better than a human; it gives us the ability to make inferences about sensitive characteristics, even from seemingly insensitive ones present in a dataset.

These advancements in database matching and machine learning simply mandate us to reevaluate our level of protection against privacy threats. Privacy was once able to be assumed out of magnitude: applying the motivated intruder test suggests an anonymized dataset's safety ought to be evaluated through its capability against a reasonably proficient adversary, not an individual with specialized resources or knowledge [26]. In the recent past, an average person without specialized equipment or knowledge would not have been able to process large scale data in the manner necessary to employ database matching or

machine learning mechanisms effectively. This is no longer a safe assumption. As we must change our approach to defense when an enemy grows stronger, so too we must transform the defense of our privacy. A new understanding of dataset evaluation has proven necessary: an understanding that accounts for the ability to process both internal and external data with the new tools available, to create data from what is not there, and to connect data that, to a human, seems unrelated.

When private data is released inappropriately, it can cause substantial harm. There are serious financial, social, or legal consequences for individuals who have data about them released to an adversary. Consider an individual who has financial data leaked, leaving them vulnerable to fraud. Or, consider the leakage of more personal information, such as an individual's sexuality, gender identity, or other personal information–in some countries, leakage of this can be a matter of life or death. It is crucial to assess the harm posed by a dataset before release, as this data has a very real impact on human lives.

Thus, this work will examine how we can account for harm arising from innovation in machine learning and database matching, and will propose a prefatory method for the quantification of harm. The work's primary contribution is to propose a method of modeling the harm of a dataset, analyzing the risk of deanonymization given these new factors. We carry out two experiments in order to demonstrate the harm accompanying modern dataset applications—the first exploring machine learning's applications to WhatsApp social network analysis, the second exploring the potential for large-scale database matching between University directories and TinderU profiles. We apply our model to these examples, to demonstrate its the efficacy in a real world setting. We show that this model can be applied as a general framework, to guide both legal regulation of data release and implementation of current anonymization methods.

Chapter 2

Background

In order to quantify threats to user privacy, it is crucial to understand current methods of anonymizing and de-anonymizing datasets, as well as the related legal frameworks that govern these practices. This section presents this necessary background for our exploration.

2.1 Privacy & Policy

While the United States does not have an all-encompassing data privacy doctrine, such as Europe's GDPR, there are several industry-specific, state-specific, and age-specific acts that govern American data practices. This legislation, originating from state and federal governing bodies, intends to protect user's Personally Identifiable Information (PII) from violation.

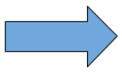
Personally Identifiable Information (PII) is a subset of personal information that can be connected to an individual's identity. In their *Requirements for Personal Information Protection*, the FTC cites examples such as "name, postal address, phone number, e-mail address, social security number and driver's license number" [23]; however, the scope of PII is extensive and challenging to define. As the U.S. General Services Association describes, PII "requires a case-by-case assessment of the specific risk that an individual can be identified using information that is linked or linkable to said individual"[6].

The greatest challenge involving PII is quantifying the impact of external information (e.g. a separate available dataset), which can transform non-PII into PII. In Figure 2.1, we observe two separate datasets. The first dataset we consider to be publicly available. This dataset contains information which could have been scraped from product reviews: names, approximate ages, and purchase IDs. The second is a dataset that we are considering releasing to an advertiser, containing non-PII data (device IDs, recent purchase IDs, and purchaser incomes). Say we consider income or device IDs to be releasable alone,

but not when associated with the name of an individual. Using purchase IDs as a key, we can combine the two databases and deduce individuals' income levels. In this way, we can see a basic example of how identifiable information can be derived through the combination of available, but not directly included, information. This external information muddies the evaluation of data release safety; one must look beyond the dataset itself to assess its dangers.

NAME	PURCHASE ID	APPROX AGE
Jane Doe	01ABCDEF	18-35
Jake Doe	12ABCDEF	65+
John Doe	23ABCDEF	18-35

DEVICE ID	INCOME	PURCHASE ID
01ABCDEF-01234567-89ABCDEF	$35k	01ABCDEF
12ABCDEF-1234567-890ABCDEF	$64k	12ABCDEF
23ABCDEF-234567-8901ABCDEF	$72k	23ABCDEF



NAME	INCOME
Jane Doe	$35k
Jake Doe	$64k
John Doe	$72k

Figure 2.1: Demonstration of Derivation of PII

Given the absence of broad federal legislation, United States privacy regulation takes place on a smaller scale, mandating data protection in specific circumstances. These circumstances could include factors such as the user's age, the user's location, or the data's relationship to the federal government.

Notably, the FTC's Children's Online Privacy Protection Rule (COPPA) provides special protection for the data of children under 13 [4], requiring parental consent prior to the collection of children's personal information, with relatively strict standards for verification [14]. Additionally, it requires that websites serving children under 13 must outline what personal information is collected, how it is used, and disclose all third parties that are collecting personal information. There are strict specifications about the circumstances in which a child's contact information may be stored, how their data may be used, and under what conditions they may be contacted [14].

Additionally, a user's residence location can affect their data protection. A number of states have introduced their own data privacy legislation, to account for the lack of federal protection. Different states' rules range in strength: most states simply require that businesses maintain "reasonable security procedures and practices...to protect from unauthorized access, destruction, use, modification, or disclosure" [40]. However, there are states with stronger regulations: California's CCPA stands out as an example of strong privacy legislation in the United States. The CCPA provides consumers with control regarding what information is collected, stored, and shared about them. It further provides con-

sumers with a private right of action in the event of a data breach containing PII [1]. This represents a significant increase in the scope of privacy protection: putting power in the hands of the user.

However, outside of these highly specific situations, there is very limited national regulation regarding the treatment of consumer PII. Without regulation, there is minimal motivation to protect PII; there is especially minimal motivation to protect PII created from inference, where a user would need to combine personal information with public information. In Section 2.3, we will explore this capability to infer PII using database matching and/or machine learning, in manners currently permissible under federal legislation. We will further model the potential harm resulting from disclosure of PII using these techniques in 3.3.

2.2 Anonymization Techniques

Researchers explore dataset modification and presentation in order to ensure anonymity. We will explore two areas of anonymization research:

1. Methods to evaluate the privacy protection a dataset provides
2. Methods to modify the dataset in a manner that increases its anonymity

2.2.1 *Evaluation*

K-anonymity is a property of a dataset that represents the strength of its protection from deanonymization, based upon the extent to which each item is indistinguishable from the whole [48]. As defined in Zhou et al., "a dataset is said to be k-anonymous ($k \geq 1$) if, on the quasi-identifier attributes (that is, the maximal set of join attributes to reidentify individual records), each record is indistinguishable from at least ($k - 1$) other records" [57]. For this reason, k-anonymity was introduced as a measurement of the extent to which a dataset allows users to remain anonymous. Much other work has built upon the initial efforts of k-anonymity; for example, the concepts of alpha-anonymity, l-diversity, and t-closeness further explore improvement in the realm of quantification of a dataset's privacy.

Alpha-anonymity, a term coined by Wong et al., attempts to protect not only the identity of a user's sensitive information, but its relationships with other data as well. The alpha value is meant to account for the, "confidence of implications from values in the quasi-identifier to the sensitive value" [53].

L-diversity, alternatively, attempts to account for an adversary's background knowledge [36][54]. L-diversity seeks to prevent inference in disclosure, by quantifying the amount of background knowledge needed to eliminate possible values, and thus disclose information. This amount is represented by the parameter ℓ.

A third alteration of k-anonymity, ***t-closeness***, examines limitations of l-diversity, separating the adversary's "information gain into two parts: that about the whole population in the released data and that about specific individuals" [32].

2.2.2 *Modification*

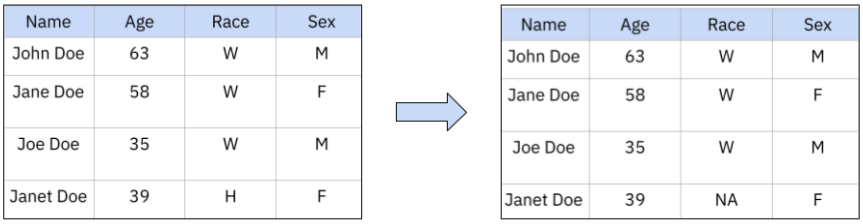
While it is important to understand measures of dataset anonymity, it is equally important to examine the methods of achieving these designations of k-anonymous, l-diverse, etc., and to examine how datasets can be modified to preserve privacy. These privacy preservation methods mainly fall into the two categories of data confidentiality and data perturbation.

Data Confidentiality

Data confidentiality (in the context of disclosure avoidance) represents the limiting or censoring of data prior to release. Notable examples include the protection of sensitive cells via methods such as cell suppression [45], recoding [7], and the use of rules to redefine sensitive cells [**?**].

Cell suppression and other similar methods of protecting sensitive cells, involve a process where "values of a variable are replaced by a missing value" [45]. In other words, if a given value doesn't fit a set of constraints, it can be suppressed, and replaced with a value such as 'NA'. Take the dataset in Figure 2.2: say we set a constraint that there may be no unique values for race. If we were looking at the race cells in the table in Figure 1, we would need to suppress Janet Doe's race, so as to not violate our constraint. These rules used to define sensitive cells can range from those as basic as in our example to those involving thresholds for k-anonymity. While effective, this method of disclosure avoidance is limiting for statisticians: missing values lead to misleading data.

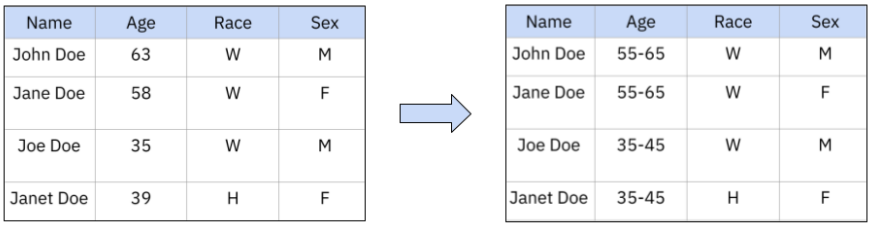
Recoding is another method of data suppression, in which the range of classifications is decreased, "by combining or grouping categories for categorical variables or constructing intervals for continuous variables" [7]. For example, in Figure 2.3, one may wish to suppress information regarding age of the individuals in the dataset. Through displaying an age range rather than an exact value, we recode, and thus generalize, the data to a range of values, introducing ambiguity. Top coding and bottom coding are very similar to recoding, except only the top or bottom extrema of the distribution are recoded [7].



Name	Age	Race	Sex
John Doe	63	W	M
Jane Doe	58	W	F
Joe Doe	35	W	M
Janet Doe	39	H	F

Name	Age	Race	Sex
John Doe	63	W	M
Jane Doe	58	W	F
Joe Doe	35	W	M
Janet Doe	39	NA	F

Figure 2.2: Demonstration of cell suppression on Janet Doe's race



Name	Age	Race	Sex
John Doe	63	W	M
Jane Doe	58	W	F
Joe Doe	35	W	M
Janet Doe	39	H	F

Name	Age	Race	Sex
John Doe	55-65	W	M
Jane Doe	55-65	W	F
Joe Doe	35-45	W	M
Janet Doe	35-45	H	F

Figure 2.3: Demonstration of recoding of age cells

Data Perturbation

Data perturbation, rather than masking the true data, involves the manipulation of the true data through methods including rounding, the addition of noise, or interchanging existing data. ***Rounding*** involves the manipulation of values to adjacent values before publication. Rounding can be random or controlled; random rounding represents the arbitrary rounding down or up at a set probability, by a randomly selected amount from a set range [45]. Alternatively, controlled rounding does not have this random aspect in adjacent value selection [45].

Data swapping involves the pairing "of records with similar attributes and then interchanging identifying or sensitive data values among the pairs." [17, 29]. For example, in Figure 2.4, we have selected Joe Doe for swapping on the variable of race. Thus, we would swap his data with an individual from another dataset, Jack Doe, who is the same age and sex as Joe, but potentially a different race. There are many variations of swapping: we may change the number of rows intended to be swapped (which is known as the swap rate), the variables being swapped, or the measure of similarity governing swap selection.

Name	Age	Race	Sex
John Doe	63	W	M
Jane Doe	58	W	F
Joe Doe	35	W	M
Janet Doe	39	H	F

Name	Age	Race	Sex
John Doe	63	W	M
Jane Doe	58	W	F
Jack Doe	35	B	M
Janet Doe	39	H	F

Figure 2.4: Demonstration of swapping between Joe and Jack Doe.

Graph modification is also a form of data perturbation, largely implemented in the context of social graphs. Zhou et al. set forth two main methods in their work of data perturbation through graph modification: clustering based methods and graph modification based methods [57].

Clustering entails grouping similar data together when sharing a dataset, clustering it into a "super vertex" [57]. There are various approaches to clustering vertices and edges into subgroups; in Zhou et al.'s overview, vertex clustering, edge clustering, vertex and edge clustering, and vertex-attribute mapping clustering are specifically mentioned [57]. Campan and Truta demonstrated how clustering leads to a need for greater seed data in order to successfully de-anonymize, thus validating its status as an anonymizing technique. However, while generalizing data through clustering does lend to a decreased ability to de-anonymize, they also demonstrated that there are quantifiable data losses in the process of clustering [12].

Alternatively, ***graph modification*** entails changing graphs on a local scale, so as to "preserve the scale and local structures" of the original graph [56]. There are several methods of graph modification, ranging from edge deletions and insertions, to randomly switching edges, to reweighing edge or node values [56].

Adding noise nodes to a dataset represents another modification method. Saptarshi and Tripathy use eigenvector centrality "to achieve k-anonymity & l-diversity by adding noise nodes in the raw data" [13]. ***Differential privacy***, as proposed in Dwork's work [21], is another semantic approach to dataset privacy–one of the best known methods involving the methodical addition of random noise. This approach "can achieve any desired level of privacy", while still "providing extremely accurate information about the database" [21]. In this way, modification of how data is presented can protect the privacy of those within a dataset.

2.3 Deanonymizing Techniques

Given the relatively weak legal framework preventing deanonymization, it is natural to examine current deanonymizing techniques that can be implemented against consumers, in order to obtain their PII. These are the techniques that require or may benefit from database matching or the use of machine learning. We will categorize de-anonymizing techniques into those involving 1. structural re-identification and those involving 2. adversary attribute knowledge.

2.3.1 Structural Re-Identification

Structural re-identification represents any attack technique where an adversary possesses information about the structure of a graph. Fu et al. set forth a framework with two main methods of social graph deanonymization: seed matching and node signatures [24]. The first method, ***seed matching***, involves taking advantage of 'seeds': nodes that are present in different graphs, but are known to represent the same individual. Through using this information, seed matching techniques are capable of anonymizing individuals at a scope far beyond the initial seed. Notably, papers including Yartseva and Grossglauer, as well as Chiasserini et al. explore the lower limit of seed amount required initially to successfully de-anonymize a network [55][22], and Naranayan explores a deanonymization algorithm based in seed identification and propagation [39]. Expanding upon this idea of seed matching, Backstrom et al. makes their own 'seed' in the network, through establishing their own nodes, and figuring out their orientation, thus gaining the ability to comprehend the structure of the network in its entirety [10]. This seed data is also extremely valuable as training data, as will be discussed further in Section 2.2.3.

The second method, ***node signatures***, involves the analysis of node features, such as its degree or sub-graphs, in order to de-anonymize. Through the application of node signatures, via the use of multi-hop neighboring nodes' information, previous works have had success in de-anonymization without needing seeds [43][30]. Others have directly observed and labeled features, such as in Ullah et al., when the authors were able to de-anonymize programmers using code features [50].

2.3.2 *Attribute Knowledge*

The second overarching deanonymization technique involves taking advantage of external attribute information. Through the possession of knowledge outside the data itself, an adversary can identify individuals within a dataset. This outside knowledge could consist of a separate dataset containing information about specific users in the original dataset, targeted knowledge about individuals present in the original dataset, or it could simply constitute outside, factual, un-targeted knowledge about information present in the original dataset.

Regarding un-targeted information about the original data, Li and Li explored mining knowledge from the dataset as a means of representing an adversary's background information based in "absolute facts" and "partial knowledge of demographic information" [33]. They provide the example of a male, who can be inferred to not have ovarian cancer, and of a young woman, who we can infer has a low risk of heart disease, based upon publicly available medical condition risk data. This type of adversary knowledge is based on reasonable assumptions derived from knowledge not specific to individuals in the database. Other works have explored disclosure associated with an attacker possessing background information that is targeted, and difficult to predict, and have modelled the harm associated under worst-case circumstances [37][34]. Targeted information is specific information possessed by the adversary about an individual or individuals in the dataset. This information could have a variety of origins, ranging from personal relationships to available data sources. The use of this outside data can be used in a manner similar to seed nodes, to identify individuals in a dataset using a process of elimination, aligning known and reported features to the individuals in the data.

2.3.3 *Role of Machine Learning*

The use of **machine learning** in deanonymization has become widespread, within both of the formerly mentioned seed mapping and attribute knowledge-based approaches. Machine learning entails the creation of computational models, through analyzing large amounts of 'training' data. The ultimate goal of machine learning is to give a correct prediction of the output class of an unseen

example of the same input type as the given training data. There are two main types of machine learning: supervised and unsupervised. ***Supervised learning*** entails creating a model when the initial training data is labeled with an associated output class. Alternatively, ***unsupervised learning*** entails creating a model of an underlying structure, when the training data provided is not labeled, with the ultimate goal of providing a summary of the data.

As outlined in Sharad and Danezis, there is clear potential for the use of "deanonymization as [a] learning task" [43]. Their study explored the use of a set of 'seeds' as training data for supervised learning applications: specfically, Sharad and Danezis use seed data as training data with random decision forests, in order to understand an anonymization algorithm [43]. Others have applied supervised learning techniques to optimize deanonymization tasks, such as Lee et al.'s use of a pseudo relevance feedback support vector machine (PRF-SVM) to optimize graph matching [30], or Li et al.'s use of kernel estimation techniques to model adversary background knowledge [34].

Alternatively, unsupervised learning techniques may also be applied to datasets for the purpose of de-anonymization. Gaihre et al., Pham and Lee, and others have employed unsupervised learning on cryptocurrency transaction records, using graph learning to de-anonymize or draw conclusions from clustering users' transaction data [25] [41].

It is clear that the potential uses of machine learning techniques in the de-anonymization process have increased, and will likely continue to do so. Aside from pure de-anonymization, however, machine learning has also allowed for the deduction of PII from personal information. For example, ethnicity is a data type that can be considered PII [15], and thus is intended to be allotted extra protections. However, as Wong et al. demonstrate, PII can be inferred from information that may not necessarily be considered PII: in their study, they were able to classify individuals' ethnicity with 91% accuracy, using only their name and province [52]. Further, sexual identity was able to be predicted using machine learning techniques such as a network classifier [28]. In this way, the power of machine learning to de-anonymize becomes evident: PII stands to be created from purely personal information.

Supervised and unsupervised learning allow researchers to optimize the process of pattern detection, to identify patterns not obvious or logical to humans, and to draw conclusions about initially innocuous data. It is thus integral to consider its potential applications when discussing de-anonymization.

2.4 Modeling Anonymity

Our analysis is motivated by former works that have modeled the ability to anonymize and de-anonymize, including Ding et al.'s model of de-anonymizing in social networks [18], Narayanan and Shmatikov's framework for analyzing privacy and anonymity in social networks [39], and Hay et al.'s presentation of models for adversary knowledge, disclosure, and social network anonymity [27].

Other works have further explored modeling the concept of adversary knowledge, notably Li et al. [31] and Li et al. [34]. This adversarial knowledge represents the ability of an attacker to use an outside database for the purpose of deanonymization.

Lastly, individuals have modeled the performance of various anonymizing and de-anonymizing techniques: papers such as Bayardo and Agarwal [11] or Mauger et al. [38] have explored the optimization of anonymization techniques, namely k-anonymity. However, to the best of our knowledge, no former model has focused on the harm resulting from de-anonymization. In this paper, we explore the nature of the relationship between data content and the capability and nature of deanonymization.

Chapter 3

Harm

Now that we have thoroughly established the necessary background, we may turn to our analysis of harm through the deanonymization of personal data. It is challenging to define harm; legal scholars and technologists alike agree it is "difficult to quantify and articulate" data's impact [19]. Largely, it is difficult to discuss harm without falling into the "creepy trap" [42]. This term, coined by Richards and Hartzog, serves to remind us that during explorations of this nature, we must take care to explore not what creeps us out, but rather, "what information we're concerned about, in what sense it is "ours," or why collecting and aggregating that information is wrong" [19].

We may feel uncomfortable knowing that a company is aware of intimate details of our lives. Take for example, the well-known case that swept through national news outlets, detailing how Target was carrying out extensive purchase analysis to determine the likelihood that a woman was pregnant [20]. Target would then use this information to provide targeted advertising to pregnant women, to encourage them to make purchases at their stores. The news story told of a high school girl receiving these ads for maternity items, prior to her parents knowing she was pregnant.

The daughter in this news story would likely feel 'creeped out' that Target realized she was pregnant before her own parents. Yet, we must differentiate between personal and "the illusion of personal" in the realm of advertising–we must examine whether data truly "contain identifiable information about the user" [19].

As Richards and Hartzog explain, "data being processed for advertising isn't "yours", in that, "it doesn't identify you as the source" [19]. It is not necessarily your data that is valuable, but rather the "preferences, habits, and transactions of large numbers of users, which are consolidated, mined, and analyzed to find patterns and common behavior" [19]. Thus, it is hard to assign blanket definitions to what content is harmful and what content is not, and who ought

to be authorized for viewership and who ought not to be. As we will further explore, this concept is dramatically influenced by the data's subject, viewer, and content. A first example of harm is investigated in the work of Hartzog and Richards, which proposes harm through the chilling effect [19]. The chilling effect represents the deterrent of "free speech and association rights protected by the First Amendment as a result of government laws or actions that appear to target expression" [9]. This indirect suppression through intimidation is put forth as a form of harm that can result from deanonymization. In an oppressive regime one may feel intimidated to publicly speak in opposition of their government; likewise, individuals may not carry out trackable actions online when they know that those actions can be traced back to their true identity, out of fear of repercussion.

Similarly, Solove and Citron explore the concept of harm through anxiety and risk [46]. Deanonymization clearly increases risks for things like identity fraud, through the deanonymization and exposure of personally identifiable information. Solove and Citron provide justification for the fact that common law recognizes "increased risk of harm as an intangible injury worthy of redress" [46]. Further, the revealing of one's true information can be anxiety inducing, particularly when one is made aware of these accompanying risks. Solove and Citron therefore argue that we can think of these concepts of anxiety and risk as a source of harm, through the emotional burden of anxiety and risk.

This leads to our current dilemma: our inability to predict future harms. While common law suggests that the risk of harm is worthy of punitive measures, there is currently no method to quantify the inherent risk of the release of a dataset. That is what this work seeks to produce, a model of privacy harm to assist in risk analysis.

As we work towards establishing this model, this section will

1. Elaborate upon the methods of generating harm (using the techniques explored in background)
2. Explore the intricacies of defining harm
3. Propose our model of this harm

3.1 Generating Harm

We seek to quantify two modern methods of generating harm: generation through the use of database operations and through the use of machine learning techniques.

3.1.1 *Database Operations*

To consider database operations in our model, we must first define a relevant database. An aset is an attribute set, represented A; a database of personal information is a set of asets: various attributes form the columns of a database table, and each row of data about an individual is represented as an aset A.

Naturally, no real database has all columns for all people; even in principle, a real database will contain only a subset of columns. For a typical web advertiser the columns might represent attributes such as interests, web sites visited, IP geolocation, etc. Furthermore, the rows of a real database—-that is, the people the asets are about—-will not be all-encompassing. For example, Acxiom, one of the largest data brokers, has records on 700,000,000 people [44]; these records have an average of 1,500 attributes per person [49]. Because collections of asets are database tables, standard database operations such as union ($\cup$), join ($\bowtie$), and select (σ) may be performed on pairs of tables. The semantics, however, are slightly different.[1]

Join Operations ($\bowtie$)

The 'join' operation represents combining database rows into one, based upon a similarity in a database column. In our case, this means merging attributes associated with the same individual, into a single row For example, if a user had information present in two different databases, a join operation would entail combining the information from the rows of both of these databases into a new row. This can be seen in Figure 3.1, where two datasets contain information. There is a similarity present in the 'email' database column, that allows for a join operation, combining the information from both databases.

In principle, two authoritative databases should have identical values for any given attribute for a given person; in practice, however, there may be discrepancies. In such situations, the result is implementation-defined; values may be

[1] From a draft paper by Steven Bellovin

Figure 3.1: Demonstration of a database join operation.

flagged with a probability value indicating a lesser degree of certainty. Ordinary join operations work by matching records based on some attributes, however, some organizations do heuristic matches based on less-certain data, to handle situations such as different renderings of a name ("Steven Bellovin" versus "Steven M. Bellovin" versus "Steve Bellovin" versus mistakes like "Stephen Bellovin"), variant transliterations from other alphabets ("Muhammed" versus "Mohammed"), common names causing mismatches (not just "John Smith" but also "John Smith, Jr." and "John Smith, Sr."), data entry errors, and so on. Therefore, although ordinary join operations elide rows from one database that do not match a row in the other, it is often desirable to include such rows but to use the symbol '$\perp$' for missing values in this use case. A heuristic join is denoted $\dot{\bowtie}$[2].

Through the use of join operations, we can leverage data from multiple databases to create stronger user profiles and more significant privacy violations.

3.1.2 *Machine Learning Techniques*

In the context of database matching, it is valuable to conclude that an individual with a given email in one database is likely the same person as an individual in a different database with the same email. While it can prove difficult to confirm the accuracy of these connections with certainty, the augmentation of a database to include a greater scope of data is very powerful for deanonymization, and thus justifies the inclusion of inferred or uncertain values. As explored in Section 2.3.3, there are numerous machine learning techniques that can be used

[2]From a draft paper by Steven Bellovin

to produce harm, especially when it comes to augmenting inference capability. Through machine learning's ability to learn complex patterns, we can connect social graphs, find potential join database operations, and bring new meaning to data within the database, extrapolating based upon information that is present.

By using supervised and unsupervised learning techniques with external and internal information from a dataset, we can predict the identity of individuals in a dataset. Similar to the aforementioned join database operation, we can train neural networks to recognize rows of data about the same individual in different datasets. However, these networks may identify heuristics that might not be obvious to an individual carrying out database matching, potentially improving performance.

More generally, machine learning augments deanonymization by inference, giving the ability to expand a database beyond its original scope. Using machine learning, we can predict an individual's interests based upon information listed about them; for example, in Wong et al.'s work, machine learning gives us the ability to infer and individual's ethnicity by their name and province, with an accuracy rate of 91% [52]. We could thus add an additional column to a given table of personal information for ethnicity, with a value of the ethnicity produced by Wong et al.'s model, and a heuristic of 91%. In this way, we can use machine learning to add or fill in additional rows, based upon the produced inference values and heuristics. Similar to the heuristic join of database operations, the likelihood of an inference in machine learning can be expressed through the probability associated with a value label. The methods for returning this label probability will vary by learning technique: for example, this likelihood value could be represented by the maximum likelihood in a likelihood function, or by a closeness measure in the context of a clustering mechanism. All of the methods outlined in Section 2.3.3, from optimization matching methods to unsupervised graph learning [30][25], expand a dataset to give it greater deanonymizing capability.

3.2 Exploring Harm

Now that we have examined new methods of generating harm, we may explore the integral qualities of harm that will be necessary to portray in any harm model; namely, that harm is user-dependent, viewer-dependent, interdependent, and cumulative.

3.2.1 *Harm is User-Dependent*

Harm is, by nature, user-dependent, as different information is private to different users. For example, consider information regarding an individual's sexual and gender identity. All users of Tinder provide this information to the app,

for the purpose of finding appropriate partners. However, the desired privacy surrounding this data varies greatly from individual to individual. A straight male living in the United States will likely have little problem with this data being released; while, a gay man in Egypt would potentially face discrimination, and even jail time—Egyptian men were jailed for their sexual identity under"indecency" laws, based upon social media and dating app presence, according to Human Rights Watch [51][16].

Similarly, the ability to indirectly disclose information (including sexuality) from one's social graph has been proven effective[28]—no longer is direct leakage of this data in its pure form the only danger. Through the use of machine learning, one can infer information about you from the information of those you interact with. This is dangerous considering our presentation of harm as user-dependent: one is not in control of the flow of information about them, even though this information is perceived as more or less private on an individual basis. Perceived harm may vary greatly, due to its status as user-dependent.

3.2.2 *Harm is Viewer-Dependent*

In the same manner, harm is viewer-dependent, to account for varying intentions and capabilities regarding data use. Take medical data as an example: while an individual may be comfortable with their medical disorder being shared with their care professional or a company that can help them to recommend appropriate treatment, there would be significantly more perceived harm if the user's medical information were revealed to, for instance, an insurance company. The content of the data is not inherently dangerous in the hands of a doctor, but is definitively so in the hands of an insurance agent, who could potentially raise their premium. In this way, the harm the data poses is viewer-dependent.

As in the aforementioned example, harm could be viewer-dependent due to the position or career of the viewer. Additionally, this dependency could relate to the background knowledge possessed by the viewer, or the relationship between the viewer and the individual. Background knowledge comes in many forms, and can be valuable in de-anonymizing through providing enough information to identify a user in a dataset. As discussed in Section 2.3.2, de-anonymized individuals can serve as seed data, allowing for larger scale deanonymization. For a simple example, if you knew that Bob's birthday was in May 1975, and only one individual in the dataset had a birthday in May 1975, you would be able to de-anonymize Bob. Knowing who Bob was in the dataset would potentially allow you to identify others in the dataset by elimination. Harm can be significantly increased when a viewer possesses this background knowledge. In this way, background knowledge is a factor that contributes to harm being viewer-dependent.

3.2.3 *Harm is Inter-Dependent*

Harm is also inter-dependent: individual data points and datasets can impact our ability to anonymize or deanonymize other data. In anonymizing, as Zhou et al. explain, "changing labels of vertices and edges may affect the neighborhoods of other vertices, and removing or adding vertices and edges may affect other vertices and edges as well as the properties of the network" [57]. Likewise, as discussed in Section 2.3.1, one individual's identity can be used to identify others in the dataset, through process of elimination. Data cannot be considered independent: it must be considered in the scope of the dataset as a whole, keeping in mind principles such as k-anonymity, and considering information contained in other public datasets.

3.2.4 *Harm is Cumulative*

Lastly, harm is cumulative. As previously discussed in the context of database matching, combining various data sources increases the risk of deanonymization, and allows for greater information to be revealed about an individual in the process. Obviously, the more information that is revealed, the more information an adversary posses, and can use to identify and harm an individual.

3.3 Modeling Harm

In this work, we seek to create a model quantifying an architecture's privacy impact. We incorporate our new understanding of harm resulting from the use of adversarial learning and database operations, and express the principles of harm in a formal manner. Through the application of this model, we hope to provide guidance in making decisions regarding the release and regulation of datasets containing information about individuals. Our hope is that this model will serve as a general framework for approaching the question of data release safety.

In our model, we consider an attribute set (aset) A, that is made up of name/value/probability triples $a_i =< n_i, v_i, p_i >$ such as <phone number, 202-555-1212, 1> or <zipcode, 10027, .9>. The type of the value is attribute-dependent; it may be a Boolean (<retired, True>), a set (<hobbies, woodworking,cooking, bicycling>), etc. This aset represents the contents of a database. From this aset A, machine learning and database operations provide the ability to create A'. The extra items A' - A are putative facts produced not by direct observation but by calculation; as such, they generally have probability values less than 1. This aspect, that we can now reason about predicted attributes rather than just observed ones, is the defining cause of our need to reconsider our current approach to harm. We can represent the formation of this aset A' by function M : A $\rightarrow$ A' such that A $\subseteq$ A'.

Through this function, we have expanded aset A', containing the contents of A, along with additional, uncertain but probable information included about individuals in the dataset. Additionally, we define a harm set, H, consisting of functions, h_j. Each of these h_j's represent a function over a subset of an aset. We define the specific attributes from A' that serve as parameters to each function h_j as σ.

Using this information, we formulate our model, representing an architecture's total potential harm to privacy:

$$\sum_{h_j \in H} h_j(\sigma_j A')$$

Alongside older methodologies surrounding data anonymity, such as privacy violations involving direct identifier association, this model reflects the harm that arises from new methods including machine learning or database matching. This model allows for the inclusion of decentralized and indirectly identifying data, which represents a majority of modern potential harm to privacy. Function M allows us to include data that was indirectly identifiable through means such as database matching or machine learning, via creating the expanded aset A' from the original aset A.

Our model reflects the user and viewer dependent nature of harm via the harm functions $h_j \in H$. A harm function could assign weights to values for a characteristic; assigning a higher weight for a values that cause more harm. For instance, in the previous example surrounding sexual identity, a higher weight could be assigned for the gay man in Egypt than to the straight man in the United States. Due to the open-ended nature of the harm function, the model is capable of accounting for the assignment of varying levels of harm based upon these more intricate and situation-dependent considerations. Additionally, the summation allows the model to reflect the cumulative nature of harm, as it creates a system where the total harm is a function of all smaller harms.

Lastly, this model is capable of describing the interdependent nature of harm, through the inclusion of σ_j as a potential input. σ_j represents not a piece of data, but rather a subset of data, that act as parameters for a given harm function h_j. This proves significant through representing the interaction of various data pieces; we are capable of assessing the dataset at varying scales–evaluating deanonymization capability based on attribute knowledge or seed data, both at a local level and a global level. σ_j allow us to identify others in the dataset, through examining the scope of the dataset as a whole. Additionally, the M function allows for the inter-dependency of the values in the dataset to be reflected in A'. For instance, two non-PII data values can be used to infer a PII data value using machine learning, as described earlier in the case of Wong et al.'s ability to infer ethnicity by name and province [52].

However, there is information regarding harm that is not reflected in our model. Namely, our model does not reflect the cost of harm: we do not represent how much work or money will be necessary to carry out the deanonymization. Just because deanonymization is possible does not mean it is likely or feasible, and our model does not account for this fact. Our model solely reflects the amount and impact of the potential privacy violations of a given architecture, not the difficulty in achieving those violations. Additionally, the model does not definitively answer whether the deanonymization or privacy violations of this type are inherently damaging in a legal sense: while we can quantify the amount of harm, we cannot say whether this harm should have legal standing, or establish a threshold for permissiblity. Rather, we propose that this equation can provide guiding principles to those making decisions related to personal data.

The status quo does not provide any regulatory framework, and largely does not consider the damaging affects of increased dataset augmentation capability. Our model seeks to introduce these considerations into the data release process, and to encourage evaluation to be formed on the basis of harm.

Chapter 4

Experiments

In order to examine potential applications of our model, we performed two experiments; the first attempting to identify machine learning usage on WhatsApp social graph metadata, and the second applying database matching between a database of Tinder user data and a University Directory. Through these experiments, we sought to identify the formerly discussed principles of harm. In Chapter 5, we will apply our model of harm to these experiments.

4.1 WhatsApp

Our first experiment (IRB-AAAT3566) sought to examine the role of inference in Facebook's creation of social graphs, using WhatsApp communications metadata. Specifically, we aimed to distinguish whether there was any identifiable connection between individuals' connections on WhatsApp and their Interest Profile on Facebook.

WhatsApp is an encrypted messaging service owned by Facebook; however, the metadata of these messages is not encrypted. Metadata is data about data: in the context of a WhatsApp message, the message text would be the data, while metadata would contain information such as who the message was sent to, what time it was sent, and where it was sent from. Although this information seems harmless, messaging metadata is vast and has significant potential for meaningful and identifying analysis. As Facebook specializes in advertising and the creation of thorough user profiles, it seemed plausible that they would be analyzing this communications metadata for advertising purposes. We therefore expected to identify WhatsApp using communications metadata and machine learning capabilities to determine connections, using structural re-identification-based methods elaborated upon in Section 2.3.1, and/or using machine learning based technology to understand what interaction qualities were valuable in quantifying connection. We set out to identify any signs that this metadata was being used by Facebook for advertising purposes.

4.1.1 Experimental Set Up

In order to show that communications were impacting advertising, we needed to create an artificial social network for exchanging communications. We therefore recruited 30 participants, from both Columbia and across the United States, to be a part of our artificial social network. Prior to the involvement of participants, we created a Facebook account for one of the researchers on a virtual machine, and establish an initial user profile through liking several Facebook pages from Alexa Top 500 Websites [3]. Additionally, we created a WhatsApp account with the same phone number to associate with the Facebook account. We allowed this account to remain unaffected by outside influence for one week. Throughout this time period, we collected the Facebook account's listed profile interests and suggested advertisements, so as to provide a 'control' user profile.

After this control period, we began meeting with participants. At the meeting, we asked each participant to provide us with their listed "User Ad Interests" on Facebook. To simplify this process we created a script that collected text from the user's screen once they had navigated to the appropriate page. We then either made them a part of the control group, or informed them they had been selected to exchanged messages on WhatsApp with the lab account. After this portion of the study, we waited 48 hours after a participant interaction, and then collected the study's Facebook account's listed profile interests. We ask participants to use the script to collect their listed profile interests, to track any potential changes in their interests as well. We additionally exchanged messages with participants selected for the 'WhatsApp' group, asking a series of questions about whether they thought their user profile representations were accurate. We continued this collection process for a period of 10 days, collecting data from each participant a total of four times. We thus sought to identify if there was a relationship between the user profile interests of the participants and the user profile interests of the controlled Facebook account. We expected that, if metadata was being used, the interests of participants would begin to appear in the interests of the study account, due to the establishment of a social connection between the two accounts.

4.1.2 Results

After ten days, the lab's account did not have any additions to the listed user profile interests, and the study was thus inconclusive. If this metadata was a factor in advertising formation, it was not significant enough of a factor to influence the study account's advertising. This does not necessarily mean that Facebook is not using WhatsApp communications metadata to target advertising. We hypothesize that the amount of communication was not significant enough to have impact on Facebook's functionality, or that the length of time in which the participants exchanged messages with the lab account was not long enough. It would be rewarding future work to repeat this experiment, with a longer time span and more variance and consistency in participant interactions.

4.2 Tinder & University Databases

Tinder is the most popular dating application in the United States, with almost eight million users in the US as of 2020 [47]. With a user base of this size and sensitive data involving gender and sexual identity, it is especially important to consider whether Tinder is doing everything they can to protect against modern threats, and further, whether they are using and managing their data responsibly.

Tinder is currently under formal EU investigation with regards to data privacy issues[35], and after investigating a CCPA release form, it is clear that Tinder is not treating users' data as they should: as personal, private information. Given recent events, including the current European Union GDPR investigation into Tinder, as well as Match Group's own admission of data breaches due to API vulnerabilities and lack of encryption, it's crucial to analyze Tinder's practices. This experiment (IRB-AAAT5771) sought to investigate the ability to perform large-scale deanonymization using information available on Tinder's API, using database matching with publicly available University directories. The availability to perform data matching of this type between databases that contain information including email, mailing and residential addresses, or place of employment, in relation to Tinder API data containing information regarding first name, sexual identity, age, and photos, is clearly dangerous in the hands of an adversary. We examine the potential danger for data misuse given Tinder's current API structure and the power of database matching.

4.2.1 Experimental Set Up

One of the study coordinators created a Tinder account, using their email, photo, identity, etc. to register. We set an age range for the account that seeks university-aged students, so as to target users with TinderU-the feature where students associate their profile with a university. We additionally set a target location: we selected the area surrounding a university town. We then collected information from the API responses about potential matches using Fiddler, a web debugger. This information consisted of the individual's first name, approximate age, bio, and frequently, university or place of employment. We identified the portion of the population that listed the specific university in their profile. Of the 1573 total profiles that we collected, 986 were eligible under these conditions.

We then used a script to search the university's database for profile matches. First, we searched this database using only first name, attempting to see how many perfect exact matches we could find, where only one person with that name existed, thus attempting to form connections between potential profiles on these different databases, using only the information provided on the Tinder API as a 'database key'. Second, we included secondary information available on the profile: sometimes, users had their major listed. Third, we sought to establish

exact name matches (not including close matches): while this introduces error, we wanted to explore potential performance.

4.2.2 Results

Our experiment puts forward a method of deanonymization with significant return at scale. For the first method, which queries the University database only based on first name, 92 out of the 980 profiles we examined were identifiable, giving us an identification rate of 10%. This means that when the first name was queried in the university database, for about 10% of the queries, only one individual with the given first name was returned.

Further, as can be seen in the first graph in Figure 4.1, for the whole of the dataset, not only were the number of exact matches high, but much of the data is largely clustered along the lower end of match numbers.

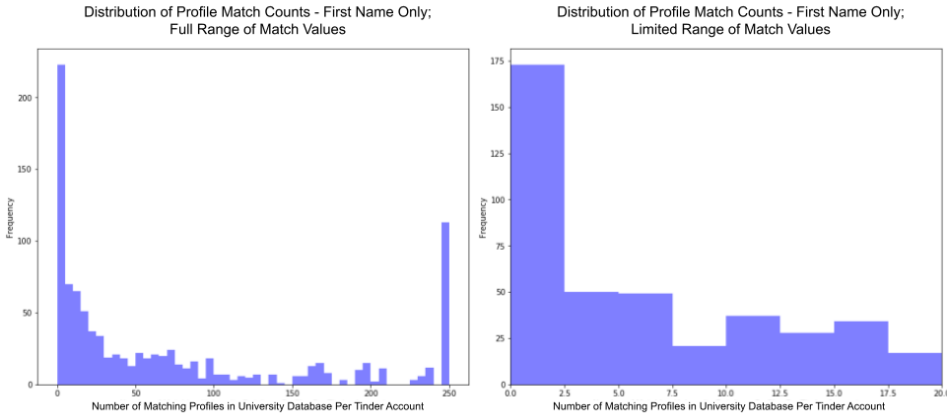


Figure 4.1: Exact Profile Matches By First Name Only

Taking a closer look at this cluster on the lower end, the second graph in Figure 4.1 displays the frequency of match numbers less than 20.

This proves to be particularly disturbing: with 471 profiles with 20 or less matches, and 284 profiles with 5 or less matches, these profiles are surely vulnerable as well through secondary methods. For example, to discern between the five potential profiles, the photos posted on the Tinder Profile could be matched with those returned by an image search engine. This is surely less straightforward than our current approach, but not out of the question, using an API such as DeepAI's Image Similarity API. It appears additional methods could supplement this method, and further increase the rate of identification.

In order to explore this, we did a minor addition to our current method, parsing the bios for listed majors to match with majors listed on the University directory and looking for exact name matches. With this new method, 117 out of 980 profiles were identifiable, increasing our capability to a 12% identification rate. The spread is displayed in Figure 4.2, showing a similar structure to Figures 4.1, for the original dataset.

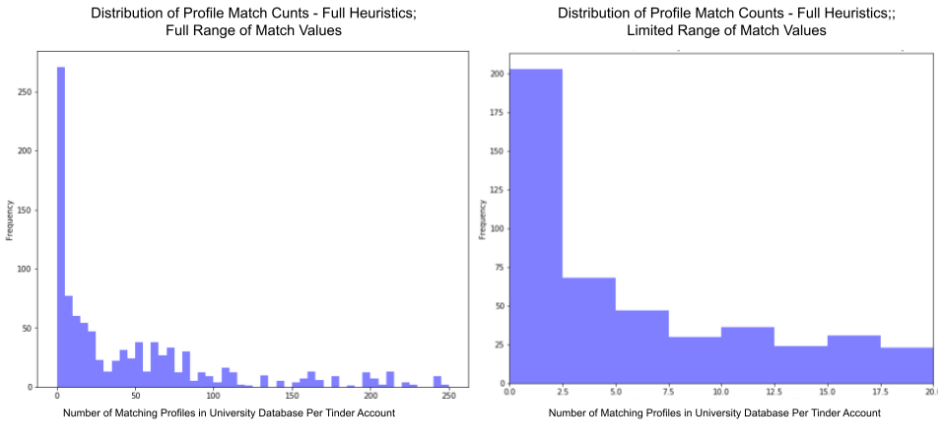


Figure 4.2: Exact Profile Matches With Increased Heuristics

It is important to think about this de-identification as more than quantities. We examine the content that is available to the public from our institution's (Columbia's) directory about a student. Published information including the student's home (dorm), mailing, and email address draws concern. When individuals sign up for Tinder accounts, they choose to publish their first name, their photos, where they go to school. However, it seems unlikely that Tinder users would willingly release information to potential romantic matches that included their home address. The alignment of personal information including sexual identity and preference with sensitive contact information is potentially dangerous in the world we live in; the ability to deanonymize this dataset is not a matter of improving functionality, but of ensuring the safety of users. When we look at other organizations that collect large scale data about topics including sexuality and race, we see significant attempts at maintaining privacy for the individuals the data is about. The U.S. Census, for example, implements in depth data perturbation techniques prior to data release; companies such as Apple investigate differentially private mechanisms in data collection and use [5] [8]. While the same scales and methods may be unreasonable, Tinder ought to be following suit in prioritizing consumer protection. At a minimum, Tinder has the ability to restructure their API in a manner that would cause this

method of large-scale database matching to be challenging, if not impossible. However, it is difficult to place the entirety of the blame upon Tinder. While Tinder's implementation allows for de-anonymization at a large scale, the publication of University databases would still allow for this process to be carried out manually, regardless of Tinder's design choices. If, for example, a stalker saw a Tinder profile associated with TinderU, they would be able to manually search a university directory with the profile information shown on their screen. In this way, the University Directory is truly the source of PII that raises concern: the posting of private information in the format of a student directory is protected under The Family Educational Rights and Privacy Act (FERPA), allowing for very little accountability in its dangerous potential uses. It is these types of considerations that our model hopes to bring to light; our framework will encourage Tinder to reconsider their API response format, and would likely discourage the publication of university directories. Through the implementation of these changes our model suggests, Tinder-using students' risk of identification or stalking is vastly decreased.

Chapter 5

Using Predictive Privacy Model

Our experiments allow us to contextualize our disclosure model; in order to comprehend how our model of dataset privacy can be used in real world decisions involving data release. As formerly mentioned, our model for predictive privacy harms leads us to examine the harm of a dataset individually and in combination with other related datasets. Therefore, as a case study, we may evaluate the datasets involved in our Tinder experiment, individually and in combination.

5.1 Contextualizing The Model

As a refresher, we model the privacy harm of a dataset with:

$$\sum_{h_j \in H} h_j(\sigma_j A')$$

where A' is the set containing the original set A and additional reasonably deduced information, h_js in harm set H are functions over subsets of A', and σ represents the specific subsets and parameters for an h_j. Given this model, we may examine how Tinder's data and the University's data would be represented by our model. We can begin with a dataset, containing information published via Tinder's API. In our model, set A represents the original dataset's contents. We can thus consider the set A to be the whole of the API data published: the first name, approximate birth year, bio, photos, school/workplace, etc. of all Tinder users. Example data is shown in Figure 5.1.

A' is a set expanded from A, representing inferrable information obtained by calculation, observation, or any method resulting in a match with an accuracy probability of less than 1. Therefore, A' would be made up of A, as well as all data that can be reasonably inferred from the original Tinder data. Because the scope of the Tinder data is so large, including personal information, this A' will

<first name, Joe, .99>, <age, 20, .98>, <university, University of CS, .99>
<first name, Jane, .99>, <age, 19, .96>, <university, University of CS, .99>
<first name, Jack, .99>, <age, 21, .98>, <university, University of CS, .99>

Figure 5.1: Example of aset data from Tinder

be quite substantial. For example, we could expand A to A' through the use of a public database, such as a database of Facebook profiles: through matching existing Facebook profiles, we could identify whether there is an existing profile that has the same first name, approximate age, and approximate location as the Tinder user. Depending on the nature of the match, we can infer that this profile is that of the individual represented in the database; therefore, with some probability that the match is correct, we can infer the additional information contained in the Facebook profile, such as the interests or relationship status listed in the profile, or the names of individuals' friends or family. Figure 5.2 demonstrates an example of this expanded aset.

<first name, Joe, .99>, <age, 20, .98>, <university, University of CS, .99>, <interests, (sports, fast food), .85>
<first name, Jane, .99>, <age, 19, .96>, <university, University of CS, .99>, <interests, (sports, smoothies), .85>
<first name, Jack, .99>, <age, 21, .98>, <university, University of CS, .99>, <interests, (theater, france), .85>

Figure 5.2: Example of expanded aset data from the original Tinder aset data using interest data

Other A' expansion methods could include the process we carried out in our experiment: instead of using Facebook, we instead used the University database to expand the asets. In this way, we can expand A' to include information such as a student's email, home, or mailing address. Figure 5.3 demonstrates an example of this expanded aset.

<first name, Joe, .99>, <age, 20, .98>, <university, University of CS, .99>, <interests, (sports, fast food), .85>, <email, joe@cs.edu, .80>, <home address, 1000 West Campus, .80>
<first name, Jane, .99>, <age, 19, .96>, <university, University of CS, .99>, <interests, (sports, smoothies), .85>, <email, jane@cs.edu, .80>, <home address, 1000 North Campus, .80>
<first name, Jack, .99>, <age, 21, .98>, <university, University of CS, .99>, <interests, (theater, france), .85>, <email, jack@cs.edu, .80>, <home address, 1000 South Campus, .80>

Figure 5.3: Example of expanded aset data from the original Tinder aset data using directory data

Additionally, as in Wong et al.'s work mentioned in 2.3.3 [52], we could use machine learning to deduce sensitive information, such as ethnicity, from names and approximate location (provided to the mile by Tinder). This would further increase the harm potential of a dataset, through the inclusion of greater PII associated with a user profile. Generally speaking, we must examine the potential of the deanonymization techniques discussed in Section 2.3, accounting for how structural re-identification and adversary attribute knowledge can be applied to a subset of A'. In this way, we can produce an extensive collection of asets originating from the original collection, A. This allows us to amass large amounts of data about Tinder users, albeit with reduced certainty.

Given this new collection, we may examine the harm functions in the context of our A'. As a reminder, a harm function, abbreviated h_j, is a function over a subset of A', that expresses specific aspects of harm a data poses. This may include the specific characteristics of harm that we discussed in Section 3, such as its user-dependent, viewer-dependent, inter-dependent, and cumulative nature. Consider harm functions we could establish for our expanded collection, A'. We may begin by considering the inter-dependent nature of harm. Specific information in collection A' becomes more revealing when used in the context of other information in A': for example, possessing only an email alone may allow a bad actor to send spam mail in a random manner. However, when one possesses then name, age, and names of family members, of the individual the data is about, a bad actor is able to effectively target a victim in manners such as phishing, impersonation attacks, etc. An email alone may not be powerful for phishing; but the more detailed information possessed about the recipient, the better.

In this way, this example also demonstrates the cumulative nature of harm, through showing the greater the amount of data, generally, the greater the ability to exploit. We could convey that this data is more valuable, say, for instance, in bulk, and create a harm function with a parameter (σ) representative of the size of the dataset.

Additionally, we may consider the viewer-dependent nature of harm arising from our expanded Tinder dataset. Tinder is a dating app; we must therefore consider that there is an increased risk for stalking-related incidents. This consideration leads us to associate increased harm with data that would be more valuable to an individual seeking to stalk: for example, we could weight information such as the home address or email to a greater degree, given the situationally more sensitive nature of this information.

Similarly, we could convey that the data is more valuable in the hands of a data broker than of a college student: while a college student may simply want to check that their Tinder date is who they say they are before agreeing to meet them in person, a data broker or stalker can do far more damage with knowledge of personal information. Likewise, some individuals on Tinder may value the privacy of their sexual orientation more than others–these harm functions allow us to represent this fact.

Using this model, we may quantify these harm functions and function parameters. For example, we could consider the initial harm function we discussed, involving the increased value with accumulated data. Given that we have an email, which we have assigned an arbitrary harm value parameter of .1, we may assign increasing harm parameters (σ) for the inclusion of other information in addition to this email: we can represent the independent nature of the data. While alone, knowledge of a user interest may not be valuable, when we have an email and learn user interests, we can represent this increased value by giving a higher harm value to the released interest data.

Similarly, we could convey that this variety of data is more valuable, for instance, in bulk, and create a harm function with a parameter (σ), that represents the size of the dataset, or a parameter that gives value to the viewer's qualities: for example, assigning a high harm value to a dataset that will be accessed by data brokers, and a lower harm value to a dataset that will be accessed by college students.

In this way, we can assign a value to the harm of a dataset. While currently the parameter values are up to the individual, interesting further work could explore more specific assignment of these values; for example, demonstrating that an email is ten times more valuable than a first name in a deanonymization task.

5.2 Applying The Model

5.2.1 Data Anonymization

We can apply our aggregate risk assessment, and its guiding principles, in design and release decisions. We can consider how the anonymization techniques discussed in Section 2.2 can be applied to our A, in order to narrow the scope of A' and the capabilities of h_js in set H. Beyond rudimentary techniques of removing identifiers such as names, the aforementioned methods intended to achieve k-anonymity or l-diversity act as methods of eliminating harm. By critically evaluating a dataset in the context of our model, we can make decisions regarding the extent of de-identification necessary, and the most effective methods for doing so. We can further determine the safety of a data release-through critically evaluating its performance.

For example, we may take our specific consideration of our dataset from Tinder. If we were Tinder, and planning to release this dataset for public accessibility, we could use the model in the aforementioned manner, producing a relative value for harm. As we have outlined above, this dataset's harm value would likely be unacceptable. This signifies to Tinder that action needs to be taken prior to release, whether that be reducing the amount or content of the information published, or using de-identifying techniques to reduce the risk of harmful data usage.

If the end use for the dataset was to understand the average age of a Tinder user, Tinder could simply practice cell suppression, removing every variable except for age, as shown in 5.4.

<age, 20>
<age, 19>
<age, 21>

Figure 5.4: Example of data for release via cell suppression

Alternatively, if the end goal for this dataset is aggregate statistical use, Tinder could implement a differentially private mechanism on the dataset. This allows for full data release, as differential privacy modifies the dataset through the addition of random noise, providing a privacy guarantee. Differential privacy would largely maintain aggregate statistical values for a dataset, as shown in 5.5, and would thus be an effective method for this use case.

When implemented correctly, differential privacy can allow for no harm to arise from release. In this way, our model helps us to consider when and how to

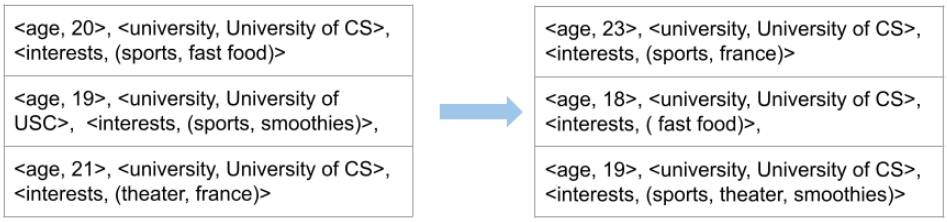


Figure 5.5: Example of data for release using differential privacy

use de-identifying mechanisms, and leads us to reflect upon whether data needs to be released in the first place.

5.2.2 Data Policy Regulation

Likewise, one may use our model to guide policy regarding harm. Today, existing data privacy legislation functions largely around disclosure. Under legislation such as the CCPA, parties must disclose what personal information is being collected. However, with the exception of special cases like health data, permissible content to be collected and released is largely unregulated. In the case of Tinder, when a user signs up, they click a button that signifies agreeing to a lengthy privacy policy. A part of this policy states:

> "We may use and share non-personal information (meaning information that, by itself, does not identify who you are such as device information, general demographics, general behavioral data, geolocation in de-identified form), as well as personal information in hashed, non-human readable form, under any of the above circumstances...We may combine this information with additional non-personal information or personal information in hashed, non-human readable form collected from other sources" [2].

In this way, under legislation such as the CCPA, users will have agreed to having their data stored, combined, and ultimately released in a non-readable form. Our model clearly shows the danger of these practices under current policy.

Rather than approaching privacy harm as a matter of user consent, our model urges us to evaluate the harm of data on a case by case basis. Consent is not enough to prevent privacy violations: it is an unfair and unrealistic standard to lay the burden of analysis on the consumer. A lay person will likely not have the time, motivation, or knowledge to thoroughly analyze every agreement they enter with a company. As we have explored, even to look at solely the original dataset itself is not enough to understand the inherent associated harm

with its release. The average consumer will likely not have a holistic view of the power of inference, and will not understand that even the information they are consenting to share is not limited to what the information they specifically provide.

Our model displays the complexity of decisions surrounding database release, and urges us to include its considerations into the data release workflow. In today's world, it should be mandated that a company consider the risks and potential harms resulting from the use of database operations or machine learning techniques on a dataset. Our model hopes to provide a framework for organizations to do so.

Chapter 6

Conclusion

Recent innovations in data analysis have made it challenging to evaluate the safety of dataset release—there are new considerations when deciding whether to release a dataset as is, to implement methods such as differential privacy, or to not collect or release data at all. While there is no panacea for privacy preservation, there exist a number of effective methods for combating adversarial attacks on datasets. From systems such as differential privacy, which provide privacy guarantees, to methods that simply suppress data cells, the anonymization techniques outlined in background together have tremendous capability to protect individuals and data from harm. The problem we currently face is how and when to implement these methods—our model hopes to contribute to solving this issue.

Without appropriately addressing the type and extent of harms a dataset poses, one cannot choose the appropriate method to protect against them, or be held liable for the harm caused if they are mishandled. Current understanding of a dataset's harm if no longer applicable. A new form of understanding of dataset harm is needed, one which considers the power of machine learning and database matching techniques to de-anonymize; these methods increase individuals' risk of emotional, financial, or even physical harm. Through understanding the harm a dataset poses, we may evaluate the need and appropriate methods for data anonymization: by understanding what we are defending and from whom, we may pick the best defense mechanism.

Thus, we sought to identify specific factors that increase risk of harm when present in a dataset.

The large-scale deanonmyization of two datasets via approximated join operations (Section 4.2) allowed us to examine the new source of harm posed by database matching. With a match rate of as high as 15% for sensitive information pertaining to gender identity, sexual identity, and residential addresses, the Tinder case study merits concern for individuals' safety and security. We demonstrated the significant potential for harm posed by database matching

with publicly available datasets, validating our need for a new framework for thinking about harm. We further investigated qualities that lead to an increased risk of harm, including the viewer-dependent, user-dependent, inter-dependent, and cumulative nature of harm. In this work, we formulated a model based upon these qualities, intended to evaluate a dataset's harm potential, accounting for these specific harm factors. Through the use of this model, we aim to provide guidance regarding a data architecture's privacy impact.

We hope our model will allow individuals to think about privacy impact quantifiably, thus giving rise to more formal best practices surrounding data de-identification, and laying the ground work for greater regulation regarding data safety. When harm from data release is thought about as more than a nebulous hypothetical, but rather as a tangible concept, reasonable guidelines can be put forth to protect consumers and citizens.

Bibliography

[1] *1.81.5. California Consumer Privacy Act of 2018 [1798.100 - 1798.199.100].*

[2] Tinder Privacy Policy. https://policies.tinder.com/privacy/intl/en.

[3] The top 500 sites on the web. https://www.alexa.com/topsites.

[4] *15 U.S. Code 6501-Children's Online Privacy Protection*, 1998. https://www.law.cornell.edu/uscode/text/15/6501.

[5] Statistical Quality Standard S1: Protecting Confidentiality, May 2015. https://www.census.gov/about/policies/quality/standards/standards1.html.

[6] *2180.2 CIO GSA Rules of Behavior for Handling Personally Identifiable Information (PII)*, 2019.

[7] Anonymization methods, 2019. https://sdcpractice.readthedocs.io/en/latest/anon_methods.html.

[8] Apple. *Differential Privacy Overview.* https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf.

[9] F. Askin. Chilling effect. *The First Amendment Encyclopedia*, 2009.

[10] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou r3579x? Anonymized social networks, hidden patterns, and structural steganography. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, page 181–190, New York, NY, USA, 2007. Association for Computing Machinery.

[11] R. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *Proceedings of the 21st International Conference on Data Engineering*, ICDE '05, page 217–228, USA, 2005. IEEE Computer Society.

[12] A. Campan and T. Truta. A clustering approach for data and structural anonymity in social networks. In *In Privacy, Security, and Trust in KDD Workshop (PinKDD*, 2008.

[13] S. Chakraborty and B. Tripathy. Privacy preserving anonymization of social networks using eigenvector centrality approach. *Intell. Data Anal.*, 20:543–560, 2016.

[14] Federal Trade Commission. Children's online privacy protection rule: A six-step compliance plan for your business.

[15] Centers For Disease Control and Prevention. What is personally identifiable information? https://www.cdc.gov/nchs/training/confidentiality/training/page581.html: :text=PII%20might%20consist%20of%20direct,unusual%20occupation%20and%20other%20details.

[16] N. Culzac. Egypt's police 'using social media and apps like Grindr to trap gay people', 2014. https://www.independent.co.uk/news/world/africa/egypt-s-police-using-social-media-and-apps-grindr-trap-gay-people-9738515.html.

[17] T. Dalenius and S. Reiss. Data-swapping: A technique for disclosure control. *Journal of statistical planning and inference*, 6(1):73–85, 1982.

[18] X. Ding, L. Zhang, Z. Wan, and M. Gu. A brief survey on de-anonymization attacks in online social networks. In *2010 International Conference on Computational Aspects of Social Networks*, pages 611–615, 2010.

[19] L. Downes. A rational response to the privacy 'crisis'. *The Cato Institute, Policy Analysis*, 2013.

[20] C. Duhigg. How companies learn your secrets, 2012.

[21] C. Dwork. Differential privacy. In *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006)*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer Verlag, July 2006.

[22] C. Fabiana, M. Garetto, and E. Leonardi. De-anonymizing scale-free social networks by percolation graph matching. In *2015 IEEE Conference on Computer Communications (INFOCOM)*, pages 1571–1579, 2015.

[23] Armaed Foundation. Requirements for Personal Information Protection Part 1:U.S. Federal Law. http://armaedfoundation.org/wp-content/uploads/2016/12/FederalPrivacy.pdf.

[24] H. Fu, A. Zhang, and X. Xie. Effective social graph deanonymization based on graph structure and descriptive information. *ACM Trans. Intell. Syst. Technol.*, 6(4), July 2015.

[25] A. Gaihre, S. Pandey, and H. Liu. Deanonymizing cryptocurrency with graph learning: The promises and challenges. In *2019 IEEE Conference on Communications and Network Security (CNS)*, pages 1–3, 2019.

[26] S. Garfinkel. *De-Identifying Government Datasets*, 2016. https://csrc.nist.gov/csrc/media/publications/sp/800-188/archive/2016-08-25/documents/sp800_188_draft.pdf.

[27] M. Hay, G. Miklau, D. Jensen, D. Towsley, and P. Weis. Resisting structural re-identification in anonymized social networks. *Proc. VLDB Endow.*, 1(1):102–114, August 2008.

[28] C. Jernigan and B. Mistree. Gaydar: Facebook friendships expose sexual orientation. *First Monday*, 14, 10 2009.

[29] P. Lavrakas. *Encyclopedia of survey research methods*. Sage Publications, 2008.

[30] W. Lee, C. Liu, S. Ji, P. Mittal, and R. Lee. Blind de-anonymization attacks using social networks. New York, NY, USA, 2017. Association for Computing Machinery.

[31] H. Li, Q. Chen, H. Zhu, D. Ma, H. Wen, and X. S. Shen. Privacy leakage via de-anonymization and aggregation in heterogeneous social networks. *IEEE Transactions on Dependable and Secure Computing*, 17(2):350–362, 2020.

[32] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115, 2007.

[33] T. Li and N. Li. Injector: Mining background knowledge for data anonymization. In *2008 IEEE 24th International Conference on Data Engineering*, pages 446–455, 2008.

[34] T. Li, N. Li, and J. Zhang. Modeling and integrating background knowledge in data anonymization. In *2009 IEEE 25th International Conference on Data Engineering*, pages 6–17, 2009.

[35] N. Lomas. Tinder's handling of data is now under EU probe, 2020.

[36] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *ICDE*, 2006.

[37] D. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Halpern. Worst-case background knowledge for privacy-preserving data publishing. In *23rd International Conference on Data Engineering, ICDE 2007*, Proceedings - International Conference on Data Engineering, pages 126–135, September 2007. 23rd International Conference on Data Engineering, ICDE 2007 ; Conference date: 15-04-2007 Through 20-04-2007.

[38] C. Mauger, G.l Le Mahec, and G. Dequen. Modeling and evaluation of k-anonymization metrics. In *PrivacyPreserving Artificial Intelligence Workshop of AAAI*, 2020.

[39] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *Proceedings of the 2009 30th IEEE Symposium on Security and Privacy*, SP '09, page 173–187, USA, 2009. IEEE Computer Society.

[40] National Conference of State Legislators. Data security laws private sector.
https://www.ncsl.org/research/telecommunications-and-information-technology/data-security-laws.aspx year = 2019.

[41] T. Pham and S. Lee. Anomaly detection in bitcoin network using unsupervised learning methods. 2017.

[42] N. Richards and W. Hartzog. Taking trust seriously in privacy law. *Stanford Technology Law Review*, 2015.

[43] K. Sharad and G. Danezis. An automated social graph de-anonymization technique. WPES '14, page 47–58, New York, NY, USA, 2014. Association for Computing Machinery.

[44] N. Singer. Mapping, and sharing, the consumer genome. *The New York Times*, 2012.

[45] Philip J Smith. *A Selective Review of Confidentiality Research Published Since 1975*. Bureau of the Census, 1985.

[46] D. Solove and D. Keats Citron. Risk and anxiety: A theory of data-breach harms. *Texas Law Review*, 96, 2017.

[47] Statista. Most popular online dating apps in the United States as of September 2019, by audience size.
https://www.statista.com/statistics/826778/most-popular-dating-apps-by-audience-size-usa/: :text=As%20of%20September%202019%2C%20Tinder,5.03%20million%20U.S.%20mobile%20users.

[48] L. Sweeny. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[49] A. Tanner. Finally you'll get to see the secret consumer dossier they have on you. *Forbes*, 2013.

[50] F. Ullah, S. Jabbar, and F. Al-Turjman. Programmers' de-anonymization using a hybrid approach of abstract syntax tree and deep learning. *Technological Forecasting and Social Change*, 159:120186, 2020.

[51] Human Rights Watch. Audacity in adversity lgbt activism in the middle east and north africa.
https://www.hrw.org/report/2018/04/16/audacity-adversity/lgbt-activism-middle-east-and-north-africa_ftn115.

[52] K. Wong, O. Zaïane, F. Davis, and Y. Yasui. A machine learning approach to predict ethnicity using personal name and census location in canada. *PloS one*, 15(11):e0241239, 2020.

[53] R. Wong, J. Li, A. Fu, and K. Wang. (, k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 754–759, New York, NY, USA, 2006. Association for Computing Machinery.

[54] X. Xiao and Y. Tao. Personalized privacy preservation. In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, SIGMOD '06, page 229–240, New York, NY, USA, 2006. Association for Computing Machinery.

[55] L. Yartseva and M. Grossglauser. On the performance of percolation graph matching. In *Proceedings of the First ACM Conference on Online Social Networks*, COSN '13, page 119–130, New York, NY, USA, 2013. Association for Computing Machinery.

[56] B. Zhou and J. Pei. Preserving privacy in social networks against neighborhood attacks. *2008 IEEE 24th International Conference on Data Engineering*, pages 506–515, 2008.

[57] B. Zhou, J. Pei, and W. Luk. A brief survey on anonymization techniques for privacy preserving publishing of social network data. *SIGKDD Explor. Newsl.*, 10(2):12–22, December 2008.