NRF: A Naive Re-identification Framework

Shubhra Kanti Karmaker Santu
University of Illinois Urbana-Champaign (UIUC)
karmake2@illinois.edu

Vincent Bindschadler
University of Illinois Urbana-Champaign (UIUC)
bindsch2@illinois.edu

ChengXiang Zhai
University of Illinois Urbana-Champaign (UIUC)
czhai@illinois.edu

Carl A. Gunter
University of Illinois Urbana-Champaign (UIUC)
cgunter@illinois.edu

ABSTRACT

The promise of big data relies on the release and aggregation of data sets. When these data sets contain sensitive information about individuals, it has been scalable and convenient to protect the privacy of these individuals by *de-identification*. However, studies show that the combination of de-identified data sets with other data sets risks *re-identification* of some records. Some studies have shown how to measure this risk in specific contexts where certain types of public data sets (such as voter roles) are assumed to be available to attackers. To the extent that it can be accomplished, such analyses enable the threat of compromises to be balanced against the benefits of sharing data. For example, a study that might save lives by enabling medical research may be enabled in light of a sufficiently low probability of compromise from sharing de-identified data.

In this paper, we introduce a *general* probabilistic re-identification framework that can be instantiated in specific contexts to estimate the probability of compromises based on explicit assumptions. We further propose a baseline of such assumptions that enable a first-cut estimate of risk for practical case studies. We refer to the framework with these assumptions as the *Naive Re-identification Framework (NRF)*. As a case study, we show how we can apply NRF to analyze and quantify the risk of re-identification arising from releasing de-identified medical data in the context of publicly-available social media data. The results of this case study show that NRF can be used to obtain meaningful quantification of the re-identification risk, compare the risk of different social media, and assess risks of combinations of various demographic attributes and medical conditions that individuals may voluntarily disclose on social media.

CCS CONCEPTS

• **Security and privacy → Privacy-preserving protocols**; **Social network security and privacy**; *Formal security models*; • **Social and professional topics → Patient privacy**;

KEYWORDS

Data Privacy, Re-identification Risk, Probabilistic Framework, Formal Privacy Model, Patient Privacy

ACM Reference Format:
Shubhra Kanti Karmaker Santu, Vincent Bindschadler, ChengXiang Zhai, and Carl A. Gunter. 2018. NRF: A Naive Re-identification Framework. In *2018 Workshop on Privacy in the Electronic Society (WPES'18), October 15, 2018, Toronto, ON, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3267323.3268948

1 INTRODUCTION

Aggregation of scattered data sets is generally desirable in data mining applications for at least two reasons. First, it enables the analysis of relations between entities whose information is contained in multiple scattered data sets; it is clearly impossible to discover such connections without combining the data sets. Second, it increases the redundancy in the data sets so that an otherwise "weak" signal in an individual data set might become "stronger" in an aggregated data set if the signal exists in multiple data sets, thus increasing the chance of discovering something statistically significant. This is why topics such as data integration, joint analysis, and heterogeneous information networks analysis have all attracted much attention in recent years in order to fully exploit "big data" to discover knowledge for many applications. A good illustration of this is the National Inpatient Sample (NIS) dataset provided by Healthcare Cost and Utilization Project (HCUP) [14], which collects data about patients from multiple hospitals to provide an aggregate data set that enables discoveries that none of the participating hospitals could accomplish on its own.

However, data sets often include sensitive information from individuals, there is a privacy risk in sharing them. This is generally addressed by a combination of technical and legal means. The technical means include *de-identification* of the data of individuals by removing key attributes like their names and addresses. Such protections have proven vulnerable to attack, so Data Use Agreements (DUAs) ask the recipients of the aggregated de-identified data not to take steps that would result in *re-identification* of individuals. While such DUAs are a valuable tool, they are virtually impossible to enforce, so it is important to estimate the risk of re-identification based on technical protections.

The reasons that motivate data aggregation in the first place are also precisely the reasons why it becomes more likely to infer private information when linking together multiple data sets with information about individuals. For example, suppose an organization publishes a de-identified data set of medical records that also contains demographic information of patients. An adversary with

WPES'18, October 15, 2018, Toronto, ON, Canada

ACM ISBN 978-1-4503-5989-4/18/10…$15.00
https://doi.org/10.1145/3267323.3268948

access to an *external data set* containing names and demographic information may join the two datasets and thereby potentially re-identify medical records of some individuals. This issue has been particularly studied for de-identified medical records with external knowledge provided by public voter rolls, but the class of such studies includes other areas like de-identified social media records with external data provided by public Web postings [2, 23, 27].

Thus it is important to assess the re-identification risk due to the aggregation of data sets, especially with consideration of the increased risk due to the emergence of large public data such as social media. Failure to measure this risk raises inherent concerns from individuals contributing data, and this creates barriers for deploying such applications.

Unfortunately, how to quantify the re-identification risk due to aggregation of data sets is a difficult scientific and practical challenge. Multiple questions must be addressed formally. First, how do we formally frame the problem scenario of re-identification involving external knowledge? Second, how can we formally define measures to quantify the risks? Third, how can we compute those measures? Finally, how can we evaluate such measures? Existing work has addressed some of these questions, but they largely remain unanswered.

In this paper, we make a first attempt to address these questions by developing a general probabilistic framework for analyzing and quantifying the re-identification risk in the context of data aggregation. The framework consists of three models: a de-identified data model, a model of external knowledge, and a risk model. From these it is possible to find the probability that an individual may be re-identified due to the availability of external knowledge. Specifically, our framework combines the adversary's external knowledge with the statistical properties of the de-identified data set to quantify the re-identification risk associated with the release. The goal is an intuitive probabilistic risk metric that can be computed on a data set (before releasing it) to assess the potential risk from releasing the data set with consideration of the adversary's external knowledge. It is then possible to find the probability that an individual may be re-identified due to the availability of external knowledge. To keep the computation of the model parameters feasible and thus, make the model practically useful, we propose an approximation of the general framework, which we call NRF, the Naive Re-identification Framework. NRF will provide a principled way to analyze how re-identification risk might be impacted by changes in the available external information. This can be done by comparing the risks associated with diverse external knowledge sources, thus facilitating research on the significance of the existing public sources such as different social media.

But why have a *naive* framework rather than a sophisticated one? A sophisticated technique that cannot be used because it cannot be instantiated will be less useful than a naive technique that gets an approximate result. We are partly inspired by Naive Bayes classifiers, which make assumptions that are only partially true in many cases, but get meaningful results anyway. A more grandiose analogy would be to describe the flight of a ball as a parabola even though it is understood that accounting for friction with the air would require a more sophisticated model.

As a case study to illustrate NRF, we show how we can apply the model to analyze the risk of aggregating a (de-identified) medical data set (the HCUP data set) and five social media data sets (Facebook, Twitter, Instagram, Pinterest, and LinkedIn). Experimental results show that the proposed framework can be used to obtain meaningful probabilities of re-identification risk and enables us to make a number of conclusions:

1. The aggregation of HCUP with social media poses high risk of re-identification, especially when multiple social media are aggregated.
2. Among all the five social media we studied Facebook poses the highest risk, followed (in order) by LinkedIn, Pinterest, Instagram, and Twitter.
3. Among the many attribute values users voluntarily "leaked" on social media, the demographic attributes pose the greatest risk of enabling re-identification, followed by the medical conditions that are disclosed in the medical data set.

Overall, these results show the great promise of the proposed framework both as a general framework that can be further refined and as a way to derive specific probabilistic risk models that can be immediately applied to help assess re-identification risks before releasing a data set and analyzing security aspects of social media.

The rest of the paper is organized as follows: Section 2 discusses the related works in the literature. Section 3 presents a brief motivating example. Section 4 defines measure r^* and describes the general risk model in detail. Section 5 presents NRF as an approximation of the general risk model by incorporating some simplifying assumptions. Section 6 describes a real life application of NRF with the details of experimental setup, parameter estimation techniques and results. Section 7 draws the conclusion and points to possible future directions.

2 RELATED WORKS

We develop a probabilistic model in order to quantify the risk of re-identification of a relational data set in association with a public relational data set. Our work is related to multiple lines of existing work, which we review briefly.

De-identification techniques are widely used to protect users' private information in public data sets. Typical examples of such sanitization procedures include removal of some records (suppression), reduction in the precision of certain attributes (generalization), and adding noise to the records (randomization) [19]. However, the success of these techniques in mitigating the risk of re-identification has not been systematically demonstrated or investigated. Furthermore, researchers have demonstrated several re-identification attacks against de-identified data sets. These attacks have been successful in re-identifying records in various kinds of de-identified data sets including medical records [27], movie ratings [23], and web search queries [1]. In fact, according to a systematic study on publications related to re-identification attacks in 2011, an overall mean proportion of successful re-identification is 0.262 for all studies and 0.338 for health data [12]. In particular, Loukides et al. [19] have shown the feasibility of re-identification attacks against patient clinical data. However, all these works on re-identification attacks are primarily case-studies where they consider some particular data set and the focus is to demonstrate its vulnerability in terms of re-identification of individuals. Our work is different from them because we propose a general and formal risk-model

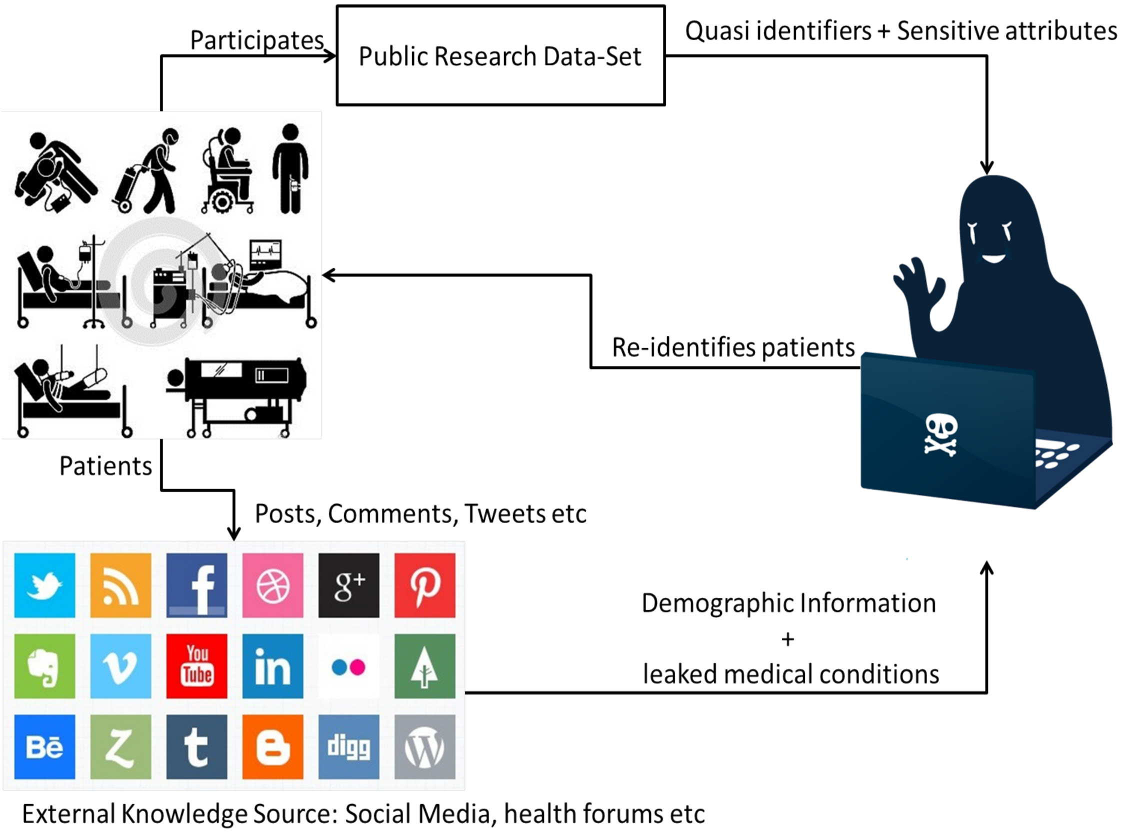

Participates
Public Research Data-Set
Quasi identifiers + Sensitive attributes
Re-identifies patients
Patients
Posts, Comments, Tweets etc
Demographic Information
+
leaked medical conditions
External Knowledge Source: Social Media, health forums etc

Figure 1: Motivating Example.

to evaluate the re-identification risks associated with the release of a de-identified data set with consideration of public external data. The proposed model can thus be applied to any relational data set. Another limitation of the previous works is that they assume the adversary knows for sure that the target individual's record is included both in the de-identified data set and the external knowledge data set. However, this assumption may not hold often in real-life. Our approach overcomes this limitation by introducing a probabilistic definition of the re-identification risk where the action of an adversary is modeled in two steps. First, the adversary finds out individual records with unique quasi-identifier values from the de-identified data set; second, he/she tries to match these individual records against the external sources being agnostic of whether the external sources actually contain those individual records or not. The risk is defined as the probability of the adversary's success in a unique matching that results in re-identification.

There are two well-established approaches to mitigate the risk of re-identification attacks on de-identified data sets; those based on syntactic metrics, and those based on differential privacy. Syntactic metrics such as k-anonymity [28], l-diversity [21], and t-closeness [17] directly quantify syntactic properties of a sanitized data set. For example, k-anonymity says that for every quasi-identifier group there should be at least k records. Such metrics can serve as a way of evaluating the privacy risk of releasing de-identified data sets when the syntactic property measured is correlated with the privacy risk. A downside of these techniques is that they largely ignore the attacker's external knowledge. Indeed, while quasi-identifiers are defined to be attributes likely to be found in external sources such as publicly available data, there is no attempt to quantify the likelihood that an adversary would know a particular attribute of a given target individual. Our model overcomes this limitation of the previous works by incorporating the disclosure probability of the quasi-identifiers (called “Disclosure-Likelihood" in our proposed model) given the external knowledge model.

Beyond techniques such as [4] which combine ideas from the work on syntactic metrics, frameworks such as ϵ-privacy [20] attempt to capture several privacy metrics. One such metric is differential privacy [8, 9]. Unlike syntactic metrics, differential privacy is independent of the protected data set and the attacker's background knowledge (i.e., it holds regardless of how much the attacker knows). While differential privacy and its many variants are useful to protect data sets, none is suitable as a quantification metric of the risk. Indeed as pointed out in [15] the ϵ in differential privacy is not a probability of identification or a measure of risk.

Evaluating the re-identification risk of health data has been studied extensively, for example in [10], [11], and [18]. Of particular interest is the expert determination guidelines of the HIPAA privacy rules [24] which recommends assessment of the risk along three axes: replicability, data source availability (sometimes called knowability), and distinguishability. If all three are low, this indicates a low risk, whereas if one or more is high, it indicates a high risk. Unfortunately, the HIPAA privacy rule does not recommend a specific method to calculate this risk or provide a threshold for it. In contrast, our risk-model captures knowability (external knowledge model) and distinguishability (data set model) and provides associated equations to calculate the risk.

There is also work on privacy risks in online social networks. For example, [30] considers the risk that patients will be re-identified due to what medical students say on social media. Biega et al. [3] proposes a probabilistic framework to evaluate the privacy risk of users' search histories. It should be noted that this work has a different goal than ours because it measures the risk of attribute disclosure, whereas we measure re-identification risk. This is not a trivial distinction because attribute disclosure can occur without re-identification or even if the target record is not in the data set. More broadly, there is work on discovering and linking the different profiles of a user across several social networks, for example [31] and [26].

3 A MOTIVATING EXAMPLE

To motivate our work and explain the rationale behind the design of the proposed probabilistic framework, we first present a toy example (refer to Figure 1) of the re-identification framework. Suppose that, there is a data set D which has been released publicly for medical research purposes. The data set contains sensitive information (e.g. DNA sequence, critical diagnosis etc) about patients along with some non-sensitive information (demographics, non-critical diagnosis etc) about the same. However, D has been de-identified to ensure that it does not contain any information that can help identify a single patient. Now, suppose that there is some adversary who is interested in identifying individual patients from D, i.e., the adversary tries to re-identify some individuals in D using some external information sources. We denote the adversaries external knowledge by E. We assume E consists of non-sensitive information like patient demographics, address, non-critical diagnosis etc. The adversary can then try to match E with D to re-identify patients and extract sensitive information that was not meant to be released and thus the patients privacy gets compromised. The schematic diagram of this re-identification model is shown in Figure 1.

To see a more concrete example, refer to Table 1. Suppose that, A and B are two patients who participates in dataset D, which is released publicly after anonymization of the individuals. Let us assume, D contains 3 attributes, namely, X, Y and Z, where, X and Y are non-sensitive attributes (also called quasi-identifiers) and Z is a sensitive attribute. Each cell represents the value of corresponding attribute of corresponding individual. For example, the value of attribute X for patient A is x_a. In absence of any other information source, the probability that some individual from this dataset will be (uniquely) re-identified by an adversary is 0.0.

However, the scenario get complicated if the adversary has access to some external information sources, e,g, social media, where the users disclose their real identities. Assume that both patient A and B use a common social forum E and the adversary can access the public data from E. Now, if $x_a \neq x_b$ or $y_a \neq y_b$ and if X, Y are publicly disclosed in E, then, for certain, the adversary will easily re-identify patients A and B from dataset D and will know that, A has $Z = z_a$ and B has $Z = z_b$. This means the re-identification probability in this case is 1.0. However, if $x_a = x_b$ and $y_a = y_b$, then, even if X, Y are publicly disclosed in E, its impossible for the adversary to pinpoint A or B from dataset D as they are no longer unique. Thus, probability of (unique) re-identification becomes 0.0. The situation becomes further complex, when the disclosure of attribute X or Y in external source E becomes uncertain. For example, if the disclosure probability of Y is zero and disclosure probability of X is 0.5, then, the probability of re-identification of either A or B becomes $1.0 - 0.5 \times 0.5 = 0.75$ (assuming $x_a \neq x_b$). Things get even more complex when the uniqueness of attribute A or B or combination of (A, B) becomes uncertain.

Table 1: Toy example data.

Patient	X	Y	Z
A	x_a	y_a	z_a
B	x_b	y_b	z_b

The proposed framework intends to formally capture the intuitions discussed above in a general way. In the next section, we present the proposed re-identification framework followed by the simplification of estimation of different parameters of the framework, which we call the *Naive Re-Identification Framework*, i.e., NRF.

4 A GENERAL RE-IDENTIFICATION FRAMEWORK

Our goal is to formally analyze the potential risk of re-identification due to releasing a de-identified data set D when the adversary has access to a general external knowledge source denoted by E. (Note that in the case when we have multiple external sources, we may simply join all of them to form one unified external knowledge source.) Such analysis is necessary before releasing any data set due to the increased risk of re-identification from aggregation of a released data set with additional ("external") data sets that an adversary may have access to (e.g., public social media). While some existing work has addressed this problem in an application-specific way (see Section 2), we hope to develop a general probabilistic framework that can be applicable to analysis of any relational data set. A probabilistic metric of risk has the advantage of being easy to interpret.

However, without any further assumption about the data, how can we formally define the re-identification risk? To address this question, we start with a systematic analysis of the process of re-identification. From the adversary's perspective, re-identification of individuals involves the following two different cases:

The first case is when the adversary has a particular target in mind. Here, the adversary would first look at what attributes the particular target has disclosed in public forums/social media, and then see if he/she can locate the corresponding target record in the de-identified data set using those disclosed attributes as *quasi-identifiers*, which we define as those attributes that are not in and of themselves unique identifiers, but which are sufficiently associated to an individual record that they can be combined with each other to yield a unique identifier.

The second case is when there is a more ambitious adversary who may not have a particular target in mind, but wants to re-identify as many individual records as possible from the de-identified data set. In this case, the adversary extracts individuals with unique combination of values for different sets of quasi-identifiers from the de-identified data set. He/she then matches these quasi-identifiers with his external knowledge source to re-identify individuals' records and link them with their real identities.

In both cases, successful re-identification attack would depend on two factors: 1) Availability of quasi-identifiers in the de-identified data set that can uniquely identify individual records and 2) Disclosure of these quasi-identifiers in the public domain (e.g., forums/social media) which the attacker can exploit as external sources of information. To capture the first factor, we introduce a *data set Model*: the more quasi-identifiers the de-identified data set contains, the higher the probability that some individual will be re-identified. To capture the second factor, we introduce an *External Knowledge Model*: the more people disclose some quasi-identifiers on the public forums, the more likely the attacker knows them and uses them to re-identify individuals. Finally, we can define the risk model based on the *re-identification risk* with respect to the *Data Set Model* and *External Knowledge Model.* Thus, the overall framework consists of three components: (1) Data set Model, (2) External Knowledge Model, and (3) Risk Model, which will be further explained in detail below.

Notation. We will use bold-uppercase letters, e.g. $\mathbf{X}$, to denote random variables, and lowercase letters, e.g., x, to denote probabilities. For example, we use $P(\mathbf{X} = x)$ to denote the probability of the event that random variable X is x. This convention is followed throughout the paper. Table 2 summarizes the notations and symbols used.

4.1 Data Set Model

We first need to model the de-identified data set D, which is to be released. D would be assumed to contain attribute values of a set of individuals that we would like to protect (e.g., patients in a medical data set). Let X be the set of all attributes in the published data set D (e.g., age, diagnosis codes, medication etc. in a medical data set). Consider any subset of X, $C \subseteq X$ (e.g., C={age, medication} in a medical data set). We introduce a binary random variable $\mathbf{U_C}$ to represent the unique identification status of any individual with respect to the set of attributes C, and define it as follows: $\mathbf{U_C} = 1$ if some random individual can be uniquely identified (within D) with respect to quasi-identifier set C (i.e., by just looking at the values of the attributes in C), and 0 otherwise.

We can now define the concept of *D-Uniqueness* (Definition 4.1) which is essentially the probability that $\mathbf{U_C} = 1$.

Definition 4.1 (D-Uniqueness). *Given a set of attributes $C \subseteq X$, the* D-Uniqueness *of C, u_C, is the probability that a randomly chosen individual from D would contain a unique combination of values for the set of attributes C within data set D, i.e., $u_C = P(\mathbf{U_C} = 1)$. The constraint $P(\mathbf{U_C} = 1) + P(\mathbf{U_C} = 0) = 1$ holds true.*

Table 2: Table of notations. All random variables and probabilities (except $\mathbf{r}^*_{\mathbf{D}}$) are defined for a random individual.

D	De-identified data set
E	External source
X	Attributes of D
x	A specific attribute of D
C	A subset of X
$\mathbf{U_C}$	Unique identification status with respect to C in D
u_C	Uniqueness with respect to C in D, i.e., $P(\mathbf{U_C} = 1)$
$\mathbf{Z_x}$	Disclosure status of attribute x
z_C	Disclosure-Likelihood of $C \subseteq X$, i.e., $P(\mathbf{Z_C} = 1)$
z_x	Disclosure-Likelihood of $x \in X$, i.e., $P(\mathbf{Z_x} = 1)$
$\mathbf{Q_C}$	Unique identification status with respect to C in E
q_C	Disclosure-Uniqueness w.r.t. $C \subseteq X$ in E, i.e., $P(\mathbf{Q_C} = \mathbf{1} \mid \mathbf{U_C} = \mathbf{1})$
r^*_D	Re-identification risk for whole data set D
$\mathbf{R_D}$	Individual re-identification status within D
r_D	Individual re-identification risk within D, i.e., $P(\mathbf{R_D} = 1)$
τ_i	Usage status of public information forum i

Our data set model is essentially a key/value pair, where each key C is a distinct subset of all attribute set X, the value is the corresponding *D-Uniqueness*, i.e., u_C. Intuitively, a data set with high *D-Uniqueness* values for a large number of different subsets of attributes is more vulnerable in terms of re-identification than a data set with low *D-Uniqueness* values for the same.

4.2 External Knowledge Model

For our purpose of analysis, the relevant background/external knowledge of the attacker (denoted by E) is any attribute-value pair that can be joined with those in D. In particular, we are interested in the probability that the value of an attribute $x \in X$ will be disclosed to the adversary via availability of E. We use $\mathbf{Z_x}$ to represent the disclosure status of attribute x, which is defined as: $\mathbf{Z_x} = 1$ if the value of attribute x has been disclosed to the adversary, and 0 otherwise.

We now introduce the notion of Disclosure-Likelihood (Definition 4.2) which is essentially the probability that $\mathbf{Z_x} = 1$.

Definition 4.2 (Disclosure-Likelihood). *The* Disclosure - Likelihood *of an attribute x, z_x, is the probability that the value of attribute x of a randomly chosen individual from D is disclosed to the attacker, i.e., $z_x = P(\mathbf{Z_x} = 1)$. The constraint $P(\mathbf{Z_x} = 1) + P(\mathbf{Z_x} = 0) = 1$ holds true.*

In general, the probability of disclosing different quasi-identifiers to the adversary is likely to be different, i.e., z_x is not the same for all x. The definition can be easily generalized to the Disclosure-Likelihood of a group of attributes C, $z_C = P(\mathbf{Z_C} = 1)$ (see equation 4 for details).

Note that, disclosing a set of attributes alone is not sufficient for a successful re-identification attack; rather, the set of values (for those attributes) disclosed must also be unique among all the records found in the external source E (We assume E as just another data set here). Otherwise, it will not be possible to uniquely match the individual records in D against the same in E. To capture this requirement, we use $\mathbf{Q_C}$ to represent the uniqueness of the values of an attribute set C in external source E, which can be defined as: $\mathbf{Q_C} = 1$ if some random individual has a unique combination of values for the attribute set C within the entire data set E (i.e., no other individuals have exactly the same values for all these attributes in C), and 0, otherwise.

Finally, we introduce the notion of *E-Uniqueness* (Definition 4.3) which is the probability that $\mathbf{Q_C} = 1$. Note that, *E-Uniqueness* is different from the notion *D-Uniqueness* defined in Section 4.1.

Definition 4.3 (E-Uniqueness). *The* E-Uniqueness *of a set of attributes $C \subseteq X$, q_C, is the* conditional probability *that a randomly chosen individual from E would have a unique combination of values for the set of attributes C among all the individuals within E given that the same combination of values for C is also unique within the de-identified data set D, i.e., $q_C = P(\mathbf{Q_C} = \mathbf{1}|\mathbf{U_C} = \mathbf{1})$.*

4.3 The Risk Model

With the models for the de-identified data set D and external knowledge E in place, we can now define the last component in the framework, i.e. the risk model. Let $\mathbf{R_D}$ be a random variable which represents the re-identification status of some random individual in D. That is, $\mathbf{R_D} = 1$ if some random individual is re-identifiable (from D) by the adversary, and 0 otherwise. We first start by introducing the definition of re-identification risk of an individual and later extend it to the re-identification risk for the entire data set.

Definition 4.4 (Individual Re-identification Risk). *Given an external knowledge model E, the* individual re-identification risk *r_D (with respect to publishing data set D) is the probability that an adversary will re-identify a random individual from D, i.e., $r_D = P(\mathbf{R_D} = 1)$.*

How can we compute this risk? To address this question, let us examine the conditions that must be satisfied in order for a re-identification to happen. Logically, successful re-identification of individual t requires the adversary to identify a set of quasi-identifiers C such that the combination of values of the attribute set C for the individual t is unique in both D and E and the values of C for this individual are disclosed to the adversary through external knowledge E. This analysis gives us the following equation of r_D :

$$r_D = 1 - \prod_{C \subseteq X} [1 - \{P(\mathbf{U_C} = 1) \cdot P(\mathbf{Z_C} = 1|\mathbf{U_C} = 1) \\ \cdot P(\mathbf{Q_C} = \mathbf{1}|\mathbf{U_C} = \mathbf{1}, \mathbf{Z_C} = \mathbf{1})\}] \quad (1)$$

where we consider all the possible C that the adversary might have leveraged by taking a product over all of them. Inside the product, there is again a product of three terms, which are intuitively very meaningful and reflect the following three conditions to be satisfied for the re-identification to happen: 1) The adversary must find a set of quasi-identifiers C such that the combination of values of the attribute set C for an individual t is unique in D (captured by $P(\mathbf{U_C} = 1)$). 2) Values of C for individual t are disclosed to the adversary through external knowledge E given that C uniquely identifies t (captured by $P(\mathbf{Z_C} = 1|\mathbf{U_C} = 1)$). 3) Combination of values of the attribute set C for t is unique in E given that t is also unique with respect to C in D and C has already been disclosed to the adversary (captured by $P(\mathbf{Q_C} = \mathbf{1}|\mathbf{U_C} = \mathbf{1}, \mathbf{Z_C} = \mathbf{1})$). Thus, the adversary will re-identify individual t only if t is unique with respect to attribute set C in both D and E and also C has been disclosed to the adversary, which the product $P(\mathbf{U_C} = 1) \cdot P(\mathbf{Z_C} = 1|\mathbf{U_C} = 1) \cdot P(\mathbf{Q_C} = \mathbf{1}|\mathbf{U_C} = \mathbf{1}, \mathbf{Z_C} = \mathbf{1})$ nicely captures. $[1 - P(\mathbf{U_C} = 1) \cdot P(\mathbf{Z_C} = 1|\mathbf{U_C} = 1) \cdot P(\mathbf{Q_C} = \mathbf{1}|\mathbf{U_C} = \mathbf{1}, \mathbf{Z_C} = \mathbf{1})]$ is simply the probability that the adversary fails to re-identify individual t based on the subset of attributes C. Finally, the outer product over all attribute set gives the probability that no subset of attributes would suffice to re-identify individual t and by subtracting it from 1, we get our desired re-identification probability.

Based on the individual re-identification risk, we can now further define the overall *re-identification risk* as the probability of re-identification of at least one individual by the adversary.

Definition 4.5 (Re-identification Risk). *Given some particular external knowledge model E, the* re-identification risk *r^*_D (with respect to publishing a data set D) is the probability that an adversary will re-identify at-least one individual whose record is in D.*

An implicit assumption made by our definition of the risk is that the re-identification of any individual would be harmful, which is reasonable. Re-identification of multiple individuals is captured by the "*at-least*" phrase in the definition.

This overall re-identification risk r^*_D can be computed based on the individual re-identification risk as follows:

$$r^*_D = 1 - \prod_{i=1}^{n} [1 - r_D] = 1 - \prod_{i=1}^{n} [1 - P(\mathbf{R_D} = 1)] \quad (2)$$

where n is the total number of participants in data set D and $\mathbf{R_D}$ is the re-identification risk of some individual in D.

Equations 1 and 2 form the general probabilistic model for assessing the risk of re-identification due to aggregation of data set D and external knowledge E. While at this point, the model is not yet "computable," the framework takes a necessary first step toward formalizing the re-identification risk in a very general way with consideration of external knowledge in probabilistic terms, and thus can serve as a roadmap for further refinement of each of the component probabilities in the framework. The three concepts we defined (i.e., *D-Uniqueness*, *Disclosure-Likelihood*, and *E-Uniqueness*) are all essential for formally characterizing the re-identification risk.

There are potentially many different ways to further refine the probability terms involved in these equations, but a thorough discussion of them is beyond the scope of this paper. Below, we discuss one line of instantiation of the framework, which we will use later for experimental study with a medical data set as D and social media as E.

5 NAIVE RE-IDENTIFICATION FRAMEWORK

A main challenge in applying the proposed formal framework to a specific application data set is to estimate all those probabilities. As always happens in statistical estimation, there is an inevitable tradeoff between the accuracy of a probabilistic model and the feasibility of parameter estimation. Specifically, a sophisticated model that does not rely on any simplification assumptions can be expected to be more accurate for modeling the re-identification risk, but such a model also tends to have many more parameters and thus require many more data for accurate parameter estimation or require data that we do not have available. Thus as a first attempt to make the estimation (computation) tractable, we explored a relatively simple instantiation of the framework where we introduced multiple independence assumptions to make it much easier to estimate the involved probabilities. As those assumptions are similar to the independence assumptions made in the popular and effective Naive Bayes Bayesian classification (i.e., the Naive Bayes classifier [16]), we call this simplified framework Naive Re-identification Framework (NRF). Although the assumptions do not actually hold in reality, the NRF may still be able to provide us with useful approximations just as the Naive Bayes classifier is still quite useful even though the independence assumptions made are not true. Of course, further exploration of how to refine the framework without making these independence assumptions is obviously a very important future work.

We introduce two independence assumptions about D-Uniqueness and E-Uniqueness.

Assumption 5.1 (Disclosure D-Uniqueness Independence). *Given a set of attributes $C \subseteq X$, the* Disclosure-Likelihood *of C is independent of its* D-Uniqueness. *Mathematically,*

$$P(\mathbf{Z_C} = 1 | \mathbf{U_C} = 1) = P(\mathbf{Z_C} = 1)$$

Assumption 5.2 (Disclosure E-Uniqueness Independence). *Given a set of attributes $C \subseteq X$, the* E-Uniqueness *of C is independent of its* Disclosure-Likelihood. *Mathematically,*

$$P(\mathbf{Q_C} = \mathbf{1} | \mathbf{U_C} = \mathbf{1}, \mathbf{Z_C} = \mathbf{1}) = P(\mathbf{Q_C} = \mathbf{1} | \mathbf{U_C} = \mathbf{1})$$

Assumption 5.1 and 5.2 are generally not true, but are fairly reasonable because, intuitively, attributes' disclosure behavior would hardly depend on how unique their value is and vice-versa. Applying assumption 5.1 and 5.2, Equation 1 reduces to the following:

$$P(\mathbf{R_D} = 1) = 1 - \prod_{C \subseteq X} [1 - u_C \cdot z_C \cdot q_C] \tag{3}$$

where, we have $u_C = P(\mathbf{U_C} = 1)$, $z_C = P(\mathbf{Z_C} = 1)$, and $q_C = P(\mathbf{Q_C} = \mathbf{1} | \mathbf{U_C} = \mathbf{1})$.

Here, the first term (u_C) in the inner product is the *D-uniqueness* of attribute set C with respect to data set D, i.e., the probability that the subset of attributes C will uniquely identify any random individual t in D. The second term (z_C) is the *Disclosure-Likelihood* of the subset of attributes C, i.e., the probability that the values of the attribute set C of individual t will be disclosed to the adversary. The third term (q_C) is the *E-uniqueness* of attribute set C with respect to the external knowledge data set E, i.e., the probability that the subset of attributes C will uniquely identify individual t within E given that the combination of values of C is unique for individual t within the de-identified data set D.

The probability $P(\mathbf{U_C} = 1)$ can be directly computed from the data set D itself, whereas $P(\mathbf{Z_C} = 1)$ and $P(\mathbf{Q_C} = \mathbf{1} | \mathbf{Z_C} = \mathbf{1})$ need to be estimated from external knowledge sources E. Due to sparseness of data used for experiments, we further make an independence assumption about disclosure of different attributes:

Assumption 5.3 (Independent Disclosure). *The probability of disclosing attribute x to the attacker is independent of the probability of disclosing attribute $x' \neq x$.*

Compared with Assumption 5.1 and 5.2, this assumption is less realistic since some attributes may tend to be mentioned together in a data set, thus their disclosures would be correlated. However, this assumption is necessary in order to alleviate the problem of data sparseness; it is straightforward to introduce more dependency between different attributes, but our data set would now allow us to collect sufficient counts for all the combinations of attribute values. Such dependent models, however, may be feasible for other data sets, and should be interesting to explore in the future.

Using Assumption 5.3, $P(\mathbf{Z_C} = 1)$ can be computed by multiplying the probabilities associated with the individual attributes (Equation 4).

$$P(\mathbf{Z_C} = 1) = \prod_{x_j \in C} P(\mathbf{Z_{x_j}} = 1) \prod_{x_j \in X \setminus C} \Big[1 - P(\mathbf{Z_{x_j}} = 1)\Big] \tag{4}$$

Here, x_j denotes a particular attribute, i.e., $x_j \in X$.

Given this, the next question is how to estimate the probabilities $P(\mathbf{Z_{x_j}} = 1)$. Without loss of generalities, let us assume that there are m different public information sources / forum (e.g., social media platforms) from which the adversary may gather information about target individuals. In this case, if some individual t whose data is in D also (voluntarily) discloses attribute x_j in any of these m public information sources, the adversary will learn x_j and may use this to re-identify t.

Let $\tau_\mathbf{i}$ be a family of random variables defined for each public information source / forum i. Given a i, $\tau_\mathbf{i}$ represents the membership status of a random individual t in forum i. That is, $\tau_\mathbf{i} = 1$ if some random individual t is a member of forum i, and 0 otherwise. With this, $P(\mathbf{Z_{x_j}} = 1)$ can be computed as follows:

$$P(\mathbf{Z_{x_j}} = 1) = 1 - \prod_{i=1}^{m} \Big[1 - P(\tau_i = 1) \cdot P(\mathbf{Z_{x_{ij}}} = 1)\Big], \tag{5}$$

where $P(\tau_i = 1)$ is the probability that some random individual t is a member of forum i, and $P(\mathbf{Z_{x_{ij}}} = 1)$ is the probability that t will disclose attribute x_j in forum i (given that t is a member of forum i).

To evaluate this special instance of the risk-model, we need to estimate the values of the parameters shown in Table 3. Note that this process depends on the specific context and application scenario. Section 6 presents the details of parameter estimation for a particular case study. Once the parameters have been estimated, they can simply be plugged into the risk model to compute the re-identification risk associated with the release of a data set D.

6 APPLICATIONS OF THE FRAMEWORK

In this section, we apply the refined risk model described in the previous section to analyze the risk from aggregating a specific

Table 3: **Parameter Summary for the Risk Model.**

Parameter	Description	**Ref. Eqn.**
n	Total participants in D	2
$P(\mathbf{U_C} = 1)$	*D-Uniqueness* of attribute set C w.r.t. D	3
$P(\tau_i = 1)$	Membership probability for social media i	5
$P(\mathbf{Z_{x_{ij}}} = 1)$	*Disclosure-Likelihood* of x_j in public forum i	5
$P(\mathbf{Q_C} = \mathbf{1} \mid \mathbf{U_C} = \mathbf{1})$	*E-Uniqueness* of attribute set C w.r.t. E	3

medical data set with social media data to demonstrate its usefulness. Our purpose is to use this example to illustrate how exactly the proposed model can be used to analyze the risk in practice and also to examine whether the results we will obtain by using the framework are meaningful and useful. The model, however, is completely general, and can thus be potentially applied to many other data sets to perform similar analysis as done here.

6.1 Problem Scenario

Consider a scenario in which a hospital or a medical center wants to encourage research by publicly releasing a de-identified data set containing patients' medical records and demographic information. Naturally, such a data set would contain information deemed non-sensitive (e.g., demographic attributes, common/temporary medical conditions such as having a cold or fever), as well as extremely sensitive information such as rare or stigmatized medical conditions (e.g., HIV, cancer). The privacy risk here comes from the fact that after an adversary acquires the de-identified data set, he/she can use the demographics and non-sensitive medical information as quasi-identifiers for a re-identification attack. In addition, the adversary may gather information from external sources such as social media, or public voter rolls, and use it to match individuals with similar or identical quasi-identifier values. This matching is possible when demographic information about individuals alongside with their real identities are publicly available. In practice, such information is frequently publicly available on online social networks or public forums, and it can be gathered by browsing user profiles. Further, people also often discuss their own illnesses and medical issues on online health forums and social networking sites. This presents another avenue through which information is disclosed to an adversary. In order to avoid prematurely releasing a data set with high risk of re-identifidcation, it is very important to be able to quantify the risk of re-identification before we actually release the data set. Below we show that our proposed model enables us to do this.

6.2 Data sets

We use two real-world data sets in our experiments. As the (de-identified) medical research data set, we used the HCUP National (Nationwide) Inpatient Sample (NIS) [13]. This data set contains 8 million de-identified inpatient medical records and 135 different attributes such as demographics, ICD-9 codes for diagnosis and procedures, and cost related attributes. For our experiments, we only consider the 4 demographic attributes and 12 frequently disclosed medical conditions (Table 4). (Those 12 medical conditions were chosen based on their high frequency of disclosure in the crawled Twitter data set.) As external source of information, we use a large number of Twitter post, crawled in 2015. This data set which contains more than 310 millions random tweets was obtained from Twitter (between February and October 2015). We use this data set as an example of external source of information that an adversary can take advantage of. Each tweet contains 27 different attributes including the tweet's text, user information, location, and language of the tweet. We only used the text attribute of the tweets and discarded the rest.

Table 4: **Demographics and Medical Conditions Attributes.**

Demographics	**Medical Conditions**
Age, Gender, Race, Location	Aches, Poisoning, Meningitis, Migraine, Arthritis, Insomnia, Diabetes, Fever, Asthma, Acne, Ulcer, Anemia

The HCUP data set contains the medical conditions (i.e., diagnosis information) as ICD-9 codes. These are 3 to 5 digit codes whose exact values are unlikely to be known by an adversary. Indeed, external information sources such as social media platforms are more likely to contain medical information in the form of commonly used keywords rather than their exact ICD-9 codes. Therefore, we consider an adversary with knowledge of medical conditions by their popular or commonly used names. To capture this, we collected all ICD-9 codes related to the 12 medical conditions considered (Table 4) and transformed each code into the general medical term to which it is most closely related. For example, ICD-9 code 493.01 and 493.02 stand for "Extrinsic asthma with status asthmaticus" and "Extrinsic asthma with (acute) exacerbation", respectively. We replaced both 493.01 and 493.02 with the commonly used term "asthma". With this transformation, multiple ICD-9 codes were merged into a single medical-condition term or concept. Note that this makes it harder for an adversary to re-identify patients as the diversity in the values of attributes decreased.

6.3 Estimation of the parameters

In order to apply the proposed model to analyze our data sets, our main task is to estimate the risk model parameters described in Table 3 based on the data sets in this scenario. We now discuss how to estimate these parameters. 1) n is simply the total number of records in D. 2) We estimate parameter $P(\mathbf{U_C} = 1)$ (Definition 4.1) based on the HCUP data set. The latter contains 16 different attributes (4 demographic and 12 medical-conditions), yielding a total of $2^{16} = 65536$ valid subsets of attributes. For each of these, we calculated the portion of the individuals who had unique set of values for that particular subset and used this proportion as an estimate of $P(\mathbf{U_C} = 1)$. 3) $P(\tau_i = 1)$ is the probability that any random patient is a member of social media/public forum i. Social media usage statistics data can provide reliable estimates for the values of τ_i [7]. Thus we set $P(\tau_i = 1)$'s directly based on their respective usage statistics (Table 5). 4) $P(\mathbf{Z_{x_{ij}}} = 1)$ is the probability that some random

Table 5: **Social media platforms with usage percentage.**

External Source (i)	Usage (τ_i)
Twitter	19%
Instagram	21%
Pinterest	22%
LinkedIn	23%
Facebook	58%

patient will disclose some attribute x_j (demographic attribute or medical-condition) in social media i.

For a demographic attribute x_j, $P(\mathbf{Z}_{\mathbf{x}_{ij}} = 1)$ can be estimated based on the portion of users sharing their demographic attribute x_j on public forum i plus proportion of users for which the attacker can infer x_j from public forum i. In our experiments, we used values obtained from [5, 6, 25, 29] and assumed that these values were the same across all social media platforms. See Table 6 for details. In the case of medical conditions, we approximate $P(\mathbf{Z}_{\mathbf{x}_{ij}} = 1)$ with the probability that an individual will disclose a disease or medical condition that he/she actually suffers from, i.e., $P(\mathbf{Z}_{\mathbf{x}_{ij}}{}^{+} = 1|x_j{}^{+})$, which can be formally written as follows:

$$P(\mathbf{Z}_{\mathbf{x}_{ij}}{}^{+} = 1|x_j{}^{+}) = \frac{P(x_j{}^{+}|\mathbf{Z}_{\mathbf{x}_{ij}}{}^{+} = 1) \cdot P(\mathbf{Z}_{\mathbf{x}_{ij}}{}^{+} = 1)}{P(x_j{}^{+})} \tag{6}$$

where we may assume $P(x_j{}^{+}|\mathbf{Z}_{\mathbf{x}_{ij}}{}^{+} = 1) = 1$, i.e., whenever an individual discloses (on public forums or on a social media platform) that he/she is suffering from some medical condition, he/she is telling he truth, i.e., he/she does indeed suffer from that condition. With this assumption, we get the following:

$$P(\mathbf{Z}_{\mathbf{x}_{ij}}{}^{+} = 1|x_j{}^{+}) = \frac{P(\mathbf{Z}_{\mathbf{x}_{ij}}{}^{+} = 1)}{P(x_j{}^{+})} . \tag{7}$$

Here, $P(\mathbf{Z}_{\mathbf{x}_{ij}}{}^{+} = 1)$ is the probability that an individual will disclose on public forum/social media i that he/she suffers from medical condition x_j. $P(x_j{}^{+})$ is the probability that any random individual actually has the medical condition x_j. Estimation of $P(x_j{}^{+})$ for the 12 medical conditions were based on the National Health Interview Survey [25]. To estimate $P(\mathbf{Z}_{\mathbf{x}_{ij}}{}^{+} = 1)$, we collected all sentences from social media i (twitter in this case) that mention word x_j (e.g., asthma) or some other synonyms of x_j. Let this set be S_{x_j} which consists of candidate sentences that may disclose that some patient suffers from condition x_j. We can then classify each sentence in S_{x_j} into one of the two classes: M_{x_j}, composed of those sentences disclosing suffering from condition x_j, and $\bar{M}_{x_j}$, composed of those sentences that do not disclose suffering from condition x_j (e.g., those to raise social awareness about x_j). Specifically, We used a lexicon-based approach to decide which class the sentence belongs to where each sentence is classified into class M_{x_j} or class $\bar{M}_{x_j}$ by looking at the different lexicon combinations of pronouns, verbs and adjectives. We searched for particular keywords related to 12 considered medical conditions, and also for possessive words such as {'I', 'have', 'got', 'mine'} to retrieve those posts which disclose information about medical conditions of the individuals. Once the candidate sentences have been classified, the final task is to estimate $P(\mathbf{Z}_{\mathbf{x}_{ij}}{}^{+} = 1)$ as:

Table 6: **Disclosure probabilities for demographic attributes.**

Demographic Attribute	**Disclosure Rate** $P(\mathbf{Z}_{\mathbf{x}_{ij}} = 1)$	Reference Paper
Age	21.6%	[6]
Gender	76.29%	[29]
Location	19.3%	[22]
Race	68.1%	[5]

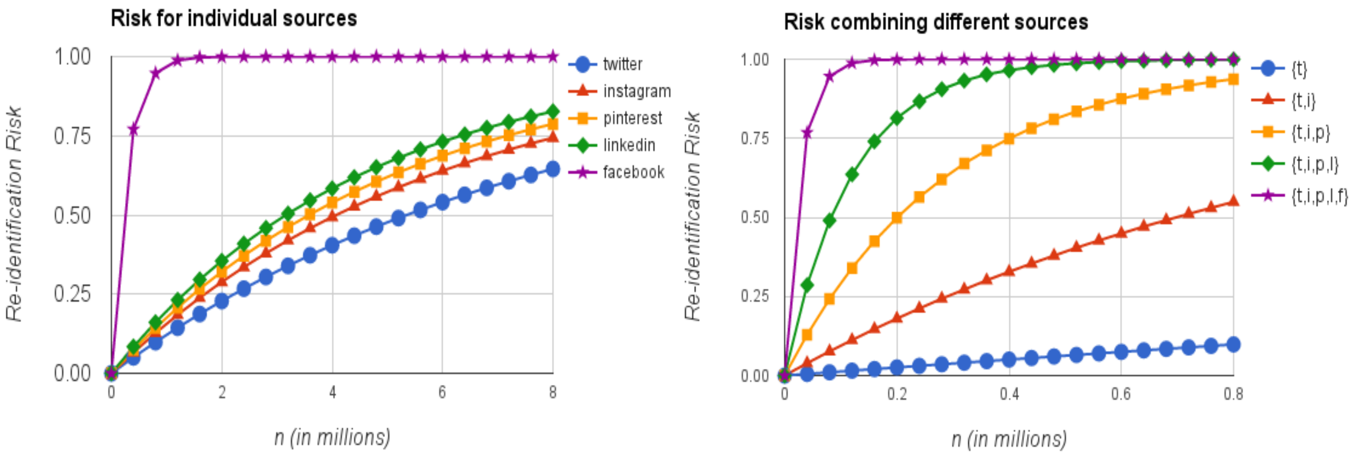
$$P(\mathbf{Z}_{\mathbf{x}_{ij}}{}^{+} = 1) = \frac{|M_{x_j}|}{|S_{x_j}|} . \tag{8}$$

Finally, we assume $P(\mathbf{Q}_{\mathrm{C}} = 1|\mathbf{U}_{\mathrm{C}} = 1) = 1$. This assumption means that, the uniqueness behavior of C is same across both D and E. Indeed, our intuition also suggests that there should be a strong correlation between U_C and Q_C. In fact, if we assume the extreme case where both D and E contains all the individuals in the universe with all information disclosed, then D and E would be identical and uniqueness in D would indeed imply uniqueness in E, i.e., $P(\mathbf{Q}_{\mathrm{C}} = 1|\mathbf{U}_{\mathrm{C}} = 1) = 1$. Thus, we argue that, this is a reasonable assumption given that de-identified data-set D is large enough to be a representative sample of the overall population. In our case, the HCUP data set contains 8 million inpatient records which we believe to be large enough to validate this assumption. However, due to this assumption, all the results reported in section 6.4 are the upper-bound estimates of the risk metric. The readers should note that this assumption does not undermine the potential of the proposed model in any way and finding reliable estimates for $P(\mathbf{Q}_{\mathrm{C}} = 1|\mathbf{U}_{\mathrm{C}} = 1)$ can be one of the interesting future directions.

6.4 Experiment Results

Ideally, we want to quantitatively evaluate the accuracy of the estimated risk of re-identification using the proposed model, but this would require ground truth about the *true* probability of re-identification. It is unclear how we can possibly create this, thus we leave this challenge as a future work, but instead would rely more on qualitative evaluation to show the usefulness of the model and discuss how the proposed model can be used to perform various interesting analyses of re-identification risk for the data sets we experimented with.

We first look at the estimated privacy risk associated with the (hypothetical) release of a de-identified version of the HCUP data set. For this we vary n, the number of individuals in the data set, from 0 to $8,000,000$ and calculate the corresponding re-identification risk, i.e., r_D^*. Figure 2a shows this risk for different social media platforms. We assume that all considered social media platforms (Table 5) exhibit the same disclosure behavior (Table 6 and 7), but have



(a) Re-identification Risk for individual external sources

(b) Re-identification Risk for combined multiple external sources

Figure 2: Re-identification Risk for different external sources. The x-axis represents total number of participants (n) and y-axis plots the corresponding risk. As legend, t, i, p, l, and f denote Twitter, Instagram, Pinterest, LinkedIn, and Facebook, respectively.

different usages (Table 5). Hence, we use the *External Knowledge Model* with parameters learned on the Twitter data set for all cases.

Figure 2a shows that as the number of individuals n increases, so does the risk of re-identifying at least one of them, indicating that the model behaves reasonably. In addition, we see that the re-identification risk is higher for more popular social media platforms. For example, when the data set is composed of 4 million patients' records, the re-identification risks when using Facebook as external data source is almost maximal, i.e., 0.9999, whereas it is only 0.35 when using Twitter. This is to be understood relative to Facebook's usage which is 58% compare to Twitter's usage which is only 19% (Table 5).

We then look at the re-identification risk when an adversary combines information from multiple social media platforms. The results are shown in Figure 2b, where t, i, p, l and f stand for Twitter, Instagram, Pinterest, LinkedIn and Facebook, respectively. We see that combining two different sources poses a greater risk of re-identification, which again shows that the model behaves as we intuitively would expect. Adding additional external knowledge sources increases the risk further. The latter is maximum when the adversary uses all five external knowledge sources as background information for a re-identification attack. It is also interesting to see that although Twitter, Pinterest, Instagram, and LinkedIn are each substantially less risky than Facebook, when they are combined, their risk would be fairly close to the risk of Facebook especially when the number of individuals considered is large.

To understand further the results, we look at specific attribute combinations and their roles in the overall risk. First, we focus on the single attribute disclosure probabilities (of both demographic and medical-condition attributes) computed using equation 5. Table 7 provides the list of attributes sorted by the descending order of *Disclosure-Likelihood*. These probabilities and all following results reported in this paper are computed for the aggregated external knowledge sources case, i.e., all five social media platforms were considered. The demographic attributes were found to be more likely to be disclosed compared to the medical condition attributes. Among the medical-condition attributes, 'meningitis' was found to be most commonly disclosed disease in the social media, whereas 'anemia' was found to be the least common.

Table 7: Attribute list sorted w.r.t. *Disclosure-Likelihood.*

Attribute	*Disclosure-Likelihood*	Attribute	*Disclosure-Likelihood*
gender	1.27633e-01	migraine	1.11424e-06
race	1.01003e-01	acne	1.08148e-06
age	1.83503e-02	diabetes	7.32283e-07
location	1.60194e-02	insomnia	5.64419e-07
meningitis	3.95810e-05	poisoning	5.31675e-07
asthma	1.75838e-05	fever	3.73356e-07
ulcer	1.72729e-06	arthritis	1.69598e-07
ache	1.36146e-06	anemia	6.02684e-08

We now turn our focus to combination of attributes with high potential for re-identification. Figure 3 provides some intuition for this. The x-axis of Figure 3 represents the cardinality of the subset of attributes considered, and the y-axis shows the corresponding average *D-Uniqueness* of the subsets with that length (i.e., the blue star-markers line). In addition, the figure also shows the average *Disclosure-Likelihood* (i.e., the green +-markers line), and the average probability of Re-identification (i.e., the red triangle-markers line). To understand the figure, recall that the probability of re-identification is basically the multiplication of *Disclosure-Likelihood* and the *D-Uniqueness*.

Figure 3 has a straightforward interpretation: as the cardinality of a particular subset of attributes increase, its D-Uniqueness, i.e., the probability to uniquely identify some individual from the data set will increase. However, at the same time, as the cardinality of the subset of attributes increases, the probability that that subset will be disclosed decreases. Thus, there is some optimal cardinality that has both high *D-Uniqueness* and high *Disclosure-Likelihood*, which in the case of our case-study lies in the range [3, 4] as shown in Figure 3 .

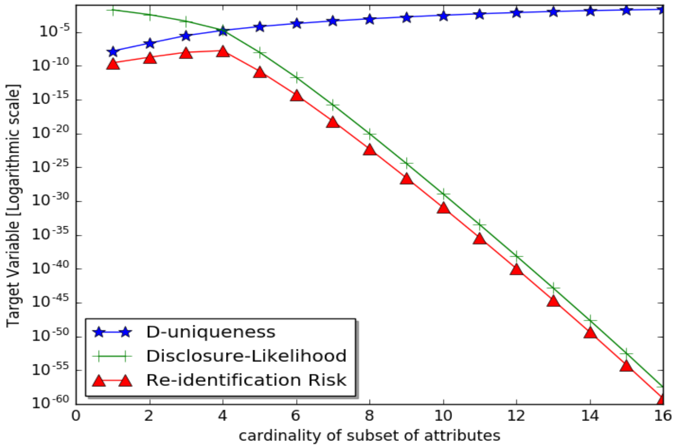


Figure 3: Relationship between cardinality of a subset of attributes and its *D-Uniqueness*, *Disclosure-Likelihood*, and Re-identification Risk

To verify that the range identified from Figure 3, i.e., cardinalities [3, 4] is indeed significant, we focus on combination of attributes (both demographic and medical-condition attributes) with the highest probabilities of re-identification. The top 20 such combinations are shown in Table 8. It can be seen that most combinations have cardinality in the range [3, 4] and many of them contain mostly demographic information. The combination found with the highest probability of re-identification was the combination of all four demographic attributes, i.e., age, location, race and gender. However, some medical condition attributes like meningitis, asthma, diabetes, and migraine were found to be quite helpful in conjunction with the demographic attributes for re-identification.

7 CONCLUSION AND FUTURE WORK

Analysis and mitigation of the re-identification risk from aggregation of multiple data sets are essential to enable privacy-preserving big data applications. To this end, we proposed a general probabilistic risk-model to quantify the risk of re-identification with consideration of external public knowledge resources. We introduced three concepts that are essential for probabilistically quantifying the risk, i.e., *D-Uniqueness*, *E-Uniqueness*, and *Disclosure Likelihood* and used them to define a probabilistic measure of risk of re-identification. The framework enables assessment of the re-identification risk before releasing a data set and detailed analysis of the impact of different kinds of public knowledge source on the potential risk, thus potentially helping mitigate the risk due to releasing a data set.

Table 8: Top 20 combinations of demographic and medical condition attributes with high re-identification probabilities.

Combination	probability of re-identification
[‘age’, ‘location’, ‘race’, ‘gender’]	3.08602e-05
[‘age’, ‘location’, ‘race’]	4.16520e-06
[‘age’, ‘location’, ‘gender’]	6.71205e-07
[‘age’, ‘race’, ‘gender’]	4.34369e-07
[‘age’, ‘race’]	8.93008e-08
[‘age’, ‘location’]	8.87072e-08
[‘location’, ‘race’, ‘gender’]	7.58389e-08
[‘age’, ‘location’, ‘meningitis’, ‘race’, ‘gender’]	3.98472e-08
[‘age’, ‘gender’]	2.96960e-08
[‘age’, ‘asthma’, ‘location’, ‘race’, ‘gender’]	2.48032e-08
[‘age’, ‘location’, ‘meningitis’, ‘race’]	6.66843e-09
[‘location’, ‘race’]	4.10303e-09
[‘age’, ‘asthma’, ‘location’, ‘race’]	3.83848e-09
[‘age’, ‘location’, ‘meningitis’, ‘gender’]	3.82483e-09
[‘age’, ‘location’, ‘race’, ‘gender’, ‘ulcer’]	2.31183e-09
[‘age’, ‘meningitis’, ‘race’, ‘gender’]	1.77845e-09
[‘ache’, ‘age’, ‘location’, ‘race’, ‘gender’]	1.77400e-09
[‘age’, ‘location’, ‘migraine’, ‘race’, ‘gender’]	1.19796e-09
[‘age’, ‘diabetes’, ‘location’, ‘race’, ‘gender’]	1.09382e-09
[‘age’, ‘asthma’, ‘location’, ‘gender’]	1.02967e-09

Using multiple independence assumptions, we also derived a specific probabilistic risk model (i.e., the Naive Re-identification Framework (NRF)) that can be estimated based on any relational data set and relational external knowledge sources for risk assessment and analysis. We further experimented with this model using a health data set (HCUP data) and multiple social media (Facebook, Twitter, Instagram, LinkedIn, and Pinterest) to demonstrate that the proposed risk-model can be used to quantify the risk of releasing HCUP in association of each social media and perform interesting analysis of factors affecting this risk. The model enabled us to make the following interesting observations: 1) The popularity of the social media platform used as the external knowledge source plays a significant role in a successful re-identification attack and all social media increase the risk of re-identification with Facebook posing the highest risk, followed by LinkedIn, Pinterest, Instagram, and Twitter. In fact, for the medical data set we considered (i.e., the HCUP data set), it is almost certain that at least one individual will be re-identified if an adversary uses only Facebook as the external knowledge source. Thus, the results indicate that such data sets require more sophisticated sanitization techniques before they can be publicly released for research purposes. 2) Among the many attribute values social media users voluntarily “leaked” on social media, the demographic attributes (especially age, location, race and gender) pose the greatest risk of re-identification, especially when the medical conditions (e.g., meningitis, asthma, and ulcer) are also disclosed in the public forums. Overall, these results show the great promise of the proposed general framework as well as the specific probabilistic risk model, and their immediate practical application in helping assess re-identification risks before releasing a data set and analyzing security aspects of social media.

Ensuring privacy protection of a released data set when an adversary has access to vast amounts of public information is a very difficult, yet also very important challenge. The proposed probabilistic framework addressed the practical need for assessing the potential risk of re-identification before releasing a data set with a theoretical approach, resulting in a useful framework and specific probabilistic models that can be applied to perform risk analysis of any relational data set and help obtain insights about how to decrease this risk. A major limitation of our work is that we have made some independence assumptions in order to address the issue of data sparseness, which has inevitably affected the accuracy of some of the estimated probabilities. Although some findings (e.g., relative comparison of different social media) are unlikely affected much by those assumptions (since those assumptions are orthogonal to the comparison we have made), the inaccuracy of the estimated probabilities clearly affects the reliability of the estimated risk. Thus in the future, we must study how we can relax these independence assumptions to obtain more accurate estimate of parameters. Another limitation of our work is the empirical evaluation. How to appropriately evaluate the proposed model is by itself a difficult challenge for at least two reasons: 1) Since our target is to quantify the risk with a probability, it is unclear how we can possibly create any gold standard for quantitative evaluation (which would require us to know the *true* probability of risk). 2) The confidentiality required by the terms and conditions prescribed in the HCUP Data-User agreement would not allow us to make any attempt to re-identify any individual in the data set. How to address those challenges is also a very important direction for future work. Another interesting future direction is to use the proposed model to develop a software tool to enable a data publisher to interactively analyze alternative configurations of data fields to be released and seek a configuration that would minimize the risk of re-identification. As we currently have no tool of this kind, such a tool can be expected to be useful even though the estimated risks may not be entirely accurate.

8 ACKNOWLEDGEMENT

This work was supported in part by NSF CNS 13-30491 and NSF CNS 14-08944. The views expressed are those of the authors only. We thank Yunhui Long for her valuable feedbacks during the preparation of the manuscript. We also thank the anonymous reviewers for their useful comments, which helped improving the quality of the paper significantly.

REFERENCES

[1] Michael Barbaro, Tom Zeller, and Saul Hansell. 2006. A face is exposed for AOL searcher no. 4417749. *New York Times* 9, 2008 (2006), 8For.

[2] Daniel C Barth-Jones. 2012. The're-identification'of Governor William Weld's medical information: a critical re-examination of health data identification risks and privacy protections, then and now. *Then and Now (June 4, 2012)* (2012).

[3] Joanna Biega, Ida Mele, and Gerhard Weikum. 2014. Probabilistic prediction of privacy risks in user search histories. In *Proceedings of the First International Workshop on Privacy and Secuirty of Big Data*. ACM, 29–36.

[4] Bee-Chung Chen, Kristen LeFevre, and Raghu Ramakrishnan. 2007. Privacy skyline: Privacy with multidimensional adversarial knowledge. In *Proceedings of the 33rd international conference on Very large data bases*. VLDB Endowment, 770–781.

[5] Aron Culotta, Nirmal Ravi Kumar, and Jennifer Cutler. 2015. Predicting the Demographics of Twitter Users from Website Traffic Data.. In *AAAI*. 72–78.

[6] Ratan Dey, Cong Tang, Keith Ross, and Nitesh Saxena. 2012. Estimating age privacy leakage in online social networks. In *IEEE INFOCOM*. IEEE, 2836–2840.

[7] Maeve Duggan, Nicole B Ellison, Cliff Lampe, Amanda Lenhart, and Mary Madden. 2015. Social media update 2014. *Pew Research Center* 19 (2015).

[8] Cynthia Dwork. 2008. Differential privacy: A survey of results. In *Theory and applications of models of computation*. Springer, 1–19.

[9] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*. Springer, 265–284.

[10] Khaled El Emam. 2013. *Guide to the de-identification of personal health information*. CRC Press.

[11] Khaled El Emam and Luk Arbuckle. 2013. *Anonymizing health data: case studies and methods to get you started*. " O'Reilly Media, Inc.".

[12] Khaled El Emam, Elizabeth Jonker, Luk Arbuckle, and Bradley Malin. 2011. A systematic review of re-identification attacks on health data. *PloS one* 6, 12 (2011), e28071.

[13] Agency for Healthcare Research and Quality. 2015. Nationwide Inpatient Sample (NIS), Healthcare Cost and Utilization Project (HCUP). www.hcup-us.ahrq.gov/nisoverview.jsp. Accessed: 2015-12-30.

[14] Agency for Healthcare Research and Quality. 2015. Overview of the National (Nationwide) Inpatient Sample (NIS). https://www.hcup-us.ahrq.gov/nisoverview.jsp. Accessed: 2015-12-28.

[15] Jaewoo Lee and Chris Clifton. 2012. Differential identifiability. In *ACM SIGKDD*. ACM, 1041–1049.

[16] David D Lewis. 1998. Naive (Bayes) at forty: The independence assumption in information retrieval. In *European conference on machine learning*. Springer, 4–15.

[17] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. 2007. t-closeness: Privacy beyond k-anonymity and l-diversity. In *ICDE*. IEEE, 106–115.

[18] Bernard Lo. 2015. Sharing clinical trial data: maximizing benefits, minimizing risk. *Jama* 313, 8 (2015), 793–794.

[19] Grigorios Loukides, Joshua C Denny, and Bradley Malin. 2010. The disclosure of diagnosis codes can breach research participants' privacy. *Journal of the American Medical Informatics Association* 17, 3 (2010), 322–327.

[20] Ashwin Machanavajjhala, Johannes Gehrke, and Michaela Götz. 2009. Data publishing against realistic adversaries. *Proceedings of the VLDB Endowment* 2, 1 (2009), 790–801.

[21] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. l-diversity: Privacy beyond k-anonymity. *ACM TKDD* 1, 1 (2007), 3.

[22] Alan Mislove, Bimal Viswanath, Krishna P. Gummadi, and Peter Druschel. 2010. You Are Who You Know: Inferring User Profiles in Online Social Networks. In *WSDM*. ACM.

[23] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*. IEEE, 111–125.

[24] US Department of Health, Human Services, et al. 2014. Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.

[25] John R Pleis, Jacqueline W Lucas, and Brian W Ward. 2009. Summary health statistics for US adults: National Health Interview Survey, 2008. *Vital and health statistics. Series 10, Data from the National Health Survey* 242 (2009), 1–157.

[26] Yilin Shen and Hongxia Jin. 2014. Controllable information sharing for user accounts linkage across multiple online social networks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, 381–390.

[27] Latanya Sweeney. 1997. Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics* 25, 2-3 (1997), 98–110.

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

[29] Kurt Thomas, Chris Grier, and David M Nicol. 2010. unfriendly: Multi-party privacy risks in social networks. In *Privacy Enhancing Technologies*. Springer, 236–252.

[30] Deva M Wells, Keren Lehavot, and Margaret L Isaac. 2015. Sounding off on social media: the ethics of patient storytelling in the modern era. *Academic Medicine* 90, 8 (2015), 1015–1019.

[31] Reza Zafarani and Huan Liu. 2013. Connecting users across social media sites: a behavioral-modeling approach. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 41–49.