

What is Privacy?

A standard dictionary definition of privacy as it pertains to data is “freedom from unauthorized intrusion”[58]. With respect to privacy-preserving data mining, this does provide some insight. If users have given authorization to use the data for the particular data mining task, then there is no privacy issue. However, the second part is more difficult: If use is not authorized, what use constitutes “intrusion”?

A common standard among most privacy laws (e.g., European Community privacy guidelines[26] or the U.S. healthcare laws[40]) is that privacy only applies to “individually identifiable data”. Combining *intrusion* and *individually identifiable* leads to a standard to judge privacy-preserving data mining: A privacy-preserving data mining technique must ensure that any information disclosed

1. cannot be traced to an individual; or
2. does not constitute an intrusion.

Formal definitions for both these items are an open challenge. At one extreme, we could assume that any data that does not give us completely accurate knowledge about a specific individual meets these criteria. At the other extreme, any improvement in our knowledge about an individual could be considered an intrusion. The latter is particularly likely to cause a problem for data mining, as the goal is to improve our knowledge. Even though the target is often groups of individuals, knowing more about a group does increase our knowledge about individuals in the group. This means we need to *measure* both the knowledge gained and our ability to relate it to a particular individual, and determine if these exceed thresholds.

This chapter first reviews metrics concerned with individual identifiability. This is not a complete review, but concentrates on work that has particular applicability to privacy-preserving data mining techniques. The second issue, what constitutes an intrusion, is less clearly defined. The end of the chapter will discuss some proposals for metrics to evaluate intrusiveness, but this is still very much an open problem.

To utilize this chapter in the concept of privacy-preserving data mining, it is important to remember that all disclosure from the data mining must be considered. This includes disclosure of data sets that have been altered/randomized to provide privacy, communications between parties participating in the mining process, and disclosure of the results of mining (e.g., a data mining model.) As this chapter introduces means of measuring privacy, examples will be provided of their relevance to the types of disclosures associated with privacy-preserving data mining.

2.1 Individual Identifiability

The U.S. Healthcare Information Portability and Accountability Act (HIPAA) defines *individually nonidentifiable data* as data “that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual”[41]. The regulation requires an analysis that the risk of identification of individuals is very small in any data disclosed, *alone or in combination with other reasonably available information*. A real example of this is given in [79]: Medical data was disclosed with name and address removed. Linking with publicly available voter registration records using birth date, gender, and postal code revealed the name and address corresponding to the (presumed anonymous) medical records. This raises a key point: Just because the individual is not identifiable in the data is not sufficient; joining the data with other sources must not enable identification.

One proposed approach to prevent this is k -anonymity[76, 79]. The basic idea behind k -anonymity is to group individuals so that any identification is only to a group of k , not to an individual. This requires the introduction of a notion of *quasi-identifier*: information that can be used to link a record to an individual. With respect to the HIPAA definition, a quasi-identifier would be anything that would be present in “reasonably available information”. The HIPAA regulations actually give a list of presumed quasi-identifiers; if these items are removed, data is considered not individually identifiable. The definition of k -anonymity states that any record must not be unique in its quasi-identifiers; there must be at least k records with the same quasi-identifier. This ensures that an attempt to identify an individual will result in at least k records that could apply to the individual. Assuming that the privacy-sensitive data (e.g., medical diagnoses) are not the same for all k records, then this throws uncertainty into any knowledge about an individual. The uncertainty lowers the risk that the knowledge constitutes an intrusion.

The idea that knowledge that applies to a group rather than a specific individual does not violate privacy has a long history. Census bureaus have used this approach as a means of protecting privacy. These agencies typically publish aggregate data in the form of contingency tables reflecting the count of individuals meeting a particular criterion (see Table 2.1). Note that some cells

Table 2.1. Excerpt from Table of Census Data, U.S. Census Bureau

		Block Group 1, Census Tract 1, District of Columbia, District of Columbia
Total:		9
Owner occupied:		3
1-person household		2
2-person household		1
	...	
Renter occupied:		6
1-person household		3
2-person household		2
	...	

list only a single such household. The disclosure problem is that combining this data with small cells in other tables (e.g., a table that reports salary by size of household, and a table reporting salary by racial characteristics) may reveal that only one possible salary is consistent with the numbers in all of the tables. For example, if we know that all owner-occupied 2-person households have salary over \$40,000, and of the nine multiracial households, only one has salary over \$40,000, we can determine that the single multiracial individual in an owner-occupied 2-person household makes over \$40,000. Since race and household size can often be observed, and home ownership status is publicly available (in the U.S.), this would result in disclosure of an individual salary.

Several methods are used to combat this. One is by introducing noise into the data; in Table 2.1 the Census Bureau warns that statistical procedures have been applied that introduce some uncertainty into data for small geographic areas with small population groups. Other techniques include cell suppression, in which counts smaller than a threshold are not reported at all; and generalization, where cells with small counts are merged (e.g., changing Table 2.1 so that it doesn't distinguish between owner-occupied and Renter-occupied housing.) Generalization and suppression are also used to achieve k -anonymity.

How does this apply to privacy-preserving data mining? If we can ensure that disclosures from the data mining generalize to large enough groups of individuals, then the size of the group can be used as a metric for privacy protection. This is of particular interest with respect to data mining results: When does the result itself violate privacy? The “size of group” standard may be easily met for some techniques; e.g., pruning approaches for decision trees may already generalize outcomes that apply to only small groups and association rule support counts provide a clear group size.

An unsolved problem for privacy-preserving data mining is the cumulative effect of multiple disclosures. While building a single model may meet the standard, multiple data mining models in combination may enable deducing individual information. This is closely related to the “multiple table” problem

of census release, or the *statistical disclosure limitation* problem. Statistical disclosure limitation has been a topic of considerable study; readers interested in addressing the problem for data mining are urged to delve further into statistical disclosure limitation[18, 88, 86].

In addition to the “size of group” standard, the census community has developed techniques to measure risk of identifying an individual in a dataset. This has been used to evaluate the release of Public Use Microdata Sets: Data that appears to be actual census records for sets of individuals. Before release, several techniques are applied to the data: Generalization (e.g., limiting geographic detail), top/bottom coding (e.g., reporting a salary only as “greater than \$100,000”), and data swapping (taking two records and swapping their values for one attribute.) These techniques introduce uncertainty into the data, thus limiting the confidence in attempts to identify an individual in the data. Combined with releasing only a sample of the dataset, it is likely that an identified individual is really a false match. This can happen if the individual is not in the sample, but swapping values between individuals in the sample creates a quasi-identifier that matches the target individual. Knowing that this is likely, an adversary trying to compromise privacy can have little confidence that the matching data really applies to the targeted individual.

A set of metrics are used to evaluate privacy preservation for public use microdata sets. One set is based on the value of the data, and includes preservation of univariate and covariate statistics on the data. The second deals with privacy, and is based on the percentage of individuals that a particularly well-equipped adversary could identify. Assumptions are that the adversary:

1. knows that some individuals are almost certainly in the sample (e.g., 600-1000 for a sample of 1500 individuals),
2. knows that the sample comes from a restricted set of individuals (e.g., 20,000),
3. has a good estimate (although some uncertainty) about the non-sensitive values (quasi-identifiers) for the target individuals, and
4. has a reasonable estimate of the sensitive values (e.g., within 10%.)

The metric is based on the number of individuals the adversary is able to correctly and confidently identify. In [60], identification rates of 13% are considered acceptably low. Note that this is an extremely well-informed adversary; in practice rates would be much lower.

While not a clean and simple metric like “size of group”, this experimental approach that looks at the rate at which a well-informed adversary can identify individuals can be used to develop techniques to evaluate a variety of privacy-preserving data mining approaches. However, it is not amenable to a simple, “one size fits all” standard – as demonstrated in [60], applying this approach demands considerable understanding of the particular domain and the privacy risks associated with that domain.

There have been attempts to develop more formal definitions of anonymity that provide greater flexibility than k -anonymity. A metric presented in [15]

uses the concept of anonymity, but specifically based on the ability to learn to distinguish individuals. The idea is that we should be unable to learn a classifier that distinguishes between individuals with high probability. The specific metric proposed was:

Definition 2.1. [15] *Two records that belong to different individuals I_1, I_2 are p -indistinguishable given data X if for every polynomial-time function $f : I \mapsto \{0, 1\}$*

$$|Pr\{f(I_1) = 1|X\} - Pr\{f(I_2) = 1|X\}| \leq p$$

where $0 < p < 1$.

Note the similarity to k -anonymity. This definition does not prevent us from learning sensitive information, it only poses a problem if that sensitive information is tied more closely to one individual rather than another. The difference is that this is a metric for the (sensitive) data X rather than the quasi-identifiers.

Further treatment along the same lines is given in [12], which defines a concept of isolation based on the ability of an adversary to “single out” an individual y in a set of points RDB using a query q :

Definition 2.2. [12] *Let y be any RDB point, and let $\delta_y = ||q - y||_2$. We say that q (c, t) -isolates y iff $B(q, c\delta_y)$ contains fewer than t points in the RDB, that is, $|B(q, c\delta_y) \cap RDB| < t$.*

The idea is that if y has at least t close neighbors, then anonymity (and privacy) is preserved. “Close” is determined by both a privacy threshold c , and how close the adversary’s “guess” q is to the actual point y . With $c = 0$, or if the adversary knows the location of y , then k -anonymity is required to meet this standard. However, if an adversary has less information about y , the “anonymizing” neighbors need not be as close.

The paper continues with several sanitization algorithms that guarantee meeting the (c, t) -isolation standard. Perhaps most relevant to our discussion is that they show how to relate the definition to different “strength” adversaries. In particular, an adversary that generates a region that it believes y lies in versus an adversary that generates an action point q as the estimate. They show that there is essentially no difference in the ability of these adversaries to violate the (non)-isolation standard.

2.2 Measuring the Intrusiveness of Disclosure

To violate privacy, disclosed information must both be linked to an individual, and constitute an intrusion. While it is possible to develop broad definitions for individually identifiable, it is much harder to state what constitutes an intrusion. Release of some types of data, such as date of birth, pose only a minor annoyance by themselves. But in conjunction with other information date

of birth can be used for identity theft, an unquestionable intrusion. Determining intrusiveness must be evaluated independently for each domain, making general approaches difficult.

What can be done is to measure the amount of information about a privacy sensitive attribute that is revealed to an adversary. As this is still an evolving area, we give only a brief description of several proposals rather than an in-depth treatment. It is our feeling that measuring intrusiveness of disclosure is still an open problem for privacy-preserving data mining; readers interested in addressing this problem are urged to consult the papers referenced in the following overview.

Bounded Knowledge.

Introducing uncertainty is a well established approach to protecting privacy. This leads to a metric based on the ability of an adversary to use the disclosed data to estimate a sensitive value. One such measure is given by [1]. They propose a measure based on the *differential entropy* of a random variable. The differential entropy $h(A)$ is a measure of the uncertainty inherent in A . Their metric for privacy is $2^{h(A)}$. Specifically, if we add noise from a random variable A , the privacy is:

$$\Pi(A) = 2^{-\int_{\Omega_A} f_A(a) \log_2 f_A(a) da}$$

where Ω_A is the domain of A . There is a nice intuition behind this measure: The privacy is 0 if the exact value is known, and if the adversary knows only that the data is in a range of width a (but has no information on where in that range), $\Pi(A) = a$.

The problem with this metric is that an adversary may already have knowledge of the sensitive value; the real concern is how much that knowledge is increased by the data mining. This leads to a conditional privacy definition:

$$\Pi(A|B) = 2^{-\int_{\Omega_{A,B}} f_{A,B}(a,b) \log_2 f_{A|B=b}(a) da db}$$

This was applied to noise addition to a dataset in [1]; this is discussed further in Chapter 4.2. However, the same metric can be applied to disclosures other than of the source data (although calculating the metric may be a challenge.)

A similar approach is taken in [14], where conditional entropy was used to evaluate disclosure from secure distributed protocols (see Chapter 3.3). While the definitions in Chapter 3.3 require perfect secrecy, the approach in [14] allows some disclosure. Assuming a uniform distribution of data, they are able to calculate the conditional entropy resulting from execution of a protocol (in particular, a set of linear equations that combine random noise and real data.) Using this, they analyze several scalar product protocols based on adding noise to a system of linear equations, then later factoring out the noise. The protocols result in sharing the “noisy” data; the technique of [14]

enables evaluating the expected change in entropy resulting from the shared noisy data. While perhaps not directly applicable to all privacy-preserving data mining, the technique shows another way of calculating the information gained.

Need to know.

While not really a metric, the reason for disclosing information is important. Privacy laws generally include disclosure for certain permitted purposes, e.g. the European Union privacy guidelines specifically allow disclosure for government use or to carry out a transaction requested by the individual[26]:

Member States shall provide that personal data may be processed only if:

- (a) the data subject has unambiguously given his consent; or
- (b) processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract; or ...

This principle can be applied to data mining as well: disclose only the data actually needed to perform the desired task. We will show an example of this in Chapter 4.3. One approach produces a classifier, with the classification model being the outcome. Another provides the ability to classify, without actually revealing the model. If the goal is to classify new instances, the latter approach is less of a privacy threat. However, if the goal is to gain knowledge from understanding the model (e.g., understanding decision rules), then disclosure of that model may be acceptable.

Protected from disclosure.

Sometimes disclosure of certain data is specifically proscribed. We may find that *any* knowledge about that data is deemed too sensitive to reveal. For specific types of data mining, it may be possible to design techniques that limit ability to infer values from results, or even to control what results can be obtained. This is discussed further in Chapter 6.3. The problem in general is difficult. Data mining results inherently give knowledge. Combined with other knowledge available to an adversary, this may give *some* information about the protected data. A more detailed analysis of this type of disclosure will be discussed below.

Indirect disclosure.

Techniques to analyze a classifier to determine if it discloses sensitive data were explored in [48]. Their work made the assumption that the disclosure was a “black box” classifier – the adversary could classify instances, but not look inside the classifier. (Chapter 4.5 shows one way to do this.) A key insight

of this work was to divide data into three classes: Sensitive data, Public data, and data that is Unknown to the adversary. The basic metric used was the Bayes classification error rate. Assume we have data (x_1, x_2, \dots, x_n) , that we want to classify x_i 's into m classes $\{0, 1, \dots, m-1\}$. For any classifier C :

$$x_i \mapsto C(x_i) \in \{0, 1, \dots, m-1\}, \quad i = 1, 2, \dots, n,$$

we define the classifier accuracy for C as:

$$\sum_{i=0}^{m-1} Pr\{C(x) \neq i | z = i\} Pr\{z = i\}.$$

As an example, assume we have n samples $X = (x_1, x_2, \dots, x_n)$ from a 2-point Gaussian mixture $(1 - \epsilon)N(0, 1) + \epsilon N(\mu, 1)$. We generate a sensitive data set $Z = (z_1, z_2, \dots, z_n)$ where $z_i = 0$ if x_i is sampled from $N(0, 1)$, and $z_i = 1$ if x_i is sampled from $N(\mu, 1)$. For this simple classification problem, notice that out of the n samples, there are roughly ϵn samples from $N(\mu, 1)$, and $(1 - \epsilon)n$ from $N(0, 1)$. The total number of misclassified samples can be approximated by:

$$n(1 - \epsilon)Pr\{C(x) = 1 | z = 0\} + n\epsilon Pr\{C(x) = 0 | z = 1\};$$

dividing by n , we get the fraction of misclassified samples:

$$(1 - \epsilon)Pr\{C(x) = 1 | z = 0\} + \epsilon Pr\{C(x) = 0 | z = 1\};$$

and the metric gives the overall possibility that any sample is misclassified by C . Notice that this metric is an “overall” measure, not a measure for a particular value of x .

Based on this, several problems are analyzed in [48]. The obvious case is the example above: The classifier returns sensitive data. However, there are several more interesting cases. What if the classifier takes both public and unknown data as input? If we assume that all of the training data is known to the adversary (including public and sensitive, but not unknown, values), the classifier $C(P, U) \rightarrow S$ gives the adversary no additional knowledge about the sensitive values. But if the training data is unknown to the adversary, the classifier C does reveal sensitive data, *even though the adversary does not have complete information as input to the classifier*.

Another issue is the potential for privacy violation of a classifier that takes public data and discloses non-sensitive data to the adversary. While not in itself a privacy violation (no sensitive data is revealed), such a classifier could enable the adversary to deduce sensitive information. An experimental approach to evaluate this possibility is given in [48].

A final issue is raised by the fact that publicly available records already contain considerable information that many would consider private. If the private data revealed by a data mining process is already publicly available, does this pose a privacy risk? If the ease of access to that data is increased

(e.g., available on the internet versus in person at a city hall), then the answer is yes. But if the data disclosed through data mining is as hard to obtain as the publicly available records, it isn't clear that the data mining poses a privacy threat.

Expanding on this argument, privacy risk really needs to be measured as the loss of privacy resulting from data mining. Suppose X is a sensitive attribute and its value for an fixed individual is equal to x . For example, $X = x$ is the salary of a professor at a university. Before any data processing and mining, some prior information may already exist regarding x . If each department publishes a range of salaries for each faculty rank, the prior information would be a bounded interval. Clearly, when addressing the impact of data mining on privacy, prior information also should be considered. Another type of external information comes from other attributes that are not privacy sensitive and are dependent on X . The values of these attributes, or even some properties regarding these attributes, are already public. Because of the dependence, information about X can be inferred from these attributes.

Several of the above techniques can be applied to these situations, in particular Bayesian inference, the conditional privacy definition of [1] (as well as a related conditional distribution definition from [27], and the indirect disclosure work of [48]. Still open is how to incorporate *ease of access* into these definitions.

Privacy Preserving Data Mining

Vaidya, J.; Clifton, C.W.; Zhu, Y.M.

2006, X, 122 p. 20 illus., Hardcover

ISBN: 978-0-387-25886-7