A Risk Management Framework for Health Care Data Anonymization

Tyrone Grandison
IBM Services Research
Murat Kantarcioglu
University of Texas at Dallas

Abstract

To facilitate many important tasks ranging from medical research to personalized medicine, micro datasets that contain sensitive patient information need to be shared. To address this important issue, there has been considerable work done on anonymization techniques that try to protect privacy. At the same time, under various assumptions about the background knowledge available to adversary, it is shown that most of the existing microdata anonymization techniques could be attacked. In this paper, instead we suggest a novel risk management framework that could be used to analyze risk in sharing anonymized data.

1. Introduction

To enable larger scale research, scientists need to share private data collections. To facilitate data sharing for research purposes, organizations in various countries (i.e., Estonia, Iceland, Japan, Mexico, Norway, Sweden, The United Kingdom, and the United States) are establishing data repositories that store person-specific biomedical records [ECO05]. At the same time, person-specific biomedical records must be shared in a manner that preserves the anonymity of the data subjects due to various social concerns [CLA03].To mitigate such concerns, data anonymization techniques have been proposed.

In general, data anonymization tools use various generalization and suppression techniques to remove unique identifiers. In addition, they modify quasi-identifiers[1] to prevent potential identification of individuals. One of the earliest examples of such techniques is *k*-anonymity [SWE02]. In *k*-anonymity, the data is modified in such a way that any given tuple is indistinguishable from any *k-1* other tuples if all the tuples are projected on their quasi-identifier attributes. Although *k*-anonymity prevents identifying individuals, by combining various publicly available data, it can be used to identify certain sensitive values, e.g. learning whether a known person is diagnosed with cancer or not. Due to the weaknesses of *k*-anonymity, various other techniques have been proposed, e.g. *l*-diversity [MAC06], *t*-closeness [Li07], etc. Unfortunately as shown in [DWO06], no anonymization technique could release useful information and still preserve privacy under all possible attack scenarios. For this reason, authors in [DWO06] propose a differential privacy definition to share statistical query results computed over private data. Still for cases where the actual microdata needs to be shared, differential privacy based solutions cannot be used.

Given the current results, it is clear that any time anonymized microdata is shared, there is a potential for attacks that could be used to violate individual privacy. Various previous work, e.g. [SKI98], has tried to estimate the "re-identification risk" when anonymized data is shared. Although this is a good starting point, we believe that this is not enough for a comprehensive risk management framework.

To achieve a comprehensive risk management framework, not only do we need to estimate re-identification risk but we also need to understand the potential losses due to re-identification. Additionally, we need to consider the utility of the data sharing. In the remainder of this paper, we will explain the details of our risk management framework.

2. Risk Management Framework for Health Care Data

To accurately manage risks in sharing anonymized data, we need to consider following factors: 1) The likelihood of re-identification for each individual *i*, l_i, due to previously released versions of the dataset and the publicly available data, 2) The severity of potentially released sensitive information, s_i, if re-identification actually happens. 3) The total expected utility (*u*) of sharing the anonymized data 4) The part of the population, po_i, that could be affected by the re-identification. For example, if only a specific subgroup of the population, e.g. American Indians, is affected by the re-

[1] Any attribute that is not an unique identifier but that could be used to identify a person (e.g. birth day) is considered a quasi-identifier

identification, this could create additional concern for this sub-population. Using the above factors, we can easily define the potential risk in sharing anonymized data as follows:

$$Risk = \bigcup_{i=1}^{m} (s_i, l_i, po_i)$$

$\bigcup$ is a monotonically increasing function that combines the inputs and transforms them into a standardized range, which is the same as the range used to measure the utility of the data. In its most basic form, it can be viewed as multiplying the severity, loss and population size for individual *i* to determine individual *i*'s risk and then adding all the risk numbers for all the individuals in the data set.

Clearly, data should be shared if utility *u* is bigger than the potential risk. Below, we discuss how various parameters could be estimated.

2.1. Estimating the parameters

To estimate re-identification risk for each individual in the anonimized data set, we need to estimate the probability of re-identifying an individual using various publicly available data including the previous releases of the data set. Existing techniques such as the ones given in [SKI98] could be used for estimating such a re-identification risk. Since, this re-identification risk depends on the background information of the adversary, various background information scenarios needs to be considered. For example, we may assume that adversary knows the quasi-identifiers for the entire society, e.g. the adversary knows the birth days and addresses of everyone from the voter registration list. In some cases, this could be too pessimistic. Instead, we can also evaluate the re-identification risk assuming that some pieces of information is not available to attacker, e.g. adversary may not know the individuals medical history. By considering these different scenarios, we can estimate the re-identification risk.

Next, we need to estimate the potential consequences of re-identification. This could be done automatically by considering the medical history of each individual. As a starting point, we suggest using various weights for different diagnosis that is shared as a part of the micro-data . For example, sensitive diagnosis could have more weight compared to non-sensitive ones.

Finally, we need to estimate whether some part of the population will be significantly affected due to potential re-identification. This could be easily estimated by combining the re-identification risk with the available sub-population information (e.g., race information).

2.2. Risk Mitigation

In some cases, the estimated risk may turn out to be more than what is deemed tolerable with respect to estimated utility. In such cases, we may try to mitigate the risk by tuning the data anonymization parameters. For example, in *k*-anonymity, *k* could be increased to reduce the privacy risk. In addition, the tuples that are highly likely to re-identified with potentially sensitive information could be suppressed.

3. Conclusions

In this paper, we propose a comprehensive risk management framework that considers various risk factors in sharing anonymized data. Compared to previous work, our framework also considers the sensitivity of the information that could be potentially leaked and the sub-population that could be affected by such leakage. Our future work involves testing various instances of the risk computation function, evaluating multiple theoretical foundations for parametric estimation, determining the taxonomy of mitigation strategies, developing, executing extensive experiments to test our framework and creating user-friendly toolkits that can be built upon.

References

[CLA03] E. Clayton, "Ethical, legal, and social implications of genomic medicine". New England J. Med., vol. 349, pp. 562–569, 2003.

[DWO06] C. Dwork. "Differential privacy". In 33rd International Colloquium on Automata. Languages and Programming- ICALP 2006, pages 1-12, 2006.

[ECO05] Anonymous, "Medicine's new central bankers". The Economist, Dec. 903 2005.

[MAC06] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. "*l*-diversity: Privacy beyond k-anonymity". ICDE 2006, page 24, 2006.

[Li07] N. Li, T. Li, and S. Venkatasubramanian. "*t*-closeness: Privacy beyond *k*-anonymity and *l*-diversity". ICDE 2007.

[SWE02] L. Sweeney. "Achieving *k*-anonymity privacy protection using generalization and suppression". International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5), 2002.

[SKI98] C.J. Skinner and D.J. Holmes. "Estimating the Re-identification Risk Per Record in Microdata". Journal of Official Statistics, Vol. 14, No. 4, 1998, pp. 361-372.