Journal of Engineering Sciences

Vol 10, Issue 9, Sept /2019
ISSN NO:0377-9254

AN ITERATIVE CLASSIFICATION SCHEME FOR SANITIZING LARGE SCALE DATASETS

[1]A.MANASA, [2]P MURALI PONAGANTI

[1]MCA Student, [2]Associate Professor

DEPARTMENT OF MCA

SREE CHAITANYA COLLEGE OF ENGINEERING, KARIMNAGAR

Abstract

Cheap ubiquitous computing enables the collection of massive amounts of personal data in a wide variety of domains. Many organizations aim to share such data while obscuring features that could disclose personally identifiable information. Much of this data exhibits weak structure (e.g., text), such that machine learning approaches have been developed to detect and remove identifiers from it. While learning is never perfect, and relying on such approaches to sanitize data can leak sensitive information, a small risk is often acceptable. Our goal is to balance the value of published data and the risk of an adversary discovering leaked identifiers. We model data sanitization as a game between 1) a publisher who chooses a set of classifiers to apply to data and publishes only instances predicted as non-sensitive and 2) an attacker who combines machine learning and manual inspection to uncover leaked identifying information. We introduce a fast iterative greedy algorithm for the publisher that ensures a low utility for a resource-limited adversary. Moreover, using five text data sets we illustrate that our algorithm leaves virtually no automatically identifiable sensitive instances for a state-of-the-art learning algorithm, while sharing over 93% of the original data, and completes after at most 5 iterations.

I. INTRODUCTION

Vast quantities of personal data are now collected in a wide variety of domains, including personal health records, emails, court documents, and the Web. It is anticipated that such data can enable significant improvements in the quality of services provided to individuals and facilitate new discoveries for society. At the same time, the data collected is often sensitive, and regulations, such as the Privacy Rule of the Health Insurance Portability and Accountability Act of 1996 (when disclosing medical records), Federal Rules of Civil Procedure (when disclosing court records), and the European Data Protection Directive often recommend the removal of identifying information. To accomplish such goals, the past several decades have brought forth the development of numerous data protection models. These models invoke various principles, such as hiding individuals in a crowd (e.g., k-anonymity) or perturbing values to ensure that little can be inferred about an individual even with arbitrary side information (e.g., ε-differential privacy). All of these approaches are predicated on the assumption that the publisher of the data knows where the identifiers are from the outset. More specifically, they assume the data has an explicit representation, such as a relational form, where the data has at most a small set of values per feature. However, it is increasingly the case that the data we generate lacks a formal relational or explicitly structured representation. A clear example of this phenomenon is the substantial quantity of natural language text which is created in the clinical notes in medical records.

To protect such data, there has been a significant amount of research into natural language processing (NLP) techniques to detect and subsequently redact or substitute identifiers. As demonstrated through systematic reviews and various competitions, the most scalable versions of such techniques are rooted in, or rely heavily upon, machine learning methods, in which the publisher of the data annotates instances of personal identifiers in the text, such as patient and doctor name, Social Security Number, and a date of birth, and the machine attempts to learn a classifier (e.g., a grammar) to predict where such identifiers reside in a much larger corpus. Unfortunately,

generating a perfectly annotated corpus for training purposes can be extremely costly. This, combined with the natural imperfection of even the best classification learning methods implies that some sensitive information will invariably leak through to the data recipient. This is clearly a problem if, for instance, the information leaked corresponds to direct identifiers (e.g., personal name) or quasi-identifiers (e.g., ZIP codes or dates of birth) which may be exploited in re identification attacks, such as the re-identification of Thelma Arnold in the search logs disclosed by AOL or the Social Security Numbers in Jeb Bush's emails. Rather than attempt to detect and redact every sensitive piece of information, our goal is to guarantee that even if identifiers remain in the published data, the adversary cannot easily find them. Fundamental to our approach is the acceptance of non-zero privacy risk, which we view as unavoidable.

This is consistent with most privacy regulation, such as HIPAA, which allows expert determination that privacy "risk is very small", and the EU Data Protection Directive, which "does not require anonymisation to be completely riskfree". Our starting point is a threat model within which an attacker uses published data to first train a classifier to predict sensitive entities based on a labeled subset of the data, prioritizes inspection based on the predicted positives, and inspects and verifies the true sensitivity status of B of these in a prioritized order. Here, B is the budget available to inspect (or read) instances and true sensitive entities are those which have been correctly labeled as sensitive (for example, true sensitive entities could include identifiers such as a name, Social Security Number, and address). We use this threat model to construct a game between a publisher, who 1) applies a collection of classifiers to an original data set, 2) prunes all the positives predicted by any classifier, and 3) publishes the remainder, and an adversary acting according to our threat model. The data publisher's ultimate goal is to release as much data as possible while at the same time redacting sensitive information to the point where reidentification risk is sufficiently low. In support of the second goal, we show that any locally optimal publishing strategy exhibits the following two properties when the loss associated with exploited personal identifiers is high: a) an adversary cannot learn a classifier with a high true positive count, and b) an adversary with a large inspection budget cannot do much better than manually inspecting and confirming instances chosen uniformly at random (i.e., the classifier adds little value).

Moreover, we introduce a greedy publishing strategy which is guaranteed to converge to a local optimum and consequently guarantees the above two properties in a linear (in the size of the data) number of iterations. At a high level, the greedy algorithm iteratively executes learning and redaction. It repeatedly learns the classifier to predict sensitive entities on the remaining data, and then removes the predicted positives, until a local optimum is reached. The intuition behind the iterative redaction process is that, in each iteration, the learner essentially checks to determine if an adversary could obtain utility by uncovering residual identifiers; if so, these instances are redacted, while the process is terminated otherwise. Our experiments on two distinct electronic health records data sets demonstrate the power of our approach, showing that 1) the number of residual true positives is always quite small, addressing he goal of reducing privacy risk, 2) confirming that the attacker with a large budget cannot do much better than uniformly randomly choosing entities to manually inspect, 3) demonstrating that most (> 93%) of the original data is published, thereby supporting the goal of maximizing the quantity of released data, and 4) showing that, in practice, the number of required algorithm iterations (< 5) is a small fraction of the size of the data. Additional experiments, involving three datasets that are unrelated to the health domain corroborate these findings, demonstrating generalizability in our approach

II. SYSTEM ANALYSIS:

Existing System

It is anticipated that such data can enable significant improvements in the quality of services provided to individuals and facilitate new discoveries for society.

These models invoke various principles, such as hiding individuals in a crowd (e.g., k-anonymity) or perturbing values to ensure that little can be inferred about an individual even with arbitrary side information.

All of these approaches are predicated on the assumption that the publisher of the data knows where the identifiers are from the outset. More specifically, they assume the data has an explicit representation, such as a relational form, where the data has at most a small set of values per feature

Proposed System

The simplest of these rely on a large collection of rules, dictionaries, and regular expressions proposed an automated data sanitization algorithm aimed at removing sensitive identifiers while inducing the least distortion to the contents of documents.

We propose a novel explicit threat model for this problem, allowing us to make formal guarantees about the vulnerability of the published data to adversarial re-identification attempts.

We provide additional theoretical analysis of the proposed GreedySanitize algorithm focusing on two questions. First, what kinds of privacy guarantees does this algorithm offer? Second, how can we generalize the privacy guarantees to account for finite sample approximations inherent in the algorithm? To address the first question, we abstract away the details of our algorithm behind the veil of its stopping condition, which turns out to be the primary driver of our results. This also allows us to state the privacy guarantees in much more general terms.

III. IMPLEMENTATION

Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective.

The implementation stage involves careful planning, investigation of the existing system and it's constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods.

Modules

1. Approaches for Anonymizing Structured Data
2. Traditional Methods for Sanitizing Unstructured Data
3. Machine Learning Methods for Sanitizing Unstructured Data

Approaches for Anonymizing Structured Data

There has been a substantial amount of research conducted in the field of privacy-preserving data publishing (PPDP) over the past several decades . Much of this work is dedicated to methods that transform well-structured (e.g., relational) data to adhere to a certain criterion or a set of criteria, such as k-anonymization, l-diversity, m- invariance, and-differential privacy, among a multitude of others. These criteria attempt to offer guarantees about the ability of an attacker to either distinguish between different records in the data or make inferences tied to a specific individual. There is now an extensive literature aiming to operationalize such PPDP criteria in practice through the application of techniques such as generalization, suppression (or removal), and randomization. All of these techniques, however, rely on a priori knowledge of which features in the data are either themselves sensitive or can be linked to sensitive attributes. This is a key distinction from our work: we aim to automatically discover which entities in unstructured data are sensitive, as well as formally ensure that whatever sensitive data remains cannot be easily unearthed by an adversary.

Traditional Methods for Sanitizing Unstructured Data

In the context of privacy preservation for unstructured data, such as text, various approaches have been proposed for the automatic discovery of sensitive entities, such as identifiers. The simplest of these rely on a large collection of rules, dictionaries, and regular expressions proposed an automated data sanitization algorithm aimed at removing sensitive identifiers while inducing the least distortion to the contents of documents. However, this algorithm assumes that sensitive entities, as well as any possible related entities, have already been labeled. Similarly, have developed the t-plausibility algorithm to replace the known (labeled) sensitive identifiers within the documents and guarantee that the sanitized document is associated with least t documents.

Machine Learning Methods for Sanitizing Unstructured Data

A key challenge in unstructured data that makes it qualitatively distinct from structured is that even identifying (labeling) which entities are sensitive is non-trivial.

A natural idea, which has received considerable traction in prior literature, is to use machine learning algorithms, trained on a small portion of labeled data, to automatically identify sensitive entities.

Our approach builds on this literature, but is quite distinct from it in several ways. First, we propose a novel explicit threat model for this problem, allowing us to make formal guarantees about the vulnerability of the published data to adversarial re-identification attempts.

IV. CONCLUSION

Our ability to take full advantage of large amounts of unstructured data collected across a broad array of domains is limited by the sensitive information contained therein. This paper introduced a novel framework for sanitization of such data that relies upon 1) a principled threat model, 2) a very general class of publishing strategies, and 3) a greedy, yet effective, data publishing algorithm. The experimental evaluation shows that our algorithm is: a) substantially better than existing approaches for suppressing sensitive data, and b) retains most of the value of the data, suppressing less than 10% of information on all four data sets we considered in evaluation. In contrast, cost-sensitive variants of standard learning methods yield virtually no residual utility, suppressing most, if not all, of the data, when the loss associated with privacy risk is even moderately high. Since our adversarial model is deliberately extremely strong - far stronger, indeed, than is plausible - our results suggest feasibility for data sanitization at scale.

REFERENCES

[1] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97–107, 2014.

[2] U.S. Dept. of Health and Human Services, "Standards for privacy and individually identifiable health information; final rule," Federal Register, vol. 65, no. 250, pp. 82 462–82 829, 2000.

[3] Committe on the Judiciary House of Representatives, "Federal Rules of Civil Procedure," 2014.

[4] European Parliament and Council of the European Union, "Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data," Official Journal of the EC, vol. 281, pp. 0031–0050, 1995.

[5] B. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy preserving data publishing : A survey of recent developments," ACM Computing Surveys, vol. 42, no. 4, p. 14, 2010.

[6] L. Sweeney, "k-anonymity: A model for protecting privacy," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 05, pp. 557–570, 2002.

[7] C. Dwork, "Differential privacy: A survey of results," in International Conference on Theory and Applications of Models of Computation, 2008, pp. 1–19.

[8] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 05, pp. 571–588, 2002.

[9] Y. He and J. F. Naughton, "Anonymization of set-valued data via top-down, local generalization," VLDB Endowment, vol. 2, no. 1,pp. 934–945, 2009.

[10] G. Poulis, A. Gkoulalas-Divanis, G. Loukides, S. Skiadopoulos, and C. Tryfonopoulos, "SECRETA: A system for evaluating and comparing relational and transaction anonymization algorithms," in International Conference on Extending Database Technology, 2014, pp. 620–623.

[11] G. Poulis, G. Loukides, A. Gkoulalas-Divanis, and S. Skiadopoulos, "Anonymizing data with

relational and transaction attributes," in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2013, pp. 353–369.

[12] M. Terrovitis, N. Mamoulis, and P. Kalnis, "Privacypreserving anonymization of set-valued data," VLDB Endowment, pp. 115–125, 2008.

[13] P. Nadkarni, L. Ohno-Machado, and W. Chapman, "Natural language processing: an introduction," Journal of the American Medical Informatics Association, vol. 18, no. 5, pp. 544–551, 2011.

[14] J. Aberdeen, S. Bayer, R. Yeniterzi, B. Wellner, C. Clark, D. Hanauer, B. Malin, and L. Hirschman, "The MITRE Identification Scrubber Toolkit: design, training, and assessment," International Journal of Medical Informatics, vol. 79, no. 12, pp. 849–859, 2010.

[15] A. Benton, S. Hill, L. Ungar, A. Chung, C. Leonard, C. Freeman, and J. H. Holmes, "A system for de-identifying medical message board text," BMC Bioinformatics, vol. 12 Suppl 3, p. S2, 2011.

[16] R. Chow, P. Golle, and J. Staddon, "Detecting privacy leaks using corpus-based association rules," in ACM International Conference on Knowledge Discovery and Data Mining, 2008, pp. 893–901.

[17] J. Gardner, L. Xiong, K. Li, and J. J. Lu, "Hide: heterogeneous information de-identification," in International Conference on Extending Database Technology: Advances in Database Technology, 2009, pp. 1116–1119.

[18] O. Ferr´andez, B. R. South, S. Shen, F. J. Friedlin, M. H. Samore, and S. M. Meystre, "BoB, a best-of-breed automated text deidentification system for vha clinical documents," Journal of the American Medical Informatics Association, vol. 20, no. 1, pp. 77–83, 2013.