JES
Journal of Engineering Sciences

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

AN ITERATIVE CLASSIFICATION SCHEME FOR SANITIZING LARGE-SCALE DATASETS

[1]KASHA KHAAVYA, mtech

Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering &Technology

[2]Mr.M.GANGAPPA, mtech,

Associate Professor

Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering &Technology

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.

Keywords: Privacy Preserving, Weak Structured Data Sanitization, Game Theory.

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

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 reidentification 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 re-identification 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

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

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.

RELATED WORK

A. 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 ε-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.

B. 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. 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.

C. 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. For example, while a structured portion of electronic medical records would generally have known sensitive categories, such as a patient's name, physician's notes do not have such labels, even though they may well refer to a patient's name, date of birth, and other potentially identifying information. While rule-based approaches, such as regular expressions, can automatically identify some of the sensitive entities, they have to be manually tuned to specific classes of data, and do not generalize well. 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. Numerous classification algorithms have been proposed for this purpose, including decision stumps, support vector machines

(SVM), conditional random fields (CRFs), hybrid

strategies that rely on rules and statistical learning models ensemble methods. Unfortunately, such PPDP algorithms fail to formally consider the adversarial model, which is crucial for the decision making of the data publisher. A recent work by Carrell considers enhancing such redaction methods by replacing removed identifiers with fake identifiers which appear real to a human reader. 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. Our model bears some relationship to a recent work by Li who also consider an adversary using machine learning to re-identify residual identifiers. However, our model combines this with a budget-limited attacker who can manually inspect instances; in addition, our publisher model involves the choice of a redaction policy, whereas Li et al. focus on the publisher's decision about the size of the training data, and use a traditional learning-based redaction approach. Second, we introduce a natural approach for sanitizing data that uses machine learning in an iterative framework. Notably, this approach performs significantly better than a standard application of CRFs, which is the leading approach for text sanitization to date, but can actually make use of arbitrary machine learning algorithms.

D. Game Theory in Security and Privacy

Our work can be seen within the broader context of game theoretic modeling of security and privacy, including a number of efforts that use game theory to make machine learning algorithms robust in adversarial environments. In both of these genres of work, a central element is an explicit formal threat (i.e., attacker) model, with the game theoretic analysis generally focused on computing defensive privacy-preserving strategies. None of this work to date, however, addresses the problem of PPDP of unstructured data with sensitive entities not known a priori.

III. MODEL

Before delving into the technical details, we offer a brief high-level intuition behind the

main idea in this paper. Suppose that a publisher uses a machine learning algorithm to identify sensitive instances in a corpus, these instances are then redacted, and the residual data is shared with an attacker. The latter, aspiring to uncover residual sensitive instances (e.g., identifiers) can, similarly, train a learning algorithm to do so (using, for example, a subset of published data that is manually labeled). At the high level, consider two possibilities: first, the learning algorithm enables the attacker to uncover a non-trivial amount of sensitive information, and second, the learning algorithm is relatively unhelpful in doing so. In the latter case, the publisher can perhaps breath freely: few sensitive entities can be identified by this attacker, and the risk of published data is low. The former case is, of course, the problem. However, notice that, in principle, the publisher can try out this attack in advance of publishing the data, to see whether it can in fact succeed in this fashion.

Data Utility

Next, we investigated the extent to which data utility can be retained in the face of a high privacy requirement. This served as motivation for GS (in comparison to simply suppressing all data), but we did not explicitly consider it in the theoretical analysis. Intuitively, GS should strike a reasonable balance: it stops immediately after a local optimum is reached. In our model, of course, there may be multiple local optima thereafter, but these would result in less data being published. Here, we evaluate the data utility of the published data using the publish ratio, which is defined as the proportion of the original number of entities in the published data. Figure 5 compares GS to cost-sensitive variants of the baseline algorithms (CRF, SVM, Adaboost, and Ensemble). GS preserves most of the data utility even when L/C is high. Specifically, in both of the EMR datasets over 98% of the data is published, even when L/C is quite high. The performance for the other three data sets is lower, but still, over 93% of the data is ultimately published, even with large L/C ratios. In contrast, when the loss due to re-identification is moderate or high, cost-sensitive algorithms essentially suppress most of the data, resulting in very low utility. GS therefore offers a far better

balance between risk and utility than the stateof-the-art alternatives.

IV. A GREEDY ALGORITHM FOR AUTOMATED DATA SANITIZATION

We can now present our iterative algorithm for automated data sanitization, which we term GreedySanitize. Our algorithm (shown as Algorithm 1) is simple to implement and involves iterating over the following steps: 1) compute a classifier on training data, 2) remove all predicted positives from the training data, and 3) add this classifier to the collection. The algorithm continues until a specified stopping condition is satisfied, at which point we publish only the predicted negatives, as above. While the primary focus of the discussion so far, as well as the stopping criterion, have been to reduce privacy risk, the nature of GreedySanitize is to also preserve as much utility as feasible: this is the consequence of stopping as soon as the re-identification risk is minimal. It is important to emphasize that GreedySanitize is qualitatively different from typical ensemble learning schemes in several ways. First, a classifier is retrained in each iteration on data that includes only predicted negatives from all prior iterations. To the best of our knowledge this is unlike the mechanics of any ensemble learning algorithm.1 Second, our algorithm removes the union of all predicted positives, whereas ensemble learning typically applies a weighted voting scheme to predict positives; our algorithm, therefore, is fundamentally more conservative when it comes to sensitive entities in the data. Third, the stopping condition is uniquely tailored to the algorithm, which is critical in enabling provable guarantees about privacy-related performance.

Algorithm 1 GreedySanitize(X), X : training data.

$H \leftarrow \{\}, k \leftarrow 0, h_0 \leftarrow \emptyset, D_0 \leftarrow X,$
repeat
  $H \leftarrow H \cup h_k$
  $k = k + 1$
  $h_k \leftarrow$LearnClassifier(D_{k-1})
  $D_k \leftarrow$RemovePredictedPositives(D_{k-1}, h_k)
until $T(H \cup h_k) - T(H) \geq 0$
return H

sufficiently successful, the publisher has a great deal to gain by redacting the sensitive entities an attacker would have found. Of course, there is no need to stop at this point: the publisher can keep simulating attacks on the published data, and redacting data labeled as sensitive, until these simulations suggest that the risk is sufficiently low. This,

indeed, is the main idea. However, many details are clearly missing: for example, what does an attacker do after training the learning algorithm, when, precisely, should the publisher stop, and what can we say about the privacy risk if data is published in this manner, under this threat model? Next, we formalize this idea, and offer precise answers to these and other relevant questions

V. 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 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.

VI. REFERENCES

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

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.

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

European Parliament and Council of the EuropeanUnion, "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.

B. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preservingdata publishing:Asurvey of recent developments,"ACM Computing Surveys, vol. 42, no. 4, p. 14, 2010.

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

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

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.

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.

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.

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