IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 10, October 2015.
www.ijiset.com
ISSN 2348 – 7968

On Syntactic Anonymity and Differential Privacy

A.V.Sriharsha, Dr. C. Parthasarathy

[1,2] Department of Information Technology, SCSVMV University,
Kancheepuram, Tamilnadu, India

Abstract

Recently, there has been a growing debate over approaches for handling and analyzing private data. Research has identified issues with syntactic approaches such as k-anonymity and l-diversity. Differential privacy, which is based on adding noise to the analysis outcome, has been promoted as the answer to privacy-preserving data mining. This paper looks at the issues involved and criticisms of both approaches. We conclude that both approaches have their place, and that each approach has issues that call for further research. We identify these research challenges, and discuss recent developments and future directions that will enable greater access to data while improving privacy guarantees.

Keywords: *Data Mining, Privacy Preserving Data Mining, Anonymity, Differential Privacy.*

1. Introduction

With the advent of tremendous growth of personal data and sophisticated, miniaturized data collecting devices, it has become a big challenge for the analysts to bring out the summary about the data without breach of privacy in the personal data. Data mining tools are being developed progressively more and more to draw trends and patterns. Of particular interest are data containing structured information on individuals. However, the use of data containing personal information has to be restricted in order to protect individual privacy.

The Official Statistics community has long recognized the privacy issues in both data publishing and release of statistics about data; Statistical Disclosure Limitation has primarily focused on tabular statistics, where a cell represents either a count of individuals matching that row/column (e.g., age range and income level), or a sum/average (e.g., years of education by race and state). Methods such as suppression (e.g., eliminating cells that reflect fewer than, say, five individuals), generalization by rounding values, or noise addition have been used to prevent individual identification. There has been extensive work for ensuring that the combinations of values from such tables cannot be "solved" to reveal exact values for individuals. Such a privacy aware release of statistics can be considered as PPDM.

Many techniques were proposed in the context of privacy preserving data publishing, including sampling, suppression, generalization (particularly of geographic details and numeric values), adding random noise, and value swapping. There has been work on showing how such methods can preserve data utility; for example, value swapping maintains univariate statistics, and if done carefully, it can also maintain controlled approximations to multivariate statistics [1]. The state of practice is based on standards for generalization of certain types of information (e.g., any disclosed geographic unit must contain at least 10,000 or 100,000 individuals). Following such standards for generalization of specific types of data, certain specifications of data generalization detail the types and specificity of data generalization that are deemed to make the data safe for releasing. A problem with this prescriptive approach is that each new domain demands new rules (e.g., due to different perceptions of the risk associated with re-identification and disclosure of data of different types, such as census data vs. health data vs. educational data). The proliferation of domains where data is being collected and may need to be published in a private manner makes this prescriptive approach impractical in the new big data world. Moreover, even this prescriptive approach does not provide a guarantee of individual privacy, but only an expectation of privacy. The result is that these prescriptive approaches are often very conservative, resulting in lower utility of the data. The fact that such standards exist, given the knowledge that they do not provide perfect privacy, suggests that PPDM and PPDP do not need to provide an absolute guarantee of privacy; adequate privacy (which may vary by domain) can be sufficient.

The Official Statistics research community has developed numerous methods for generating privacy-protected microdata, but this has not resulted in a standard approach to PPDP. One difficulty is that much of the work emphasizes methods to produce microdata sets, often for a particular domain. This makes the work difficult to generalize. There has recently been an explosion of attempts in the computing research community to provide formal mathematical definitions that either bound the probability of identification of individuals, or the specificity of information released about individuals. While much of the earlier (and current) work in Statistical Disclosure Limitation is highly relevant, a comprehensive survey and comparative analysis of those methods is beyond the scope of this paper. Herein, we focus only on

ISSN 2348 – 7968

the recent definitions offered by the computing research community, and indicate claims or interpretations that we perceive as misunderstandings that are impacting the progress of research in this field.

Probably the first formal mathematical model to achieve wide visibility in the computing research community was k-anonymity, proposed by Samarati and Sweeney. This model requires that each of the released records be indistinguishable from at least k - 1 other records when projected on the quasi-identifier attributes. As a consequence, each individual may be linked to sets of records of size at least k in the released anonymized table, whence privacy is protected to some extent. This is accomplished by modifying table entries. The above seminal studies, and the majority of the subsequent studies, modify data by generalizing table entries.

However, other techniques have also been suggested to achieve record indistinguishability. All those techniques first partition the data records into blocks, and then release information on the records within each block so that the linkage between quasi-identifier tuples and sensitive values within a given block is fully blurred.

Despite the enhanced privacy that those models offer with respect to the basic model of k-anonymity, they are still susceptible to various attacks. As a result of those attacks, it seems that part of the research community has lost faith in those privacy models. The emergence of differential privacy [2], a rigorous notion of privacy based on adding noise to answers to queries on the data, has revolutionized the field of PPDM. There seems to be a widespread belief that differential privacy and its offsprings are immune to those attacks, and that they render the syntactic models of anonymity obsolete. In this paper we discuss the problems with syntactic anonymity and argue that, while all those problems are genuine, they can be addressed within the framework of syntactic anonymity. We further argue that differential privacy too is susceptible to attacks, as well as having other problems and (often unstated) assumptions that raise problems in practice.

2. Syntactic Models of Anonymity

Herein we survey some of the main models of syntactic anonymity. Most of those models are based on generalizing table entries. Such data distortion preserves the truthfulness of data, in the sense that a generalized value defines a group of possible original values. (Generalization also includes, as a special case, the operation of suppression; suppression also defines a group of possible original values since usually the dictionary of possible values for each attribute is known.) Those models provide privacy for the data subjects by rendering some sort of record indistinguishability.

k-Anonymity and most of the models that evolved from it are based on partitioning the database records to blocks and then anonymizing the records so that those that appear in the same block become indistinguishable. In k-anonymity, all blocks are required to be of size at least k, and the records within each block are replaced with their closure, being the minimal generalized record that generalizes all of them.

3. Differential Privacy

3.1 Tables and Figures

In the middle of the previous decade, the research community began exploring new privacy notions that are not based on a syntactic definition of privacy, most prominent among which is differential privacy. Differential Privacy is a formal definition relating uncertainty at an individual level to the noise or randomization used in a privacy mechanism.

Owing to its rigorous approach and formal privacy guarantees, differential privacy has started to be adopted by a growing part of the academic community as the only acceptable definition of privacy, sometimes to the extent that it is viewed as rendering previous privacy models obsolete.

A key value of differential privacy is that it is proof against an attacker with strong background knowledge. The strong attacker assumed by differential privacy knows all records in the database except for one record, but is still unable to violate the privacy of the individual behind that record: the query result would be essentially indistinguishable (modulo $e\varepsilon$) whether that individual's record was or was not in the data. The second breakthrough made by differential privacy is in formulating a general mechanism for adding noise to any continuous-valued query towards meeting that privacy measure. Another merit of differential privacy is that it is composable, in the sense that it can support multiple queries on the data.

4. PPDM and PPDP

There is a fundamental difference between the assumptions that underlie differential privacy and those that underlie syntactic privacy models. In fact, those two seemingly competing approaches are targeting two

ISSN 2348 – 7968

different playgrounds. k-anonymity and other syntactic notions of anonymity target PPDP. A typical scenario of PPDP is that in which a hospital wishes to release data about its patients for public scrutiny of any type. The hospital possesses the data and is committed to the privacy of its patients. The goal is to publish the data in an anonymized manner without making any assumptions on the type of analysis and queries that will be executed on it. Once the data is published, it is available for any type of analysis.

Differential privacy, on the other hand, typically targets PPDM. In PPDM, as opposed to PPDP, the query that needs to be answered must be known prior to applying the privacy preserving process. In the typical PPDM scenario, the data custodian maintains control of the data and does not publish it. Instead, the custodian responds to queries on the data, and ensures that the answers provided do not violate the privacy of the data subjects. In differential privacy this is typically achieved by adding noise to the data, and it is necessary to know the analysis to be performed in advance in order to calibrate the level of noise to the global sensitivity of the query and to the targeted differential privacy parameter ε. While some differential privacy techniques (e.g., private histograms) are really intermediate analysis rather than a final data mining model, it is still necessary for the data custodian to know what analysis is intended to be performed.

Data publishing is a widespread practice hence, it is important to develop appropriate techniques for PPDP. First of all, even if the data custodian knows that the data will be used for classification, it may not know how the user may analyze the data. The user often has application-specific bias towards building the classifier. For example, some users prefer accuracy while others prefer interpretability, or some prefer recall while others prefer precision. In other cases, visualization or exploratory analysis of the data may guide the user toward the right approach to classification for their particular problem. Publishing the data provides the user a greater flexibility for data analysis. It should be noted that while data publishing techniques can be customized to provide better results for particular types of analysis data which is published towards a specific data mining goal can still be used for other data mining goals as well.

5. Summary and Conclusions

This study examined two types of privacy models: syntactic models of anonymity and differential privacy. Those two approaches are sometimes perceived as competing approaches, and that one can be used instead of the other. The first point that we made in this study is that the above conception is wrong. We explained that the syntactic models are designed for privacy-preserving data publishing (PPDP) while differential privacy is typically applicable for privacy-preserving data mining (PPDM). Hence, one approach cannot replace the other, and they both have a place alongside the other.

Our conclusion is that while differential privacy is a valuable weapon in the fight to both maintain privacy and foster use of data, it is not the universal answer. It provides a way to deal with a previously unanswered question in PPDM: how to ensure that the model developed does not inherently violate privacy of the individuals in the training data? While there are still issues related to both privacy and utility to be resolved.

In conclusion, in both paradigms, the issues raised should be viewed as opportunities for future research, rather than a call for abandoning one approach or the other. Advances in both paradigms are needed to ensure that the future provides reasonable protections on privacy as well as supporting legitimate learning from the ever-increasing data about us.

Sincere Acknowledgements: All the seminal works of the Transaction on Data Privacy lead by Latanya Sweeny, Cynthia, D'Work, Tamir Tassa, Samarati and et. al.

References

[1] R. A. Moore, Jr. "*Analysis of the kim-winkler algorithm for masking microdata files – how much masking is necessary and sufficient? conjectures for the development of a controllable algorithm*". Statistical Research Division Report Series RR 96-05, U.S. Bureau of the Census, Washington, DC., 1996.

[2] C. Dwork. *Differential privacy*. In ICALP (2), pages 1–12, 2006.

[3] *Standard for privacy of individually identifiable health information*. Federal Register, Special Edition:768–769, Oct. 1 2007. 45 CFR 164.514(b)(2).

[4] K. E. Emam, F. K. Dankar, R. Issa, E. Jonker, D. Amyot, E. Cogo, J.-P. Corriveau, M. Walker, S. Chowdhury, R. Vaillancourt, T. Roffey, and J. Bottomley. "*A globally optimal k-anonymity method for the de-identification of health information*". Journal of the American Medical Informatics Association, 16:670–682, 2009.

[5] L. Sweeney, "*k-anonymity: a model for protecting privacy*", International Journal of Uncertainty, Fuzziness and Knowledge based Systems, vol. 10, no. 5, pp. 557–570, 2002.

[6] L. Sweeney. Achieving "*k-anonymity privacy protection using generalization and suppression*". International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 571- 588.

[7] Noman Mohammed, Dima Alhadidi, Benjamin C. M. Fung, and Mourad Debbabi, "*Secure Two-Party Differentially Private*

ISSN 2348 – 7968

Data Release for Vertically-Partitioned Data", IEEE Transactions On Depedable And Secure Computing, Vol. 11, No. 1, pp. 59 © January/February 2014.

[8] Papadimitriou, S., Li, F., Kollios, G., Yu, P. S. 2007. "*Time series compressibility and privacy*". In Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB), pp. 459–470.

[9] R. Agrawal, A. Evfimievski, R. Srikant, "*Information Sharing Across Private Databases*", Proc. ACM-SIGMOD, pp. 86 - 97, 2003.

[10] Raymond Chi Wing Wong, Jiuyong Li, Ada Wai Chee Fu, Ke Wang, "(α,k)-*Anonymity: An Enhanced kAnonymity Model for Privacy Preserving Data Publishing*", KDD'06, Philadelphia, Pennsylvania, USA. © 2006 ACM, August 20–23, 2006 pp 236-242.

[11] Richard H. Rand, Dieter Armbruster, "*Perturbation Methods, Bifurcation Theory and Computer Algebra*", © 1987 by Springer-Verlag New York Inc.

[12] R. J. Bayardo and R. Agrawal, "*Data privacy through optimal k-anonymization*", in Proceedings of the 21st International Conference on Data Engineering (ICDE), Tokyo, Japan, April 2005, pp. 217–228.

[13] Samarati, P. 2001. "*Protecting respondents' identities in microdata release*". IEEE Transactions on Knowledge and Data Engineering. 13, 6, 1010–1027.

[14] Aristides Gionis, Tamir Tassa, "*k-Anonymization with Minimal Loss of Information*", IEEE Transactions on Knowledge and Data Engineering, Vol 21. No.2 February 2009.

[15] Thomas A. Lasko, Staal A. Vinterbo, "*Spectral Anonymization of Data*", IEEE Transactions on Knowledge and Data Engineering, Vol 22. No.3 March 2010.

[16] Christos Dimitrakakis, Aris Gkoulalas-Divanis, Aikaterini Mitrokotsa Vassilios S. Verykios, Yücel Saygin (Eds.), "*Privacy and Security Issues in Data Mining and Machine Learning*", © Springer-Verlag Berlin Heidelberg, September 2010.

[17] Cynthia Dwork, F. McSherry, K. Nissim, and A. Smith, "*Calibrating Noise to Sensitivity in Private Data Analysis*," in TCC 2006.

[18] Daniel Shiffman, "*The Nature of Code*", http://natureofcode.com/ © 2012.

[19] Elisa Bertino, Beng Chin Ooi, Yanjiang Yang, Robert H. Deng, "*Privacy and Ownership Preserving of Outsourced Medical Data*", Proceedings of the 21st International Conference on Data Engineering (ICDE 2005) 1084-4627/05.

[20] Fosca Giannotti, Laks V. S. Lakshmanan, Anna Monreale, Dino Pedreschi, and Hui (Wendy) Wang, "*Privacy-Preserving Mining of Association Rules From Outsourced Transaction Databases*", IEEE Systems Journal, Vol. 7, No. 3, Pp. 385 © September 2013.

[21] V. S. Iyengar, "*Transforming data to satisfy privacy constraints*", in Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, AB, Canada, July 2002, pp. 279–288.

[22] Wenliang Du, Zhijun Zhan, "*A Practical Approach to Solve Secure Multi-Party Computation Problems*", (2002). Electrical Engineering and Computer Science. Paper 19. pp. 655-664.

[23] Yaping Li, Minghua Chen, Qiwei Li, and Wei Zhang, "*Enabling Multilevel Trust in Privacy Preserving Data Mining*", IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 9, Pp. 1598, © September 2012.

[24] Machanavajjhala, A.,Kifer, D., Abowd, J. M.,Gehrke, J., Vilhuber, L. 2008. "*Privacy: Theory meets practice on the map*". In Proceedings of the 24th IEEE International Conference on Data Engineering (ICDE). 277–286.

[25] Meyerson, A., Williams, R. 2004. "*On the complexity of optimal k-anonymity*". In Proceedings of the 23rd ACM SIGMOD-SIGACT-SIGART PODS. ACM, New York, 223–228.

[26] Michal Sramka, Reihaneh Safavi-Naini, J¨org Denzinger, "*An Attack on the Privacy of Sanitized Data That Fuses the Outputs of Multiple Data Miners*", KDD'09, © 2009 ACM, pp 432-440.

[27] Noman Mohammed, Dima Alhadidi, Benjamin C. M. Fung, and Mourad Debbabi, "*Secure Two-Party Differentially Private Data Release for Vertically-Partitioned Data*", IEEE Transactions On Dependable and Secure Computing, Vol. 11, No. 1, Pp. 59 © January/February 2014.

A.V. Sriharsha, has completed his B.Tech in Computer Science & Engineering from Andhra University and M.Tech in Information Technology from Sathyabhama University, Chennai. He is currently working as Associate Professor in the Department of CSE, Sree Vidyanikethan Engineering College, A. Rangampet, Tirupati, A.P. He is pursuing his Ph.D. His main research interest includes Data Mining, Information Retrieval and DBMS.

Dr. C. Parthasarathy, is a Assistant Professor in Department of Information Technology in SCSVMV University, Kancheepuram, He has acquired M.C.A., M.Phil., M.Tech., PhD and has publish 22 national and international papers in reputed conferences and journals. His interested areas are cryptography, security in networks, mobile ad hoc networks and data mining.