Report on International Workshop on Privacy and Anonymity in the Information Society (PAIS 2008)

Li Xiong
Department of Math & Computer Science
Emory University
lxiong@mathcs.emory.edu

Traian Marius Truta
Department of Computer Science
Northern Kentucky University
trutat1@nku.edu

Farshad Fotouhi
Department of Computer Science
Wayne State University
fotouhi@wayne.edu

1 Introduction

While the ever increasing computational power together with the huge amount of individual data collected daily by various agencies is of great value for our society, they also pose a significant threat to individual privacy. As a result legislators for many countries try to regulate the use and the disclosure of confidential information. Various privacy regulations (such as USA Health Insurance Portability and Accountability Act, Canadian Standard Association Model Code for the Protection of Personal Information, Australian Privacy Amendment Act 2000, etc.) have been enacted in many countries all over the world. Data privacy and protecting individual anonymity have become a mainstream avenue for research. While privacy is a topic discussed everywhere, data anonymity recently established itself as an emerging area of computer science. Its goal is to produce useful computational solutions for releasing data, while providing scientific guarantees that the identities and other sensitive information of the individuals who are the subjects of the data are protected.

The Workshop on Privacy and Anonymity in the Information Society (PAIS 2008)[1] was held on March 29, 2008, co-located with the 11th International Conference on Extending Database Technology (EDBT 2008) in Nantes, France. It was the first in its series and the mission of the workshop is to provide an open yet focused platform for researchers and practitioners from computer science and other fields that are interacting with computer science in the privacy area such as statistics, healthcare informatics, and law to discuss and present current research challenges and advances in data privacy and anonymity research.

[1] http://cscdb.nku.edu/pais/

2 Workshop Themes and Program

The workshop program included a keynote speech and 8 paper presentations. The paper presentations were divided into 3 sessions. There were 30 participants who attended the workshop. The workshop was very interactive, with the audience raising many questions for the speakers and a lively discussion following the technical presentations. Several overall themes emerged from the presentations and discussions, including location privacy, distributed privacy protection, query auditing, k-anonymization and its applications, and micro-aggregation. We report and discuss each of them below.

2.1 Location Privacy

The proliferation of mobile communications is leading to new services based on the ability of providers to determine, with increasing precision, the geographic location of the accessing device. While these applications and services promise enormous consumer benefit, privacy concerns abound, and must be addressed

before new services and applications are accepted by consumers.

The keynote speech addressed the timely topic of location privacy. The talk was given by Josep Domingo-Ferrer, professor of Computer Science from Rovira i Virgili University of Tarragona, Catalonia, and the UNESCO Chair in Data Privacy. The talk was titled "Location privacy via unlinkability: an alternative to cloaking and perturbation" [4]. In the talk, he summarized that the usual approach to location privacy is to cloak and/or perturb the positions or trajectories of the mobile objects. However, he argued that if unlinkability of the various interactions between a mobile object and the service or control system can be afforded and achieved, neither cloaking nor perturbation is unnecessary. The unlinkability results in higher privacy for the mobile object and better accuracy of the aggregated mobility information gathered by the service/control system. He illustrated the feasibility of the approach in the scenario of car-to-car communication.

2.2 Distributed Privacy Protection

Distributed privacy-preserving data mining deals with data sharing across multiple distributed data sources for specific mining tasks. The problem is a specific example of the secure multi-party computation (MPC) problem. In MPC, a given number of participants, each having a private data, wants to compute the value of a public function. A MPC protocol is secure if no participant can learn more from the description of the public function and the result of function. While there are general secure MPC protocols, they require substantial computation and communication costs and are impractical for multi-party large database problems.

The paper titled "Distributed Privacy Preserving k-Means Clustering with Additive Secret Sharing" [3] considered a distributed privacy preserving data mining scenario where the data is partitioned vertically over multiple sites and the involved sites then perform clustering without revealing their local databases. For this setting, the authors proposed a new protocol for privacy preserving k-means clustering based on additive secret sharing. They showed that the new protocol is more secure than the current state of the art while the communication and computation cost is considerably less which is crucial for data mining applications.

Protecting privacy for content-sharing P2P networks is another distributed privacy protection problem of increasing importance. The privacy issues in this context include anonymity of uploaders and downloaders, linkability (correlation between uploaders and downloaders), content deniability, data encryption and authenticity, and data disclosure.

The paper titled "Design of PriServ, a privacy service for DHTs" [5] addressed the data disclosure problem in P2P networks. When sharing data for different purposes, data privacy can be easily violated by untrustworthy peers which may use data for unintended purposes. A basic principle of data privacy is purpose specification which states that data providers should be able to specify the purpose for which their data will be collected and used. The work applied the Hippocratic database principles to P2P systems to enforce purpose-based privacy. They focused on Distributed Hash Tables (DHTs), and proposed PriServ, a privacy service which prevents privacy violation by prohibiting malicious data access. The performance evaluation of the approach through simulation shows that the overhead introduced by PriServ is small.

2.3 Query Auditing

Research in statistical databases is focused on enabling queries on aggregate information (e.g. sum, count) from a database without revealing individual records. A key technique is query restriction which includes schemes that check for possible privacy breaches by keeping audit trails and controlling overlap of successive aggregate queries.

The paper titled "A Bayesian Approach for on-Line Max and Min Auditing" [1] considered the online max and min query auditing problem. Given a private association between fields in a data set, a sequence of max and min queries that have already been posed about the data, their corresponding answers and a new query, the objective is to deny the answer if a private information is inferred or give the true answer otherwise. The authors gave a probabilistic definition of privacy and demonstrated that max and min queries, without no duplicates assumption, can be audited by means of a Bayesian network. Moreover, the auditing approach is able to manage user prior-knowledge.

2.4 K-Anonymization and its Applications

Data anonymization has been extensively studied in recent years and a few principles have been proposed that serve as criteria for judging whether a published dataset provides sufficient privacy protection. Notably, the seminal work of k-anonymity, requires a set of k records (entities) to be indistinguishable from each other based on a quasi-identifier set. A large body of work contributes to transforming a dataset to meet a privacy principle (dominantly k-anonymity) using techniques such as generalization, suppression (removal), permutation and swapping of certain data values while minimizing certain cost metrics. The workshop features a few papers on k-anonymization and its applications in novel domains.

The paper titled "Protecting Privacy in Recorded Conversations" [2] considered the problem of privacy protection in the domain of speech technology. While speech corpora are important to professionals in the field of speech technology, they are often prevented from being shared due to privacy rules and regulations. Efforts to scrub this data to make it shareable can result in data that has been either inadequately protected or data that has been rendered virtually unusable due to the loss resulting from suppression. This work attempted to address these issues by developing a scientific workflow that combines proven techniques in data privacy with controlled audio distortion resulting in corpora that have been adequately protected with minimal information loss.

The paper titled "Data Utility and Privacy Protection Trade-off in K-Anonymization" [7] revisited the issue of balancing between data utility and privacy for k-anonymization. While existing methods try to maximize utility while satisfying a required level of protection, their work attempted to optimize the trade-off between utility and protection. The authors introduced a measure that captured both utility and protection, and an algorithm that exploited this measure using a combination of clustering and partitioning techniques. The author showed that the method is capable of producing k-anonymization with required utility and protection trade-off and with a performance scalable to large datasets.

The paper titled "An Efficient Clustering Method for k-Anonymization" [6] proposed a new clustering method for k-anonymization. The authors argued that in order to minimize the information loss due to anonymization, it is crucial to group similar data together and then anonymize each group individually. The work proposed a clustering based anonymization method and compared it with another state-of-the-art clustering based k-anonymization method. Their experiments and analysis showed that the proposed method outperforms the existing method with respect to time complexity, information loss and resilience to outliers.

2.5 Microaggregation

Microaggregation is a hot topic in the field of Statistical Disclosure Control (SDC) and one of the most employed microdata protection methods. The main idea is to build clusters of at least k original records, and then replace them with the centroid of the cluster. This is one way to achieve k-anonymity.

The paper titled "Attribute Selection in Multivariate Microaggregation" [8] addressed the issue of attribute grouping for microaggregation. When the number of attributes is large, a common practice is to split the dataset into smaller blocks of attributes. This paper showed that, besides the specific microaggregation method employed, the value of the parameter k, and the number of blocks in which the dataset is split, there exists another factor which can influence the quality of microaggregation: the way in which the attributes are grouped to form the blocks. When correlated attributes are grouped in the same block, the statistical utility of the protected dataset is higher. In contrast, when correlated attributes are dispersed into different blocks, the achieved anonymity is higher. The authors also presented quantitative evaluations of such statements.

The paper titled "Micro-aggregation-Based Heuristics for p-sensitive k-anonymity: One Step Beyond" [9] adapted micro-aggregation based techniques for p-sensitive k-anonymity, a recently defined sophistication of k-anonymity. The p-sensitive k-anonymity requires that there be at least p different values for each confidential attribute within the records sharing a combination of key attributes. While the original algorithm was based on generalizations and suppressions, this work show that k-anonymity and p-sensitive k-anonymity can be achieved in numerical data sets by means of micro-aggregation heuristics properly adapted to deal with this task. The authors presented and evaluated two heuristics for p-sensitive k-anonymity which, being based on micro-aggregation, overcame most of the drawbacks resulting from the generalization and suppression method.

3 Final Remarks

PAIS 2008 workshop is among several series of recent workshops focusing on the various issues of data privacy and security including: Secure Data Management Workshop (SDM) co-located with VLDB, International Workshop on Privacy, Security and Trust (PinKDD) co-located with SIGKDD, International Workshop on Privacy Data Management (PDM) co-located with ICDE, Workshop on Privacy Aspects of Data Mining (PADM) co-located with ICDM, and Practical privacy preserving data mining (P3DM) co-located with SDM.

The PAIS workshop is unique in that it focuses on the topic of privacy and in particular anonymity in the general information society. The workshop organizers and attendees envision a series of workshops building upon the success of this workshop.

4 Acknowledgements

Putting together PAIS 2008 was a team effort. First of all, we would like to thank our keynote speaker and the authors for providing the quality content of the program. In addition, we would like to express our gratitude to the program committee, who worked very hard in reviewing papers and providing suggestions for their improvements. Finally, we would like to thank the UNESCO Chair in Data Privacy for their travel support and the EDBT 2008 conference for their support of the workshop.

References

[1] G. Canfora and B. Cavallo. A bayesian approach for on-line max and min auditing. In *PAIS*, pages 12–20, 2008.

[2] S. Cunningham and T. M. Truta. Protecting privacy in recorded conversations. In *PAIS*, pages 26–35, 2008.

[3] M. C. Doganay, T. B. Pedersen, Y. Saygin, E. Savas, and A. Levi. Distributed privacy preserving k-means clustering with additive secret sharing. In *PAIS*, pages 3–11, 2008.

[4] J. Domingo-Ferrer. Location privacy via unlinkability: an alternative to cloaking and perturbation. In *PAIS*, pages 1–2, 2008.

[5] M. Jawad, P. Serrano-Alvarado, and P. Valduriez. Design of priserv, a privacy service for dhts. In *PAIS*, pages 21–25, 2008.

[6] J.-L. Lin and M.-C. Wei. An efficient clustering method for k-anonymization. In *PAIS*, pages 46–50, 2008.

[7] G. Loukides and J. Shao. Data utility and privacy protection trade-off in k-anonymisation. In *PAIS*, pages 36–45, 2008.

[8] J. Nin, J. Herranz, and V. Torra. Attribute selection in multivariate microaggregation. In *PAIS*, pages 51–60, 2008.

[9] A. Solanas, F. Sebé, and J. Domingo-Ferrer. Micro-aggregation-based heuristics for p-sensitive k-anonymity: one step beyond. In *PAIS*, pages 61–69, 2008.