Fraud, Anonymization, and Privacy in the Internet: A Database Perspective

Divyakant Agrawal   Amr El Abbadi

Department of Computer Science, University of California, Santa Barbara
Santa Barbara, CA 93106-5110, USA
{agrawal, amr}@cs.ucsb.edu

Data is everywhere, and manifests itself in various formats. Data can be publicly available and can be privately owned. Data can be persistent on a server or ephemeral in a data stream. Society depends on data and hence the security, privacy and reliability of the data are critical in diverse ways. In this talk I will touch on various security aspects of data in different contexts that arise in diverse internet applications.

1 Internet Advertising Fraud Detection

In Internet advertising networks, Internet advertisers pay Internet publishers to display advertisements on their Web sites and drive traffic to the advertisers from surfer's clicks. Since publishers are paid by the traffic they drive to advertisers, there is an incentive for dishonest publishers to *inflate* the number of impressions (renderings of advertisements) and clicks their sites generate [2].

Analysis of traffic streams can be executed on aggregate levels, such that the individuals' identities are not revealed, and still satisfactorily detect fraud. Due to the problem scale, fraud detection in advertising networks is attractive as a quintessential streaming application. The approach was introduced in [7], where a simple Bloom-Filter-based algorithm was proposed to detect duplicates in a stream of impressions or clicks. Experiments on real data were revealing. More than 27% of the clicks were suspicious. Interestingly, one of the advertisements was clicked 10,781 times by the same surfer in one day. Even more shocking is that the fraud was performed using a primitive attack. In [8], a detection algorithm was proposed for the sophisticated *hit inflation* attack in [2], which involves a coalition of dishonest publishers. The detection algorithm entails identifying associations in a HTTP stream. In [9], we model discovering single publisher attacks as a problem of finding correlations in multi-dimensional data, and an algorithm is devised for detecting such attacks in their most general forms.

2 Anonymizing Social Networks

The increasing popularity of social networks has initiated a fertile research area in information extraction and data mining. Although such analysis can facilitate better understanding of sociological, behavioral, and other interesting phenomena, there is growing concern about personal privacy being breached, thereby requiring effective anonymization techniques. In this context we will describe methods for anonymizing the relevant contents of the social network graph while preserving critical properties of the graph.

Recently, there has been considerable interest in the analysis of the *weighted* network model where the social networks are viewed as weighted graphs. The weighted graph model is used for analyzing various social phenomena as well as traditional applications on weighted graphs such as *shortest paths*, *spanning trees*, *k-Nearest Neighbors (kNN)* etc. The semantics of the edge weights depend on the application (such as "degree of friendship", "trustworthiness", etc.), or the property being modeled.

Our solution [3] to the problem of edge weight anonymization is to model the weighted graph based on the property to be preserved, and then reassign edge weights to obtain the anonymized graph satisfying the model. To be specific, we preserve any graph property that is expressible in terms of inequalities involving linear combinations of edge weights. Many algorithms make decisions based on the actual numerical values of the edge weights and we model this decision in terms of variables. Decisions made at each step of the algorithm can simply be expressed as inequalities involving the edge-weights. Thus, the execution of an algorithm processing a graph can be modeled as a set of linear inequalities involving the edge weights as *variables*, and this results in a system of linear inequalities. If the edge weights are reassigned as any solution of the system of inequalities, this would ensure that the properties of the graph remain unchanged w.r.t the algorithm being modeled. The model can therefore be formulated as a Linear Programming (LP) problem. Our approach is therefore independent of the semantics of edge-weights, and is general enough to encompass many important algorithms.

15th International Conference on Management of Data
COMAD 2009, Mysore, India, December 9–12, 2009

3 Privacy preserving operations on Data

The advent of cloud computing, as well as the ubiquitous availability of internet information service providers has resulted in different settings where users are concerned about the privacy of their data. This may arise when private data is outsourced to a service provider, and the client wishes to retrieve some of this private data in an efficient privacy preserving manner, or when a client wishes to retrieve publicly available data in a way that does not reveal the specific interests of the client to the service provider.

Most approaches proposed for secure and private data outsourcing are based on data encryption [11]. Unfortunately, the computational complexity of encrypting and decrypting data to execute a query increases the query response time [1]. Agrawal et al. [1] show that computing a privacy preserving intersection using encryption results in a very high time complexity. Furthermore, executing a query over encrypted data is a significant challenge.

In [5, 4], we proposed using Shamir's secret sharing algorithm [10] to execute privacy preserving operations among a set of distributed data warehouses. In [5], a middleware, Abacus, was developed to support selection, intersection and join operations. In [4], this work was generalized to any type of database, and used distributed third parties and Shamir's secret sharing algorithm to secure information and support privacy preserving selection, intersection, join and aggregation operations (such as SUM and MIN/MAX operations) on a set of distributed databases.

Private retrieval of public data is useful when a client wants to query a public data service without revealing the specific query data to the server. A general and promising use is in personalized search and recommendation subscriptions through big internet information service providers such as Google, Yahoo, and Microsoft. On one hand, users need these public services in their daily lives. On the other hand, users are concerned that their private profile data or their personal tastes might be disclosed or compromised through analysis or inferences.

Computational Private Information Retrieval (cPIR) [6] is able to achieve complete privacy for a client, but is deemed impractical since it involves expensive computation on all the data on the server. Besides, it is inflexible if the server charges clients based on the exposed data. k-Anonymity [12], on the other hand, is flexible and cheap for anonymizing the querying process, but is vulnerable to privacy and security threats. We propose a practical and flexible approach for the private retrieval of public data called *Bounding-Box PIR* (bbPIR) [13]. Using bbPIR, a client specifies both privacy requirements and the service charge budget. The server satisfies the client's requirements, and at the same time achieves overall good performance in terms of computation and communication costs. bbPIR generalizes cPIR and k-Anonymity in that the bounding box can include as much as all the data on the server or as little as just k data items. The effectiveness of bbPIR compared to cPIR and k-Anonymity is verified using experimental evaluation.

Acknowledgement

Thanks to Ahmed Metwally, Fatih Emekçi, Shyam Antony, Sudipto Das and Shiyuan Wang for the many hours, days and years working together on the topics discussed here. Also thanks to Ömer Eğecioğlu for his collaboration on the anonymization work. Special thanks to Sudipto Das for his assistance with this presentation.

References

[1] R. Agrawal, A. Evfimievski, and R. Srikant. Information sharing across private databases. In *Proc. of the 2003 ACM SIGMOD international conference on on Management of data*, pages 86–97, 2003.

[2] V. Anupam, A. Mayer, K. Nissim, B. Pinkas, and M. Reiter. On the Security of Pay-Per-Click and Other Web Advertising Schemes. In *Proceedings of the 8th WWW International Conference on World Wide Web*, pages 1091–1100, 1999.

[3] S. Das, Ömer Eğecioğlu, and A. El Abbadi. Anonymizing Weighted Social Network Graphs. In *ICDE*, 2010.

[4] F. Emekçi, D. Agrawal, A. El Abbadi, and A. Gulbeden. Privacy preserving query processing using third parties. In *ICDE*, page 27, 2006.

[5] F. Emekci, D. Agrawal, and A. E. Abbadi. Abacus: A distributed middleware for privacy preserving data sharing across private data warehouses. In *ACM/IFIP/USENIX 6th International Middleware Conference*, 2005.

[6] E. Kushilevitz and R. Ostrovsky. Replication is not needed: Single database, computationally-private information retrieval. In *FOCS*, pages 364–373, 1997.

[7] A. Metwally, D. Agrawal, and A. El Abbadi. Duplicate Detection in Click Streams. In *Proceedings of the 14th WWW International World Wide Web Conference*, pages 12–21, 2005.

[8] A. Metwally, D. Agrawal, and A. El Abbadi. Using Association Rules for Fraud Detection in Web Advertising Networks. In *Proceedings of the 31st VLDB International Conference on Very Large Data Bases*, pages 169–180, 2005.

[9] A. Metwally, F. Emekci, D. Agrawal, and A. El Abbadi. Sleuth: Single Publisher attack detection using correleation Hunting . In *Proceedings of Very Large Data Base Endowment*, volume 1, pages 1217–1228, 2008.

[10] A. Shamir. How to share a secret. *Commun. ACM*, 22(11):612–613, 1979.

[11] R. Sion. Secure data outsourcing. In *Proceedings of the 33rd International Conference on Very Large Data Bases*, pages 1431–1432, 2007.

[12] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.

[13] S. Wang, D. Agrawal, and A. El Abbadi. Generalizing PIR for Practical Private Retrieval of Public Data. Technical Report 2009-16, Computer Science, UC Santa Barbara, October 2009.