Statement of Research — Alexandre Evfimievski

Overview

My prior research has been mainly in the area of privacy preserving data mining. It included such topics as: using randomization for preserving privacy of individual transactions in association rule mining; secure two-party computation of joins between two relational tables, set intersections, join sizes, and supports of vertically partitioned itemsets; improving space and time efficiency in privacy preserving algorithms by means of error-correcting codes and pseudorandomness; and developing probability-based methodology for privacy evaluation, centered on the notion of privacy breach. Some earlier work included updating files with polylogarithmic communication and mining hidden information in data streams.

In the future, I would like to move more towards machine learning and data mining, extracting hidden information and structure from large amounts of data. Meanwhile, I may continue working on privacy, for example developing efficient and secure algorithms for information integration across multiple private databases and exploring further the statistical approach to privacy and its applications.

Research Experience

Privacy Preserving Mining of Association Rules. Consider a company that collects product preferences data from its clients, for mining statistically significant associations. We have one server and many clients, each client having a set of items (a "transaction"). The server wants to find item-sets that occur frequently in transactions, and then find association rules. However, the clients do not want to release their private transactions to the server. We proposed that the clients randomize their transactions by deleting some items and inserting other items, thereby hiding private information within randomness, and submit these new transactions to the server. The server uses statistical estimation to efficiently recover original item-set frequencies from randomized data; the recovered frequencies and their variances are given to an Apriori-like mining algorithm. The degree of randomization depends on the needed amount of privacy. The algorithm was implemented and tested on real-life data [1, 2, 3]. This work was done during my summer internship at IBM Almaden Research Center in 2001, joint with Dr. Ramakrishnan Srikant and Dr. Rakesh Agrawal, and later with Prof. Johannes Gehrke at Cornell University.

Information Sharing Across Private Databases. Generally speaking, given a database query spanning multiple private databases, we wish to efficiently compute the answer to the query while revealing as little extra information as possible, apart from the query result. In our work, we have two servers, Sender and Receiver, each having a relational table; the two tables have a discrete attribute in common. Receiver wants to compute the equijoin between the two tables. Sender permits Receiver to have the join, but wants to keep private all records with the values of the common attribute unmatched by Receiver's table. A special case is when Sender and Receiver have two sets, and Receiver wants to compute the intersection between these sets. We have developed cryptographic protocols, based on commutative encryption and random hash functions, that achieve private computation of join and intersection while disclosing essentially nothing extra. As examples of applications for these protocols we suggested selective document sharing between enterprises and medical research, and evaluated efficiency of the protocols for these applications [4]. This work was done during my summer internship at IBM Almaden Research Center in 2002, joint with Dr. Ramakrishnan Srikant and Dr. Rakesh Agrawal.

Limiting Privacy Breaches in Data Randomization. We looked more generally at using randomization for preserving privacy. Consider a server and a client with a private record; the client does not want to disclose its record to the server. Given a certain randomization procedure for the client's record, what can we claim about the privacy protection it achieves? A new methodology, called *amplification*, has been developed in [5] that answers this question in terms of a limit on *privacy breaches*. A privacy breach is a situation when some property of a private record is unlikely (to the server) in the absence of any knowledge, but becomes likely once the server receives the randomized record. The new methodology has been applied to the case of association rules and shown to yield practically useful results. The generality of this methodology has allowed us to extend privacy guarantees to randomization with compression based on error-correcting codes. This work is done at Cornell University, joint with Prof. Johannes Gehrke and Dr. Ramakrishnan Srikant.

Ongoing Research. We are working to combine the techniques for approximate query evaluation over data streams, called *sketches*, with cryptography and randomization in order to develop more efficient privacy preserving protocols. We developed an efficient protocol for secure two-party approximate scalar product computation, with applications to equijoin size computation and item-set frequency computation in vertically partitioned market-basket data. The algorithm performs only polylogarithmic number of cryptographic operations and communications, with respect to the dataset size, and randomizes the approximate answer to hide its exact value. To analyze privacy-preserving properties of the algorithm, we extended the notion of privacy breaches and defined *probabilistic* privacy breaches that occur only if there is a non-negligible probability of a (deterministic) privacy breach. The amplification approach [5] was extended accordingly. In this framework, we are making steps to develop a "statistical" approach to privacy, as opposed to a "computational" approach where security of cryptographic protocols is based on computational intractability of disclosure of private information. This is a joint work with Prof. Johannes Gehrke, Dr. Ramakrishnan Srikant, and Dr. Rakesh Agrawal.

Earlier Research. One result concerns communication complexity of updating a large file over a slow communication link. Consider two servers, SENDER and RECEIVER; SENDER has a newer version of some large file, whereas RECEIVER has an older version of the same file. Assume that the newer version can be obtained from the older one by a short sequence of common text editor operations. This sequence is not assumed to be known. The result consists in a randomized protocol that updates RECEIVER's file (with high probability) by actually sending a polylogarithmic amount of information over the link, with respect to the file size [6]. This work was part of my undergraduate studies at the Dept. of Mathematics and Mechanics in Moscow State University, under the supervision of Prof. Nikolai Vereshchagin and Alexander Shen.

Another direction was mining for hidden information. Consider a data stream of multidimensional Boolean vectors; suppose that there is a small set of hidden variables whose values determine each vector. We want to find these hidden variables. The dependency of visible vectors on hidden variables was given by a low-depth Boolean circuit. Our algorithm was based on k-means clustering and feature selection, and was tested on artificial data. This was done under the supervision of Prof. Johannes Gehrke at Cornell University.

Future Research Interests

In the future, I would like to move from "strictly" privacy preserving data mining towards data mining and machine learning in general. I am interested in developing efficient algorithms for learning hidden information within large amounts of data. To achieve this, I would rather use sound mathematical approach that emphasizes understanding and digs to the bottom, while still avoiding pure scholasticism, keeping in mind practical applications. Here are some examples of problems I find interesting:

- In a database of graphs or other data structures, find records similar (in structure) to a given record;
- When doing reinforcement learning over complex multidimensional data, combine it with unsupervised learning in order to discover structure and dependencies in the data, so that the reinforcement information is used more efficiently;
- Unsupervised learning with background knowledge: Given a dataset and some already mined information about it (such as a separation into clusters), mine something new, "orthogonal" to what is known;
- After some information about the data has been collected via data mining, make it precise (as a mathematical statement) so that it can be accumulated and used in reasoning about the data, as well as in further data mining.

I am ready to work on any problem that looks fundamental for machine learning and relevant to understanding the nature of intelligence.

In privacy preserving data mining, it is interesting to further develop the theory of statistical privacy, possibly making connections to other areas of statistics such as information theory and statistical learning theory. Privacy evaluation methods are needed that would allow to deal with efficient privacy preserving protocols that are not "sterile", protocols that do disclose some extra information. For such protocols, the disclosure should be bounded, proven to be very unlikely, or proven to be irrelevant to any given sensitive question. Then these methods may be used for constructing scalable algorithms for private database processing. A related direction is to combine conventional cryptography (based on computational hardness) and secure multiparty computation with this statistical approach to privacy.

References

[1] Alexandre Evfimievski, Ramakrishnan Srikant, Rakesh Agrawal, and Johannes Gehrke. Privacy preserving mining of association rules. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery in Databases and Data Mining*, pages 217–228, Edmonton, Alberta, Canada, July 23–26 2002.

[2] Alexandre Evfimievski, Ramakrishnan Srikant, Rakesh Agrawal, and Johannes Gehrke. Privacy preserving mining of association rules (invited journal version). *Information Systems*, 29(4):343–364, June 2004.

[3] Alexandre Evfimievski. Randomization in privacy-preserving data mining. *SIGKDD Explorations: Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining*, 4(2):43–48, December 2002.

[4] Rakesh Agrawal, Alexandre Evfimievski, and Ramakrishnan Srikant. Information sharing across private databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 86–97, San Diego, California, USA, June 9–12 2003.

[5] Alexandre Evfimievski, Johannes Gehrke, and Ramakrishnan Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the 22-nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 211–222, San Diego, California, USA, June 9–11 2003.

[6] Alexandre Evfimievski. A probabilistic algorithm for updating files over a communication link. In *Proceedings of the 9-th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 300–305, San Francisco, California, USA, January 25–27 1998.