Anonymized Data: Generation, Models, Usage

Graham Cormode Divesh Srivastava
AT&T Labs–Research
{graham,divesh}@research.att.com

ABSTRACT

Data anonymization techniques have been the subject of intense investigation in recent years, for many kinds of structured data, including tabular, item set and graph data. They enable publication of detailed information, which permits ad hoc queries and analyses, while guaranteeing the privacy of sensitive information in the data against a variety of attacks. In this tutorial, we aim to present a *unified* framework of data anonymization techniques, viewed through the lens of data uncertainty. Essentially, anonymized data describes a set of possible worlds, one of which corresponds to the original data. We show that anonymization approaches such as suppression, generalization, perturbation and permutation generate different working models of uncertain data, some of which have been well studied, while others open new directions for research. We demonstrate that the privacy guarantees offered by methods such as k-anonymization and ℓ-diversity can be naturally understood in terms of similarities and differences in the sets of possible worlds that correspond to the anonymized data. We describe how the body of work in query evaluation over uncertain databases can be used for answering ad hoc queries over anonymized data in a principled manner. A key benefit of the unified approach is the identification of a rich set of new problems for both the Data Anonymization and the Uncertain Data communities.

Categories and Subject Descriptors

H.2.0 [**DATABASE MANAGEMENT**]: Security, Integrity and Protection

General Terms

Algorithms, Security

Keywords

Anonymization, Privacy, Uncertain Databases

1. MOTIVATION

In the 21st Century, the data management research community has become acutely aware of the need for privacy of data concerning individuals. As the ability to collect and store more and more information about every single action in life has grown, huge amounts of details about individuals are now recorded in database systems. This is both a blessing and a curse: a blessing, since this allows many possibilities for enriching people's lives, through new and improved services, and a greater knowledge of people's preferences and desires; a curse, since often this knowledge concerns very private aspects of personal lives which can be damaging if widely publicized. For example, knowledge of people's queries to search engines can enable a new early warning system for fast-spreading flu outbreaks[1], but even an inadvertent public disclosure of this detailed data may violate basic privacy rights of individuals. Similarly, knowledge of people's locations, along with their preferences, can enable a variety of useful location-based services, but public disclosure of an individual's movements over time can have serious consequences for privacy.

This leads to a question at the heart of modern data processing: how to take advantage of this explosion of data to the betterment of society as a whole, without violating the right to privacy of any individual. To address this question, the community has expended great effort in designing techniques to effectively *anonymize* data so that the detailed results can be published and shared with others. The aim is that a malicious party should be unable to use this published data to infer anything "private" about the entities represented, while an honest party should still be able to perform a variety of ad hoc analyses and find results which are close to their true values on the original data. The data anonymization techniques proposed in the literature can be classified in several dimensions:

1. **Nature of data**: Techniques have been proposed for (a) tabular data, which represents information about entities (e.g., people), their quasi-identifiers (e.g., age, gender, zip code), and their sensitive information (e.g., salary, disease); (b) item set data, which represents transactional (or "market basket") data, associating people with the sets of items purchased in a transaction; and (c) graph data, which represents sensitive associations between entities (e.g., people in social networks).

2. **Anonymization approaches**: Proposed anonymization techniques use a variety of approaches, including (a) suppression, where information (e.g., gender) is removed from the data; (b) generalization, where information (e.g., age) is coarsened into sets (e.g., into age ranges); (c) perturbation, where noise is added to the data (e.g., salary); and (d) permutation, where sensitive associations between entities (e.g., purchase of medication by a person) are swapped.

3. **Anonymization objectives**: Various privacy goals are achieved by applying particular approaches (as above) until the resulting data has certain properties, such as (a) k-anonymity, where each individual in the database must be indistinguishable from $k-1$ others; (b) ℓ-diversity, which seeks to en-

SIGMOD'09, June 29–July 2, 2009, Providence, Rhode Island, USA.
ACM 978-1-60558-551-2/09/06.

[1] http://www.nytimes.com/2008/11/12/technology/internet/12flu.html

sure sufficient diversity in the sensitive information associated with individuals; and (c) other methods which aim to prevent certain inferences based on assumptions about knowledge held by an attacker.

A unified framework for these facets of data anonymization techniques would clearly be very useful. In this tutorial, we aim to present such a unified framework, viewed *through the lens of data uncertainty*. In recent years, motivated by quality-aware scenarios like imprecise observations, there has been a growth of interest in models and algorithms for handling Uncertain Data, i.e. data describing many alternatives (or "possible worlds"), instead of the single world described by regular data. In particular, several working models of uncertain data have been proposed, which compactly describe many possible worlds by outlining the alternatives for possible events, and the correlations or independencies between them. Given data presented in such models, there has been much effort in studying how to efficiently evaluate queries and perform analysis over the uncertain data, and come up with a compact description of the possible answers to the queries.

We observe that there is an important connection between the topics of Data Anonymization and Uncertain Data. The process of data anonymization introduces uncertainty into data that was initially certain. Essentially, anonymized data describes a set of possible worlds, one of which corresponds to the original data. Our unified framework provides the following perspectives on the problems of data anonymization, developed in detail in the tutorial.

- Anonymization approaches such as suppression, generalization, perturbation and permutation generate different working models of uncertain data, some of which have been well studied by the Uncertain Data community, while others open new directions for research.

- Privacy guarantees offered by methods such as k-anonymization and ℓ-diversity equate to requiring that it is not possible to deduce certain facts about the original data with high confidence from the anonymized data: the fraction of possible worlds (or likelihood, in a probabilistic setting) must be low.

- The lens of data uncertainty also relates to another important challenge: how to answer ad hoc queries on anonymized data in a principled manner. Now, if the anonymized data is viewed as existing in an uncertain data model, then answering queries on the data translates to evaluating queries in that uncertain model: the topic of much recent study.

This unified framework provides benefits to both the Data Anonymization and the Uncertain Data communities, providing the former with principled methods for query evaluation, the latter with a natural application area for uncertain data, and both with a rich set of challenging problems.

2. TUTORIAL OUTLINE

Our tutorial is example driven, and organized as follows.

2.1 Data Anonymization: Basics

We first introduce the problem of *anonymization* of private data: using a variety of techniques to modify the original data in such a way that the original "sensitive" data is masked. The need for anonymization is motivated by many legal and ethical requirements for protecting private, personal data. The intent is that anonymized data can be shared freely with other parties, who can perform their own analysis and investigation of the data. We will present examples which show the dangers of data release without rigorous anonymization, such as the AOL Search Data example and attacks on Netflix data.

Once the goal of anonymization is formalized, a fundamental trade-off is established between two aspects: the privacy goals of the data owners, and the utility goals of the data users. Most work in this area fixes a particular privacy requirement, and then tries to optimize the utility while guaranteeing this level of privacy. We will discuss various definitions of what is meant by both "privacy" and "utility", using examples.

2.2 Uncertain Data: Basics

We present various models of *uncertain data*: data best characterized by many possible worlds, instead of the single world described by regular data. Both complete models (those that can represent any set of possible worlds) and working models (intuitive models that capture common forms of uncertainty) will be presented, and illustrated via examples [12, 1]. The role of constraints in uncertain data models will be highlighted: these allow the encoding of various correlations and exclusions between different items within the possible worlds.

2.3 Generating Anonymized/Uncertain Data

We discuss techniques for dealing with data that is best structured as a table. We also present methods for data with different structural properties, such as item set or graph data. We present the key works on these problems in a common framework, identifying the key approaches and paradigms. We also discuss their limitations, such as assumptions about the strength of the attacker or susceptibility to certain attacks.

Entities and Tabular Data. Recent interest in anonymizing data about entities (such as people) and their attributes, which can be represented in the form of single tables, began with work on k-anonymization [14, 11]. The key techniques here are *generalization*, which replaces an exact item with a *set* of possibilities, naturally seen as a description of possible worlds, and *suppression*, which conceptually replaces an item with the universe of possibilities. Various enhancements have been proposed, such as ℓ-diversity [10], which place additional requirements on the generated possible worlds.

A second class of techniques are based on the concept of *perturbations*. For example, recent work by Aggarwal [2] has proposed adding noise from a specified distribution to yield a probabilistic model of k-anonymity.

A third class of techniques use the concept of *permutations* [16, 19]. These vertically partition the table, and state that there are bijections between identified groups of tuples on either side of the partition. This generates an uncertain data model which interestingly does not seem to have been well-studied.

Associations and Graph Data. Single tables are not the best way to represent associations between complex entities, and sensitive associations are not well anonymized by tabular anonymization techniques. An example that has attracted significant recent study is of data from *social networks*. Here, the data records information about individuals but also the connections *between them*. This data may better be thought of as a *graph* over the various entities. Simply treating this data as a table ignores the graph structure, and results in anonymized data with little or no value; instead, anonymization must take into account the particular graph nature of the data.

A first class of work tries to thwart attackers by modifying the link structure of the graph by adding or removing edges. This is motivated by results of Backstrom *et al.* [3] which show that a powerful attacker with significant background knowledge can learn information about some individuals from the unlabeled graph. First efforts in this direction modify the graph by edge additions (or deletions) so that there are at least k nodes with the same degree [8]; or the stronger requirement that each node must have k others with the same (labeled) neighborhood [21].

A second class of work avoids altering the link structure, and instead masks the details of the links by generalization and permutation techniques. The attacker can be modeled as a machine learning algorithm, the aim being to limit the ability of such algorithms to correctly predict links [20]. Techniques to mask the graph data include forming nodes into groups and revealing only the number of edges between pairs of groups [7], and masking the mapping from nodes to their corresponding entities via a permutation-based approach [6].

Associations and Item Sets. Another example of associations arises with item sets, which represents transactional (or "market basket") data, associating people with the sets of items purchased in a transaction. This can also be represented by a table, but this table becomes very large and sparse, and tabular anonymization similarly fails to preserve the utility; it may be better represented as a collection of *sets* or a *bipartite graph* between transactions and items. Expectations about what background knowledge a malicious attacker might have are also different in this case [15, 17].

2.4 Using Anonymized/Uncertain Data

Given data described in uncertain models, there has been a lot of effort in studying how to efficiently evaluate a variety of queries (including relational algebra and top-k) over the uncertain data, and come up with a compact description of the possible answers to the queries. Deterministic approaches lead to upper and lower bounds on the query answer values or sets. Probabilistic approaches lead to either an expected answer to the query, or tail bounds on its possible values. We will survey methods that are useful for query answering in the uncertain models generated from data anonymization [13, 4]. Here, there is further potential to link the two areas, via more explicit definition of the model in which anonymized data is published, and more application of query answering on such data sets.

2.5 Research Directions and Open Problems

Our framework provides benefits to both the Data Anonymization and the Uncertain Data communities, providing the former with principled methods for query evaluation, the latter with a natural application area for uncertain data, and both with a rich set of challenging problems. We identify a few in our tutorial: (a) modeling and query evaluation over working models of uncertain data that can represent permutations; (b) a study of (deterministic and probabilistic) query evaluation techniques for ad hoc analyses over anonymized/uncertain data, including OLAP queries, clustering and association rules; and (c) anonymizing unstructured (text) data and semi-structured (XML) data.

Many other related topics are beyond the scope of this tutorial, and will be mentioned briefly if at all. These include (a) "Anonymity-preserving data collection" [18, 5] where individuals hold their own data, and do not trust anyone else to perform the anonymization; instead, they use cryptographic techniques to reveal an arbitrary permutation between the data items and the individuals (b) *Interactive querying* relies on having an active participant accepting queries from researchers, and deciding whether to answer them, and how much noise to add to the results. (c) *Differential privacy* is a paradigm developed from the theoretical community which aims to ensure that the inclusion or exclusion of a single individual from the data set make no statistical difference to the results found.

3. INTENDED AUDIENCE AND COVERAGE

Our goal is to present a coherent starting point for SIGMOD attendees who are interested in understanding the rapidly growing area of structured data anonymization, as well as uncertain data. We will not assume any background in either data anonymization or uncertain data models, but will introduce both together, giving broad coverage of many of the key ideas, making it appropriate for graduate students seeking new areas to study and researchers active in the field alike.

4. ABOUT THE TUTORS

Graham Cormode is a Principal Member of Technical Staff in the Database Research Department at AT&T Labs-Research. His Ph.D. is from the University of Warwick, and he completed postdoctoral studies at the DIMACS Center in Rutgers University. He has presented tutorials on data stream computations in Bristol, UK and Caserta, Italy, and on "Querying and Tracking Distributed Data Streams" (with Minos Garofalakis) at VLDB 2006, SIGMOD 2007 and EDBT 2008.

Divesh Srivastava is the head of the Database Research Department at AT&T Labs-Research. He received his Ph.D. from the University of Wisconsin, Madison, and his B.Tech. from the Indian Institute of Technology, Bombay. He has presented tutorials on "Data Stream Query Processing" (with Nick Koudas) at VLDB 2003 and ICDE 2005, and on "Record Linkage: Similarity Measures and Algorithms" (with Nick Koudas and Sunita Sarawagi) at VLDB 2005 and SIGMOD 2006.

5. REFERENCES

[1] S. Abiteboul, P. C. Kanellakis, and G. Grahne. On the representation and querying of sets of possible worlds. *Theor. Comput. Sci*, 78(1):158–187, 1991.

[2] C. C. Aggarwal. On unifying privacy and uncertain data models. In *IEEE International Conference on Data Engineering*, 2008.

[3] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore are thou R3579X? Anonymized social networks, hiddern patterns and structural steganography. In *International Conference on World Wide Web (WWW)*, 2007.

[4] O. Benjelloun, A. D. Sarma, C. Hayworth, and J. Widomn. An introduction to ULDBs and the Trio system. *IEEE Data Engineering Bulletin*, 29(1):5–16, Mar. 2006.

[5] J. Brickell and V. Shmatikov. Efficient anonymity-preserving data collection. In *ACM SIGKDD*, 2006.

[6] G. Cormode, D. Srivastava, T. Yu, and Q. Zhang. Anonymizing bipartite graph data using safe groupings. In *International Conference on Very Large Data Bases*, 2008.

[7] M. Hay, D. Jensen, G. Miklau, D. Towsley, and P. Weis. Resisting structural re-identification in anonymized social networks. In *International Conference on Very Large Data Bases*, 2008.

[8] K. Liu and E. Terzi. Towards identity anonymization on graphs. In *ACM SIGMOD International Conference on Management of Data*, 2008.

[9] L. Liu. From data privacy to location privacy: Models and algorithms. In *International Conference on Very Large Data Bases*, 2007.

[10] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. ℓ-diversity: Privacy beyond k-diversity. In *IEEE International Conference on Data Engineering*, 2006.

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

[12] A. D. Sarma, O. Benjelloun, A. Y. Halevy, and J. Widom. Working models for uncertain data. In *IEEE International Conference on Data Engineering*, 2006.

[13] D. Suciu and N. N. Dalvi. Foundations of probabilistic answers to queries. In *ACM SIGMOD International Conference on Management of Data*, 2005.

[14] L. Sweeney. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based systems*, 10(5):557–570, 2002.

[15] M. Terrovitis, N. Mamoulis, and P. Kalnis. Privacy-preserving anonymization of set-valued data. In *International Conference on Very Large Data Bases*, 2008.

[16] X. Xiao and Y. Tao. Anatomy: Simple and effective privacy preservation. In *International Conference on Very Large Data Bases*, 2006.

[17] Y. Xu, K. Wang, A. W.-C. Fu, and P. S. Yu. Anonymizing transaction databases for publication. In *ACM SIGKDD*, 2008.

[18] Z. Yang, S. Zhong, and R. N. Wright. Anonymity-preserving data collection. In *ACM SIGKDD*, 2005.

[19] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu. Aggregate query answering on anonymized tables. In *IEEE International Conference on Data Engineering*, 2007.

[20] E. Zheleva and L. Getoor. Preserving the privacy of sensitive relationships in graph data. In *International Workshop on Privacy, Security and Trust in KDD (PinKDD)*, 2007.

[21] B. Zhou and J. Pei. Preserving privacy in social networks against neighborhood attacks. In *IEEE International Conference on Data Engineering*, 2008.