Anonymized Data: Generation, Models, Usage

Graham Cormode, Divesh Srivastava

AT&T Labs-Research, Florham Park, NJ 07932 USA
{graham,divesh}@research.att.com

***Abstract*— Data anonymization techniques 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 that include the original data. We show that anonymization approaches generate different working models of uncertain data, and that the privacy guarantees offered by k-anonymization and ℓ-diversity can be naturally understood in terms of the sets of possible worlds that correspond to the anonymized data. Work in query evaluation over uncertain databases can hence be used for answering ad hoc queries over anonymized data. We identify new research problems for both the Data Anonymization and the Uncertain Data communities.**

I. Motivation

The data engineering 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.

The goal is to take advantage of this data without violating the right to privacy of any individual. Accordingly, the community has expended great effort in designing techniques to effectively *anonymize* data so that 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. Existing data anonymization techniques 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 ensuring the published 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 ensure sufficient diversity in the sensitive information associated with individuals; and (c) other goals which aim to prevent certain inferences based on assumptions about knowledge held by an attacker.

In this tutorial, we aim to present a unified framework for these facets of anonymization, viewed *through the lens of data uncertainty*. In recent years there has been a growth of interest in models and algorithms for handling Uncertain Data which describes 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. Much effort has been made in studying how to efficiently evaluate queries and perform analysis over the uncertain data, and give a compact description of the possible answers to the queries. We observe an important connection between 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. This framework gives perspectives on the problems of data anonymization, developed in detail in the tutorial:
(a) Existing anonymization approaches generate different working models of uncertain data, some of which have been well studied; others open new research directions.
(b) Privacy guarantees equate to requiring that it is not possible to deduce certain facts about the original data with high confidence from the anonymized data: the corresponding fraction of possible worlds must be low.
(c) 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.

II. Tutorial Outline

The tutorial presents a coherent starting point for understanding the areas of structured data anonymization and uncertain data. We assume no background in either, and instead introduce both together, giving broad coverage of key ideas. `http://tinyurl.com/anon09` has links to the slides and references.

A. Data Anonymization and Uncertain Data: Basics

We introduce the problem of *anonymization*: to modify the original data so that the original "sensitive" data is masked. This is motivated by many legal and ethical requirements for protecting such 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 present examples which show the pros and cons of data release including the Netflix prize data and the AOL search data. Once the goal is formalized, there is a trade-off between the privacy goals of the data owners, and the utility goals of the data users.

We then present various models of *uncertain data*: data representing many possible worlds. Complete models (that can represent any set of possible worlds) and working models (incomplete models that capture common forms of uncertainty) are presented, and illustrated via examples. Constraints in uncertain data models are highlighted: these encode correlations or exclusions between items in the possible worlds.

B. Generating Anonymized/Uncertain Data

Entities and Tabular Data. Recent interest in anonymizing data about entities and their attributes, which can be represented in the form of single tables, began with work on k-anonymization. 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, which place additional requirements on the generated possible worlds. A second class of techniques are based on the concept of *perturbations*. A third class of techniques use the concept of *permutations*; these vertically partition the table, and state that there are bijections between identified groups of tuples on either side of the partition.

Associations: Graph Data and Itemsets. 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 the connections *between* individuals, which can be thought of as a *graph*. Viewing this as a table ignores the graph structure, and gives anonymized data with little value; instead, anonymization must take into account the particular graph nature of the data.

A first class of work modifies the link structure of the graph, motivated by results which show that a powerful attacker with significant background knowledge can learn information about some individuals from the unlabeled graph. Efforts in this direction use edge additions (or deletions) so that there are at least k nodes with the same degree or so that for each node there are $k-1$ others with the same (labeled) neighborhood.

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. Techniques to mask the graph data include forming nodes into groups and revealing only the number of edges between pairs of groups, and masking the mapping from nodes to their corresponding entities via a permutation-based approach.

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 the background knowledge of a malicious attacker are also different in this case.

C. Using Anonymized/Uncertain Data

There has been much effort in studying how to efficiently evaluate queries (including relational algebra and top-k) over uncertain data, and come up with a compact description of the possible answers. 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 survey methods that are useful for query answering in the uncertain models generated from data anonymization. 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.

D. Research Directions and Open Problems

This 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. These include: (a) modeling and query evaluation over working models optimized for anonymized data; (b) studying query evaluation techniques for ad hoc analyses, such as OLAP queries, clustering and association rules; and (c) anonymizing unstructured (text) data and semi-structured (XML) data.

Other related topics discussed include: (a) "Anonymity-preserving data collection" where individuals hold their own data, work together to perform the anonymization; (b) *Interactive querying* where an active participant accepts queries and decides how much noise to add to the results; (c) *Differential privacy*, a paradigm which ensures that the inclusion or exclusion of a single individual from the data set make no statistical difference to the results.