Privacy-Preserving Data Publishing

Privacy-Preserving Data Publishing

Bee-Chung Chen

Yahoo! Research
USA
beechun@yahoo-inc.com

Daniel Kifer

Penn State University
USA
dkifer@cse.psu.edu

Kristen LeFevre

University of Michigan
USA
klefevre@eecs.umich.edu

Ashwin Machanavajjhala

Yahoo! Research
USA
mvnak@yahoo-inc.com

now
the essence of knowledge

Boston – Delft

Foundations and Trends® in Databases

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is B.-C. Chen, D. Kifer, K. LeFevre and A. Machanavajjhala, Privacy-Preserving Data Publishing, Foundation and Trends® in Databases, vol 2, nos 1–2, pp 1–167, 2009

ISBN: 978-1-60198-276-6

Foundations and Trends® in Databases

Volume 2 Issues 1–2, 2009

Editorial Scope

Foundations and Trends® in Databases covers a breadth of topics relating to the management of large volumes of data. The journal targets the full scope of issues in data management, from theoretical foundations, to languages and modeling, to algorithms, system architecture, and applications. The list of topics below illustrates some of the intended coverage, though it is by no means exhaustive:

- Data Models and Query Languages
- Query Processing and Optimization
- Storage, Access Methods, and Indexing
- Transaction Management, Concurrency Control and Recovery
- Deductive Databases
- Parallel and Distributed Database Systems
- Database Design and Tuning
- Metadata Management
- Object Management
- Trigger Processing and Active Databases
- Data Mining and OLAP
- Approximate and Interactive Query Processing
- Data Warehousing
- Adaptive Query Processing
- Data Stream Management
- Search and Query Integration
- XML and Semi-Structured Data
- Web Services and Middleware
- Data Integration and Exchange
- Private and Secure Data Management
- Peer-to-Peer, Sensornet and Mobile Data Management
- Scientific and Spatial Data Management
- Data Brokering and Publish/Subscribe
- Data Cleaning and Information Extraction
- Probabilistic Data Management

Information for Librarians

Foundations and Trends® in Databases, 2009, Volume 2, 4 issues. ISSN paper version 1931-7883. ISSN online version 1931-7891. Also available as a combined paper and online subscription.

Foundations and Trends® in
Databases
Vol. 2, Nos. 1–2 (2009) 1–167
© 2009 B.-C. Chen, D. Kifer, K. LeFevre and
A. Machanavajjhala
DOI: 10.1561/1900000008

the essence of knowledge

Privacy-Preserving Data Publishing

Bee-Chung Chen[1], Daniel Kifer[2], Kristen LeFevre[3] and Ashwin Machanavajjhala[4]

[1] *Yahoo! Research, USA, beechun@yahoo-inc.com*
[2] *Penn State University, USA, dkifer@cse.psu.edu*
[3] *University of Michigan, USA, klefevre@eecs.umich.edu*
[4] *Yahoo! Research, USA, mvnak@yahoo-inc.com*

Abstract

Privacy is an important issue when one wants to make use of data that involves individuals' sensitive information. Research on protecting the privacy of individuals and the confidentiality of data has received contributions from many fields, including computer science, statistics, economics, and social science. In this paper, we survey research work in privacy-preserving data publishing. This is an area that attempts to answer the problem of how an organization, such as a hospital, government agency, or insurance company, can release data to the public without violating the confidentiality of personal information. We focus on privacy criteria that provide formal safety guarantees, present algorithms that sanitize data to make it safe for release while preserving useful information, and discuss ways of analyzing the sanitized data. Many challenges still remain. This survey provides a summary of the current state-of-the-art, based on which we expect to see advances in years to come.

Contents

1

Introduction

I have as much privacy as a goldfish in a bowl.

— Princess Margaret

Privacy is an important issue when one wants to make use of data that involve individuals' sensitive information, especially in a time when data collection is becoming easier and sophisticated data mining techniques are becoming more efficient. It is no surprise that research on protecting the privacy of individuals and the confidentiality of data has received many contributions from many fields such as computer science, statistics, economics, and social science. With the current rate of growth in this area it is nearly impossible to organize this entire body of work into a survey paper or even a book. Thus we have proceeded with a more modest goal. This survey describes research in the area of privacy-preserving data publishing. We are mainly concerned with data custodians such as hospitals, government agencies, insurance companies, and other businesses that have data they would like to release to analysts, researchers, and anyone else who wants to use the data. The overall intent is for the data to be used for the public good: in the evaluation of economic models, in the identification of social trends, and in the pursuit of the state-of-the-art in various fields. Usually, such data contain personal information such as medical records, salaries, and so on, so that a straightforward release of data is not appropriate. One approach to solving this problem is to require data users to sign non-disclosure agreements. This solution will need significant legal resources and enforcement mechanisms and may be a barrier to wide dissemination of the data. Furthermore, this cannot protect against data theft even when the victim takes reasonable precautions. Thus, it is important to explore technological solutions which anonymize the data prior to its release. This is the focus of this survey.

In Section 1, we begin by describing the information-protection practices employed by census bureaus (Section 1.1), and we motivate the importance of considering privacy protection in data publishing through a number of real-world attacks (Section 1.2). We then use a simple example (Section 1.3) to introduce the problem and its challenges (Section 1.4). Section 2 is devoted to formal definitions of privacy, while Section 3 is devoted to ways of measuring the utility of sanitized data or the information lost due to the sanitization process. In Section 4, we present algorithms for sanitizing data. These algorithms seek to output a sanitized version of data that satisfies a privacy definition and has high utility. In Section 5, we discuss how a data user can make use of sanitized data. Then, in Section 6, we discuss how an adversary might attack sanitized data. In Section 7, we cover emerging applications and their associated research problems and discuss difficult problems that are common to many applications of privacy-preserving data publishing and need further research.

Having explained what this survey is about, we will now briefly mention what this survey is not about. Areas such as access control, query auditing, authentication, encryption, interactive query answering, and secure multiparty computation are considered outside the scope of this paper. Thus we do not discuss them except in places where we deem this to be necessary. We also focus more on recent work as many of the older ideas have already been summarized in book and survey form [4, 263, 264]. Unfortunately, we cannot cover every technique in detail and so the choice of presentation will largely reflect the authors' bias. We have tried to cover as much ground as possible and regret any inadvertent omissions of relevant work.

1.1 Information Protection in Censuses, Official Statistics

The problem of privacy-preserving data publishing is perhaps most strongly associated with censuses, official processes through which governments systematically collect information about their populations. While emerging applications such as electronic medical records, Web search, online social networks, and GPS devices have heightened concerns with respect to collection and distribution of personal information, censuses have taken place for centuries, and considerable effort has focused on developing privacy-protection mechanisms in this setting. Thus, we find it appropriate to begin this survey by describing some of the diverse privacy-protection practices currently in place at national census bureaus and affiliated statistical agencies around the world.

1.1.1 Public-Use Data

Most related to the topic of this survey is the problem of releasing public-use data sets. Worldwide, many (though not all) governmental statistical agencies distribute data to the public [54, 58, 133, 234] to be used, for example, in demographic research. However, it is also a common belief that these public-use data sets should not reveal information about individuals in the population. For example, in the United States, Title 13 of the US Code requires that census information only be collected to produce statistics, and that census employees be sworn to protect confidentiality.

Thus, over the years, government statistical agencies have developed a variety of mechanisms intended to protect individual privacy in public-use data. (This research area is commonly known as *statistical disclosure limitation* or *confidentiality*, and it is a subset of the broader field of *official statistics*.) Historically, this work has focused on two main classes of data that are commonly released by governmental agencies:

- **Aggregate count data (*contingency tables*)** Contingency tables contain frequency count information, tabulated on the basis of one of more variables.[1] For example, a contingency table might contain a population count based on *Zip Code*, *Age Range*, and *Smoking Status*; i.e., in each zip code and each age range, how many people smoke?

- **Non-aggregate data (*Microdata*)** Microdata are simply conventional (non-aggregate) data, where each row refers to a person in the population.

In order to limit the possibility that an individual could be identified from the public-use data, statistical agencies commonly use a combination of techniques [54, 58, 59, 95, 133, 234, 257]; however, statistical disclosure limitation experts at statistical agencies do not typically provide details of the mechanisms used for confidentiality, only generic descriptions. A recent report [95] outlines, in general terms, the practices of the various federal agencies in the United States. (We will describe some of these techniques in more detail in Section 4.)

- **Cell suppression and noise addition (for contingency tables)** In contingency tables, it is common to suppress cells with small counts (*primary suppression*), as well as additional cells that can be inferred using marginal totals (*complementary suppression*). Similarly, it is common to make small perturbations to the counts.
- **Data swapping (for microdata and contingency tables)** Data swapping is a method of making controlled changes to microdata; modified contingency tables can also be re-computed from the results. This technique was used in the United States during the 1990 and 2000 censuses [101].
- **Sampling, geographic coarsening, and top/bottom-coding (for microdata)** For microdata, it is common to only release a subset of respondents' data (e.g., a 1% sample). In addition, it is common to restrict geographic identifiers to regions containing at least a certain population. (In the United States, this is typically 100,000 [257].) It is also common to "top-code" and "bottom-code" certain values. For example, if there are sufficiently few respondents

[1] In SQL, this is analogous to releasing the answer to a COUNT(*) query with one or more attributes in the GROUP BY clause.

over age 90, then a top-coding approach would replace all ages ≥ 90 with the value 90.

- **Synthetic data (for microdata)** Finally, sometimes synthetic data are generated. The idea is to produce data with similar distributional characteristics to the original microdata. The US Census Bureau is considering using a synthetic data approach to release microdata following the 2010 census [272].

Many of the above-mentioned mechanisms for microdata and contingency table sanitization, respectively, have been implemented in the μ- and τ- Argus software packages [127, 128]; these packages have also been used extensively by Statistics Netherlands.

The US Census Bureau also provides an online (real-time) system called the American FactFinder Advanced Query System [122], which provides custom tabulations (count queries) from the census data. Disclosure control in this system is done primarily by applying queries to the sanitized (e.g., swapped) microdata, and also by imposing cell suppression and top-coding rules to the results.

1.1.2 Restricted-Use Data, Research Data Centers, and Remote Servers

While many statistical agencies release sanitized public-use data sets, there is also a commonly held belief that certain data (e.g., high-precision geographical units) cannot be sanitized enough to release, or that the process would yield the data useless for certain kinds of research. For these reasons, federal agencies in the United States [256, 225], Canada [46], and Germany [219] have also set up secure *research data centers* to allow outside researchers to access more precise and detailed data. The idea is to provide a secure physical facility, staffed by census personnel, in which vetted researchers can carry out approved studies using computers with limited external access. In the United States, there are approximately a dozen such locations. Before conducting a study, a researcher must undergo a background check and provide a sworn statement. Before removing results or data from the center, the results must undergo a strict disclosure review, which is conducted by Census Bureau personnel. Similarly, a variety of countries provide "virtual" secure research data centers (also known as *remote access servers*) that serve a similar purpose [214].

While secure facilities and data centers are not the topic of this survey, this example highlights the multifaceted nature of the privacy-protection problem. Technical tools for privacy-preserving data publishing are one weapon in a larger arsenal consisting also of legal regulation, more conventional security mechanisms, and the like. In addition, this example highlights a (perceived and sometimes formal) tradeoff between *privacy* and *utility*, a theme that has been repeated throughout the literature and that will be repeated throughout this survey.

1.2 Real-World Attacks and Attack Demonstrations

A number of real-world attacks and demonstrations indicate the importance of taking privacy into consideration when publishing personal data. In this section, our goal is to briefly recap some notable recent events and attacks, which serve to illustrate the challenges in developing privacy-preserving publishing tools.

One published attack on (purportedly) de-identified data was described by Sweeney [241]. The dataset in consideration was collected by the Group Insurance Commission (GIC) and contained medical records of Massachusetts state employees. Since the data did not contain identifiers such as names, social security numbers, addresses, or phone numbers, it was considered safe to give the data to researchers. The data did contain demographic information such as birth date, gender, and zip code. Unfortunately, it is not common for two individuals to have the same birth date, less common for them to also live in the same zip code, and less common still for them to also have the same gender. In fact, according to the Massachusetts voter registration list (available at the time for $20), no one else had the same combination of birth date, gender, and zip code as William Weld, who was then the governor. Thus, his medical records were easy to identify in the data provided by GIC. This sort of attack, where external data are combined with an anonymized data set, is called a *linking attack*.

Not all linking attacks are as simple as performing a join between the GIC data and the voter registration list. This is especially true for text.

As an example, consider the case of AOL. On Sunday, August 6, 2006, AOL released a 2 GB file containing approximately 20 million search queries from 650,000 of its users, which were collected over a period of three months [24]. In addition to the queries themselves, the data set contained information such as which URL from the search results was clicked and what was its ranking. Although the data set was withdrawn within a few hours, it had already been widely downloaded. The anonymization scheme used to protect the data consisted of assigning a random number (pseudonym) to each AOL user and replacing the user id with this number. Three days later, two New York Times reporters [28] found and interviewed user number 4417749 from the data set. They tracked down this user based on the semantic information contained in her search queries: the name of a town, several searches with a particular last name, age-related information, etc. In the case of AOL, there was no single authoritative table (such as a voter list) to link against; instead, there were many scattered sources of information that were used. The privacy breach occurred since AOL failed to reason about these sources and about the semantic content of search queries. We will return to a more detailed discussion of state-of-the-art privacy protection tools for search logs in Section 7.2.

A few months later, Netflix, a movie rental service, announced the Netflix Prize for the development of an accurate movie recommendation algorithm. To aid participants in their research efforts, Netflix also released a data set of 100 million ratings for 18,000 movie titles collected from 480,000 randomly chosen users. Personal information had been removed, and user ids were replaced with pseudonyms, as in the AOL data. This data set contained movie ratings and the dates when the ratings were created [191]. The high-dimensionality of the data set proved to be a tempting target and an attack on such a data set was anticipated by Frankowski et al. [105], who showed that movie ratings can be linked to posts in an online forum. The Netflix data were attacked shortly after it came out by Narayanan and Shmatikov [186], who showed that external information (such as IMDB reviews) can indeed be linked to the Netflix data set using techniques that are commonly known as *record linkage*. Record linkage was first formalized in the 1960s by Fellegi and Sunter [96]; for a survey, see [270]. Record linkage techniques are frequently used to estimate *re-identification probabilities*: the probabilities that users in a data set can be re-identified through auxiliary data [268]. These techniques can often handle varying amounts of noise in the auxiliary data, and are also commonly used for the purpose of data cleaning.

Finally, even further illustrating the vulnerability of public personal data sets, several recent attacks have been demonstrated on (purportedly) de-identified social network graphs. Social networks describe a set of people (nodes) and the relationships between them (edges). As in the cases of search logs and movies, a graph can be considered naively anonymized if all identifying characteristics of the people (e.g., names, etc.) have been removed and replaced with pseudonyms. Interestingly, though by this point perhaps unsurprising, a series of attacks have illustrated the fallacy of this approach. Using data from LiveJournal (a blogging site), Backstrom et al. [26] demonstrated that it is often possible for a particular user to re-identify himself in a social network graph, and with minimal collusion, he can frequently re-identify a large fraction of users. Hay et al. [123] and Narayanan and Shmatikov [187] both took this observation a step further, observing that users can often be re-identified using various forms of structural auxiliary information; these results were demonstrated using a real e-mail graph from Enron Corporation [123] and social network graphs from LiveJournal, Twitter, and Flickr [187]. We will return to an in-depth discussion of the state-of-the-art in privacy protection for social network graphs in Section 7.1. In addition to these examples, attacks on purportedly de-identified data sets have been illustrated in domains as diverse as GPS traces [120, 145] and genomic records [125, 170, 171, 172].

Note that not all attacks need to involve linking. Some involve reconstructing the original data to uncover pieces of information that are considered confidential. One such example was discussed by Meyer and Kadane [177] in relation to the 1990 decennial census. Two important uses of census data are distribution of federal funds and reapportionment (the assignment of seats in the House of Representatives to different states). Thus, undercounting different segments of the population (including minorities) is a serious political issue, and there is a debate about whether to adjust the census data to control for undercounting.

In 1991, the Commerce Department decided not to use the adjusted census data. It also refused to release the adjusted data. Following a congressional subpoena, a compromise was reached and the Commerce Department released adjusted population counts for every other census block and for all blocks whose adjusted population was at least 1,000 [177]. The leaders of the Florida House of Representatives asked Meyer and Kadane to reconstruct these missing values based on the actual census counts and on the released adjusted counts. Later, due to a lawsuit, the rest of the adjusted data was released and Meyer and Kadane were able to evaluate the accuracy of their reconstruction. Using relatively simple techniques based on comparisons of unadjusted counts for various blocks (see [177] for more details), they were able to obtain remarkably accurate results. For the 23 congressional districts of Florida that existed at the time, their estimate of the adjusted population differed from the official adjusted counts by at most 79 people. Meanwhile, the difference between the adjusted and unadjusted counts was on the order of several thousand people. Thus the Commerce Department's naive use of suppression ended up concealing less information than they intended.

Algranati and Kadane [19] discuss another example of data reconstruction. This time it involves the U.S. Department of Justice. In 2000, the U.S. Department of Justice released a report [248] about death penalty statistics for federal crimes. When a federal crime has been committed, the U.S. Attorney in charge of the case must make a recommendation on whether or not to seek the death penalty. The case is also reviewed by the Department of Justice, which also submits a recommendation. Finally, the Attorney General reviews the case and makes the final decision about this process (for more details about the circumstance of the report and the nature of the decisions, see [19, 248]). The Attorney General's decision is made public but the recommendations made by the U.S. Attorney and the Department of Justice are confidential. Algranati and Kadane focused on the 682 cases from 1995 to 2000 that are contained in this report. This report contains eight measured variables: the federal district, defendant's race, victim's race, the crime, whether or not there were multiple victims, and the recommendations made by the U.S. Attorney, the Department of Justice, and the Attorney General. The data were released as a set of lower-dimensional tables of counts. Using some simple combinatorial techniques, Algranati and Kadane were able to fully recover 386 out of 682 records. They were also able to recover the combination of defendant race, federal district and all three recommendations for all of the 682 cases. Again, a naive release of data allowed for the recovery of most of the information that was considered confidential.

All of these examples serve to illustrate the challenges and importance of developing appropriate anonymization measures for published data.

1.3 Running Example

To prevent privacy breaches, organizations that want to publish data must resolve possible privacy issues before releasing data. We introduce privacy issues in data publishing by the following example scenario. A centralized *trusted* data collection agency, say Gotham City Hospital, collects information from a set of patients. The information collected from each patient consists of identifying information like name; demographic information like age, gender, zip code, and nationality; and the patient's medical condition. The data are put into a table like Table 1.1. Researchers in Gotham City University, who study how

Table 1.1. Medical record table.

	Name	Age	Gender	Zip Code	Nationality	Condition
1	Ann	28	F	13053	Russian	Heart disease
2	Bruce	29	M	13068	Chinese	Heart disease
3	Cary	21	F	13068	Japanese	Viral infection
4	Dick	23	M	13053	American	Viral infection
5	Eshwar	50	M	14853	Indian	Cancer
6	Fox	55	M	14750	Japanese	Flu
7	Gary	47	M	14562	Chinese	Heart disease
8	Helen	49	F	14821	Korean	Flu
9	Igor	31	M	13222	American	Cancer
10	Jean	37	F	13227	American	Cancer
11	Ken	36	M	13228	American	Cancer
12	Lewis	35	M	13221	American	Cancer

diseases correlate with patients' demographic attributes, can benefit substantially from analyzing these data and have made a request to the hospital for releasing the table. Now, the question is whether releasing Table 1.1 is safe. In fact, the hospital has a privacy policy that prevents it from releasing patients' identifying information. Obviously, releasing Table 1.1, which contains names, would violate this policy. However, does removal of names from Table 1.1 make the table safe for release? Consider a researcher, say Mark, who is a friend of Eshwar and knows that Eshwar is a 50-year-old Indian male having zip code 14853. He also knows that Eshwar visited Gotham City Hospital several times. If Mark saw this table with names removed, he would be almost sure that his friend Eshwar got cancer, because the fifth record is the only record that matches Mark's knowledge about Eshwar. Age, gender, zip code, and nationality are called *quasi-identifier attributes*, because by looking at these attributes an adversary may potentially identify an individual in the data set.

One way to prevent Mark from being able to infer Eshwar's medical condition is to make sure that, in the released data, no patient can be distinguished from a group of k patients by using age, gender, zip code, and nationality. We call a table that satisfies this criterion a k-anonymous table. Table 1.2 is a modified version of the medical record table that is 4-anonymous, where names have been removed, age values have been generalized to age groups, gender values have been generalized to Any, zip codes have been generalized to first few digits and nationality values have been generalized to different geographical granularities. Now, when Mark sees this generalized table, he only knows that Eshwar's record is in the second group and is not sure whether Eshwar had flu or cancer. However, as will be seen later, this table is still not safe for release.

For now, let us assume that Gotham City Hospital somehow decides to consider 4-anonymous tables to be safe for release; but in addition to Table 1.2, there are many 4-anonymous tables which can be derived from the medical record table. Table 1.3 is another 4-anonymous table derived from the original medical record table. Which one should Gotham City Hospital choose to release? Intuitively, the hospital should choose the one that is the most useful for the researchers who request

Table 1.2. Generalized medical record table.

		Age	Gender	Zip Code	Nationality	Condition
(Ann)	1	20–29	Any	130**	Any	Heart disease
(Bruce)	2	20–29	Any	130**	Any	Heart disease
(Cary)	3	20–29	Any	130**	Any	Viral infection
(Dick)	4	20–29	Any	130**	Any	Viral Infection
(Eshwar)	5	40–59	Any	14***	Asian	Cancer
(Fox)	6	40–59	Any	14***	Asian	Flu
(Gary)	7	40–59	Any	14***	Asian	Heart disease
(Helen)	8	40–59	Any	14***	Asian	Flu
(Igor)	9	30–39	Any	1322*	American	Cancer
(Jean)	10	30–39	Any	1322*	American	Cancer
(Ken)	11	30–39	Any	1322*	American	Cancer
(Lewis)	12	30–39	Any	1322*	American	Cancer

[a] No record can be distinguished from a group of four based on Age, Gender, Zip Code, and nationality.

[b] Names are removed. Age values are generalized to age groups. Gender values are generalized to Any. Zip codes are generalized to first few digits. Nationality values are generalized to different geographical granularities.

Table 1.3. Another generalized medical record table.

		Age	Gender	Zip Code	Nationality	Condition
(Ann)	1	20–59	F	1****	Any	Heart disease
(Helen)	8	20–59	F	1****	Any	Flu
(Cary)	3	20–59	F	1****	Any	Viral infection
(Jean)	10	20–59	F	1****	Any	Cancer
(Eshwar)	5	20–59	M	1****	Asian	Cancer
(Fox)	6	20–59	M	1****	Asian	Flu
(Gary)	7	20–59	M	1****	Asian	Heart disease
(Bruce)	2	20–59	M	1****	Asian	Heart Disease
(Igor)	9	20–39	M	13***	American	Cancer
(Dick)	4	20–39	M	13***	American	Viral infection
(Ken)	11	20–39	M	13***	American	Cancer
(Lewis)	12	20–39	M	13***	American	Cancer

[a] The second record has been swapped with the eighth record, and the fourth record has been swapped with the tenth record.

for the data. Assume that the primary objective of the researchers is to understand how diseases correlated with genders. Thus, the researchers want as little replacement of a gender value by Any as possible. It should be easy to see that Table 1.3 is a better choice than Table 1.2 in terms of the number of replacements of gender values by Any.

1.4 Overview

Given a data set, privacy-preserving data publishing can be intuitively thought of as a game among four parties:

- **Data user**, like the researchers in Gotham City University, who wants to utilize the data.
- **Adversary**, like Mark in the running example, who wants to derive private information from the data.
- **Data publisher**, like Gotham City Hospital, who collects the data and wants to release the data in a way that satisfies the data user's need but also prevents the adversary from obtaining private information about the individuals in the data.
- **Individuals**, like Eshwar, whose data are collected by the data publisher. In some cases, the individuals agree with the data publisher's privacy policy, trust the data publisher and give the data publisher all the requested information. In these cases, it is the data publisher's responsibility to ensure privacy preservation. In other cases, the individuals do not trust the data publisher and want to make sure that the data publisher cannot precisely identify their sensitive information (e.g., by adding noise to their data records so that the data publisher can only have accurate aggregate statistics, but noisy individual data values). Although the primary focus of this paper is on trusted data publishers, we will also discuss untrusted data publishers in Section 4.2.

There is a fundamental tradeoff between privacy and utility. At one extreme, the data publisher may release nothing so that privacy is perfectly preserved; however, no one is able to use the data. At the other extreme, the data publisher may release the data set without any modification so that data utility can be maximized; however, no privacy protection is provided. For the data publisher to release useful data in a way that preserves privacy, the following three components need to be defined.

- **Sanitization mechanism:** Given an *original data set*, e.g., Table 1.1, a sanitization mechanism sanitizes the data set by making the data less precise. This mechanism defines the space of possible "snapshots" of the original data set that are considered as candidates for release. We call such a snapshot a *release candidate*. Generalization is an example sanitization mechanism. Tables 1.2 and 1.3 are two release candidates of such a mechanism when applied to Table 1.1. We will first introduce some common sanitization mechanisms in Section 1.5 and have an in-depth discussion in Section 4.
- **Privacy criterion:** Given a release candidate, the privacy criterion defines whether the release candidate is safe for release or not. k-Anonymity is an example privacy criterion. Privacy criteria are the focus of Section 2.
- **Utility metric:** Given a release candidate, the utility metric quantifies the utility of the release candidate (equivalently, the information loss due to the sanitization process). For example, the researchers in Gotham City University use the number of replacements of gender values by Any as their utility measure. We survey utility metrics in Section 3.

Given the above three components, one approach to privacy-preserving data publishing is to publish the most useful release candidate that satisfies the privacy criterion. An algorithm that takes an original data set and generates a release candidate that satisfies a given privacy criterion while providing high utility[2] is called an *anonymization (or sanitization) algorithm*. The terms "anonymization" and "sanitization" will be used interchangeably. A selected list of interesting anonymization algorithms is presented in Section 4.

After the data publisher finds a good release candidate and makes it public, the data user will use it for good and the adversary will attack it. Because the sanitization mechanism has perturbed the data to make it less precise and less sensitive, the data user may not be able

[2] Note that providing the maximum utility among all release candidates may not be algorithmically feasible and may also be undesirable because it gives an adversary an additional avenue of attack (see Section 6).

to use the data in a straightforward manner. For example, suppose that Table 1.3 is released, and the data user wants to know the fraction of patients with ages between 20 and 30 who have heart disease. This query cannot be answered precisely based on Table 1.3, but may be answered probabilistically. A methodology is needed to answer such queries in a meaningful and consistent manner. In addition to database queries, the data user may also want to build machine-learning models (for a prediction task) or conduct statistical analysis (to test whether a finding from a sanitized data set is statistically significant). We will discuss how to do so in Section 5 and point the readers to related literature.

From the adversary's point of view, although the released data satisfy a privacy criterion (or a few criteria), it is still possible to uncover some individuals' sensitive information. This is because each privacy criterion has its own assumption and sometimes only protects data against a few types of attacks. For example, Table 1.2 satisfies the k-anonymity criterion. However, it is vulnerable to a *homogeneity attack*: although no one cannot distinguish Jean's record from the other three records (Igor's, Ken's, and Lewis') based on the quasi-identifier attributes, we are 100% sure that she has cancer (if we know her quasi-identifier attributes and the fact that her data are in Table 1.2). Furthermore, some anonymization algorithm have special behavior that may allow the adversary to make further inference about the data, and the adversary may have more background knowledge than a privacy criterion assumes. We review interesting attacks against sanitized data in Section 6.

We note that there can potentially be multiple data users with different data needs, multiple adversaries with different purposes and knowledge about individuals in the data, and multiple data publishers (whose data sets may overlap with each other) who would like to release versions of their data. A single data publisher may also want to release different versions of the data at different times. Furthermore, the original data set may not be a single table; it may be a relational database (that contains multiple tables), a market-basket database (in which each record is a set of items), a search log (in which each record is a search query with some metadata), a social network (relating individuals), and so on. These variations all add to the complexity of the problem and will be addressed with different levels of details (in proportion to the progress that has been made on these problems). In particular, we discuss social network privacy in Section 7.1, search log privacy in Section 7.2, location privacy of mobile applications in Section 7.3, and challenges for future research in Section 7.4.

1.5 Examples of Sanitization Mechanisms

Before proceeding to the next chapter, we will first briefly introduce a number of common sanitization mechanisms to facilitate our discussion. It is important to have a basic idea of such mechanisms because a privacy criterion is defined on the output of such a mechanism, an adversary breaches privacy by analyzing such an output, and a data user studies such an output. However, we do not try to cover all of the sanitization mechanisms here. An in-depth discussion of mechanisms and algorithms will be presented in Section 4.

Recall that a sanitization mechanism defines the space of all possible release candidates in an application of privacy-preserving data publishing. An anonymization algorithm finds a release candidate that is both useful and safe (according to a given privacy criterion) from this space. To simplify our discussion, we consider the original data set to be a table (e.g., Table 1.1), in which each column is an attribute and each row is the data record of an individual. Other kinds of data (sets of items, text data, graph and network data, and others) will be discussed later (primarily in Section 7).

Generalization: The generalization mechanism produces a release candidate by generalizing (coarsening) some attribute values in the original table. We have seen two examples of such release candidates in Tables 1.2 and 1.3. The basic idea is that, after generalizing some attribute values, some records (e.g., Ann's record and Bruce's record in Table 1.2) would become identical when projected on the set of quasi-identifier (QI) attributes (e.g., age, gender, zip code, and nationality). Each group of records that have identical QI attribute values is called an *equivalence class*.

Suppression: The suppression mechanism produces a release candidate by replacing some attribute values (or parts of attribute values) by a special symbol that indicates that the value has been suppressed (e.g., "*" or "Any"). Suppression can be thought of as a special kind of generalization. For example, in Table 1.2, we can say that some digits of zip codes and all the gender values have been suppressed.

Swapping: The swapping mechanism produces a release candidate by swapping some attribute values. For example, consider Table 1.1. After removing the names, the data publisher may swap the age values of Ann and Eshwar, swap the gender values of Bruce and Cary, and so on.

Bucketization: The bucketization mechanism produces a release candidate by first partitioning the original data table into non-overlapping groups (or buckets) and then, for each group, releasing its projection on the non-sensitive attributes and also its projection on the sensitive attribute. Table 1.4 is a release candidate of the bucketization mechanism when applied to Table 1.1. In this case the Condition attribute is considered to be sensitive and the other attributes are not. The idea is that after bucketization, the sensitive attribute value of an individual would be indistinguishable from that of any other individual in the same group. Each group is also called an *equivalence class*.

Table 1.4. Bucketized medical record table.

	Age	Gender	Zip Code	Nationality	BID	BID	Condition
(Ann)	28	F	13053	Russian	1	1	Heart disease
(Bruce)	29	M	13068	Chinese	1	1	Heart disease
(Cary)	21	F	13068	Japanese	1	1	Viral infection
(Dick)	23	M	13053	American	1	1	Viral infection
(Eshwar)	50	M	14853	Indian	2	2	Cancer
(Fox)	55	M	14750	Japanese	2	2	Flu
(Gary)	47	M	14562	Chinese	2	2	Heart disease
(Helen)	49	F	14821	Korean	2	2	Flu
(Igor)	31	M	13222	American	3	3	Cancer
(Jean)	37	F	13227	American	3	3	Cancer
(Ken)	36	M	13228	American	3	3	Cancer
(Lewis)	35	M	13221	American	3	3	Cancer

[a] Three buckets are created and identified by their bucket IDs (BID).

[b] A patient's condition in a bucket is indistinguishable from any other patient's condition in the same bucket.

Table 1.5. Randomized medical record table.

		Age	Gender	Zip code	Nationality	Condition
(Ann)	1	30	F	13073	Russian	Heart disease
(Bruce)	2	28	M	13121	American	Heart disease
(Cary)	3	22	M	13024	Japanese	Cancer
(Dick)	4	20	M	13030	American	Viral infection
...		...	...	...	...	...

[a] Names are removed. Random noise is added to each attribute value. For numeric attributes (age and zip code), Gaussian noise is added. For categorical attributes (gender, zip code, and nationality), with some probability, an attribute value is replaced by a random value in the domain.

Randomization: A release candidate of the randomization mechanism is generated by adding random noise to the data. The sanitized data could be sampled from a probability distribution (in which case it is known as *synthetic data*) or the sanitized data could be created by randomly perturbing the attribute values. For example, Table 1.5 is such a release candidate for Table 1.1, where random noise is added to each attribute value. We add Gaussian noise with mean 0 and variance 4 to age and also Gaussian noise with 0 mean and variance 500 to zip code. For gender, nationality, and condition, with probability 1/4, we replace the original attribute value with a random value in the domain; otherwise, we keep the original attribute value. Note that, in general, we may add different amounts of noise to different records and different attributes. Several application scenarios of randomization can be distinguished. In *input randomization*, the data publisher adds random noise to the original data set and releases the resulting randomized data, like Table 1.5. In *output randomization*, data users submit queries to the data publisher and the publisher releases randomized query results. In *local randomization*, individuals (who contribute their data to the data publisher) randomize their own data before giving their data to the publisher. In this last scenario, the data publisher is no longer required to be trusted.

Multi-view release: To increase data utility, the data publisher may release multiple views of a single original data set, where the released views are outputs of one (or more) of the above sanitization mechanisms. For example, a release candidate could be a set of generalized tables. As a special case of multiple generalized tables, we show an

Table 1.6. An example of multi-marginal release.

(a) Marginal on gender, nationality

Gender	Nationality	Count
F	Russian	1
F	Japanese	1
F	Korean	1
F	American	1
M	Chinese	2
M	American	4
M	Indian	1
M	Japanese	1

(b) Marginal on gender, condition

Gender	Condition	Count
F	Heart disease	1
F	Viral infection	1
F	Flu	1
F	Cancer	1
M	Heart disease	2
M	Viral infection	1
M	Flu	1
M	Cancer	4

example of *multi-marginal release* in Table 1.6, which consists of two views of the original data Table 1.1. Each view is generated by projecting the original data table on a subset of attributes and computing the counts. Such a view is called a marginal table or a histogram on the subset of attributes.

References

[1] J. M. Abowd and S. D. Woodcock, "Disclosure limitation in longitudinal linked data," *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, pp. 215–277, 2001.

[2] J. M. Abowd and S. D. Woodcock, "Multiply-imputing confidential characteristics and file links in longitudinal linked data," in *Privacy in Statistical Databases*, 2004.

[3] O. Abul, F. Bonchi, and M. Nanni, "Never walk along: Uncertainty for anonymity in moving objects databases," in *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, 2008.

[4] N. Adam and J. Wortmann, "Security-control methods for statistical databases," *ACM Computing Surveys*, vol. 21, no. 4, pp. 515–556, 1989.

[5] E. Adar, "User 4xxxxx9: Anonymizing query logs," in *Query Log Analysis Workshop at WWW*, 2007.

[6] C. C. Aggarwal, "On k-anonymity and the curse of dimensionality," in *Proceedings of the 31st International Conference on Very Large Databases (VLDB)*, 2005.

[7] C. C. Aggarwal, "On randomization, public information and the curse of dimensionality," in *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, 2007.

[8] C. C. Aggarwal, "On unifying privacy and uncertain data models," in *Proceedings of the 24th International Conference on Data Engineering (ICDE)*, pp. 386–395, 2008.

[9] C. C. Aggarwal, J. Pei, and B. Zhang, "On privacy preservation against adversarial data mining," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2006.

[10] C. C. Aggarwal and P. S. Yu, "A condensation approach to privacy-preserving data mining," in *Proceedings of the 9th International Conference on Extending Database Technology (EDBT)*, 2004.

[11] C. C. Aggarwal and P. S. Yu, "On privacy-preservation of text and sparse binary data with sketches," in *SDM*, 2007.

[12] C. C. Aggarwal and P. S. Yu, *Privacy-Preserving Data Mining: Models and Algorithms*. Springer, 2008.

[13] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu, "Approximation algorithms for k-anonymity," *Journal of Privacy Technology (JOPT)*, 2005.

[14] G. Aggarwal, T. Feder, K. Kenthapadi, R. Panigrahy, D. Thomas, and A. Zhu, "Achieving anonymity via clustering in a metric space," in *Proceedings of the 25th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, 2006.

[15] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," in *Proceedings of the 20th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, 2001.

[16] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2000.

[17] R. Agrawal, R. Srikant, and D. Thomas, "Privacy Preserving OLAP," in *Proceedings of the 23th ACM SIGMOD Conference on Management of Data*, 2004.

[18] S. Agrawal and J. R. Haritsa, "A framework for high-accuracy privacy-preserving mining," in *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, 2005.

[19] D. J. Algranati and J. B. Kadane, "Extracting confidential information from public documents: The 2000 department of justice report on the federal use of the death penalty in the United States," *Journal of Official Statistics*, vol. 20, no. 1, pp. 97–113, 2004.

[20] P. Allison, "Multiple imputation for missing data: A cautionary tale," *Sociological Methods and Research*, vol. 28, no. 3, pp. 301–309, 2000.

[21] N. Alon, Y. Matias, and M. Szegedy, "The space complexity of approximating the frequency moments," in *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing (STOC)*, 1996.

[22] S. Andrews and T. Hofmann, "Multiple-instance learning via disjunctive programming boosting," in *Advances in Neural Information Processing Systems 16 (NIPS)*, (S. Thrun, L. Saul, and B. Schölkopf, eds.), Cambridge, MA: MIT Press, 2004.

[23] S. Arora and B. Barak, *Computational Complexity: A Modern Approach*. Cambridge University Press, 2009.

[24] M. Arrington, "Aol proudly releases massive amounts of private data," TechCrunch: http://www.techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data/, August 6, 2006.

[25] F. Bacchus, A. J. Grove, J. Y. Halpern, and D. Koller, "From statistics to beliefs," in *National Conference on Artificial Intelligence AAAI*, 1992.

[26] L. Backstrom, C. Dwork, and J. Kleinberg, "Wherefore art thou r3579x?: Anonymized social networks, hidden patterns, and structural steganography," in *Proceedings of the 16th International Conference on World Wide Web (WWW)*, 2007.

[27] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar, "Privacy, accuracy and consistency too: A holistic solution to contingency table release," in *Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, 2007.

[28] M. Barbaro and T. Zeller, "A face is exposed for AOL Searcher no. 4417749," *New York Times*, August 9 2006.

[29] R. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymity," in *Proceedings of the 21st International Conference on Data Engineering (ICDE)*, 2005.

[30] R. Benedetti and L. Franconi, "Statistical and technological solutions for controlled data dissemination," in *New Techniques and Technologies for Statistics*, 1998.

[31] A. R. Beresford and F. Stajano, "Location privacy in pervasive computing," *Pervasive Computing*, vol. 2, no. 1, 2003.

[32] C. Bettini, X. Wang, and S. Jajodia, "Protecting privacy against location-based personal identification," in *Proceedings of the VLDB Workshop on Secure Data Management*, 2005.

[33] K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "Nearest Neighbor" meaningful?," in *Proceedings of the 10th International Conference on Database Theory (ICDT)*, pp. 217–235, 1999.

[34] J. Bi and T. Zhang, "Support vector classification with input data uncertainty," in *Advances in Neural Information Processing Systems (NIPS)*, 2004.

[35] Y. M. Bishop, S. E. Fienberg, and P. W. Holland, *Discrete Multivariate Analysis: Theory and Practice.* The MIT Press, 1977.

[36] U. Blien, H. Wirth, and M. Muller, "Disclosure risk for microdata stemming from official statistics," *Statistica Neerlandica*, vol. 46, no. 1, pp. 69–82, 1992.

[37] A. Blum, C. Dwork, F. McSherry, and K. Nissim, "Practical privacy: The SuLQ framework," in *Proceedings of the 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, 2005.

[38] G. E. P. Box and N. R. Draper, *Empirical Model-Building And Response Surface.* Wiley, 1987.

[39] J. Brickell and V. Shmatikov, "The cost of privacy: Destruction of data-mining utility in anonymized data publishing," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008.

[40] S. Bu, L. V. Lakshmanan, R. T. Ng, and G. Ramesh, "Preservation Of patterns and input–output privacy," in *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, 2007.

[41] D. Burdick, P. M. Deshpande, T. S. Jayram, R. Ramakrishnan, and S. Vaithyanathan, "Efficient allocation algorithms for OLAP over imprecise data," in *Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB)*, pp. 391–402, VLDB Endowment, 2006.

[42] D. Burdick, P. M. Deshpande, T. S. Jayram, R. Ramakrishnan, and S. Vaithyanathan, "OLAP over uncertain and imprecise data," *VLDB Journal*, vol. 16, no. 1, pp. 123–144, 2007.

[43] D. Burdick, A. Doan, R. Ramakrishnan, and S. Vaithyanathan, "OLAP over imprecise data with domain constraints," in *Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB)*, pp. 39–50, VLDB Endowment, 2007.

[44] J. Burridge, "Information preserving statistical obfuscation," *Statistics and Computing*, vol. 13, pp. 321–327, 2003.

[45] J.-W. Byun, Y. Sohn, E. Bertino, and N. Li, "Secure anonymization for incremental datasets," in *SIAM Conference on Data Mining (SDM)*, 2006.

[46] S. Canada, "The research data centres (RDC) program," http://www.statcan.gc.ca/rdc-cdr/network-reseau-eng.htm, June 9, 2009.

[47] G. Casella and R. L. Berger, *Statistical Inference*. Duxbury, 2nd ed., 2002.

[48] J. Castro, "Minimum-distance controlled perturbation methods for large-scale tabular data protection," *European Journal of Operational Research*, vol. 171, 2004.

[49] K. Chaudhuri and N. Mishra, "When random sampling preserves privacy," in *Proceedings of the International Cryptology Conference*, 2006.

[50] B.-C. Chen, L. Chen, D. Musicant, and R. Ramakrishnan, "Learning from agggregate views," in *Proceedings of the 22nd International Conference on Data Engineering (ICDE)*, 2006.

[51] B.-C. Chen, K. LeFevre, and R. Ramakrishnan, "PrivacySkyline: Privacy with multidimensional adversarial knowledge," in *Proceedings of the 33rd International Conference on Very Large Databases (VLDB)*, 2007.

[52] C.-Y. Chow and M. Mokbel, "Enabling private continuous queries for revealed user locations," in *Advances in Spatial and Temporal Databases*, 2007.

[53] R. Christensen, *Log-Linear Models and Logistic Regression*. Springer-Verlag, 1997.

[54] C. A. W. Citteur and L. C. R. J. Willenborg, "Public use microdata files: Current practices at national statistical bureaus," *Journal of Official Statistics*, vol. 9, no. 4, pp. 783–794, 1993.

[55] A. Cooper, "A survey of query log privacy-enhancing techniques from a policy perspective," *ACM Transactions on the Web*, vol. 2, pp. 19–27, 2008.

[56] D. A. Cox, J. Little, and D. O'Shea, *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Springer, 3rd ed., 2008.

[57] L. H. Cox, J. P. Kelly, and R. Patil, "Balancing quality and confidentiality for multivariate tabular data," *Domingo-Ferrer J, Torra V, editors, Privacy in Statistical Databases*, 2004.

[58] L. H. Cox, S.-K. McDonald, and D. Nelson, "Confidentiality issues at the United States bureau of the census," *Journal of Official Statistics*, vol. 2, no. 2, pp. 135–160, 1986.

[59] L. H. Cox and L. V. Zayatz, "An Agenda for research in statistical disclosure limitation," *Journal of Official Statistics*, vol. 11, no. 2, pp. 205–220, 1995.

[60] T. Dalenius and S. Reiss, "Data swapping: A technique for disclosure control," *Journal of Statistical Planning and Inference*, vol. 6, pp. 73–85, 1982.

[61] N. N. Dalvi, G. Miklau, and D. Suciu, "Asymptotic conditional probabilities for conjunctive queries," in *Proceedings of the 10th International Conference on Database Theory (ICDT)*, 2005.

[62] N. N. Dalvi and D. Suciu, "Efficient query evaluation on probabilistic databases," *VLDB Journal*, vol. 16, no. 4, pp. 523–544, 2007.

[63] R. A. Dandekar, "Cost effective implementation of synthetic tabulation (a.k.a. controlled tabular adjustments) in legacy and new statistical data publication systems," *Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Luxembourg*, 2003.

[64] R. A. Dandekar, "Maximum utility-minimum information loss table server design for statistical disclosure control of tabular data," *Domingo-Ferrer J, Torra V, editors, Privacy in statistical databases*, 2004.

[65] R. A. Dandekar, M. Cohen, and N. Kirkendall, "Sensitive micro data protection using latin hypercube sampling technique," in *Inference Control in Statistical Databases, From Theory to Practice*, 2002.

[66] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," in *Proceedings of the 12th International Conference on World Wide Web (WWW)*, 2003.

[67] N. de Freitas and H. Kück, "Learning about individuals from group statistics," in *Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence (UAI)*, pp. 332–339, 2005.

[68] A. de Waal and L. C. R. J. Willenborg, "Statistical disclosure control and sampling weights," *Journal of Official Statistics*, vol. 13, no. 4, pp. 417–434, 1997.

[69] D. Defays and M. N. Anwar, "Masking microdata using micro-aggregation," *Journal of Ofcial Statistics*, vol. 14, no. 4, 1998.

[70] D. E. Denning, P. J. Denning, and M. D. Schwartz, "The tracker: A threat to statistical database security," *ACM Transactions on Database Systems*, vol. 4, no. 1, pp. 76–96, 1979.

[71] D. E. Denning and J. Schlörer, "A fast procedure for finding a tracker in a statistical database," *ACM Transactions on Database Systems*, vol. 5, no. 1, pp. 88–102, 1980.

[72] P. Diaconis and B. Sturmfels, "Algebraic algorithms for sampling from conditional distributions," *Annals of Statistics*, vol. 1, pp. 363–397, 1998.

[73] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1–2, pp. 31–71, 1997.

[74] I. Dinur and K. Nissim, "Revealing information while preserving privacy," in *Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, 2003.

[75] A. Dobra, "statistical tools for disclosure limitation in multiway contingency tables," PhD thesis, Carnegie Mellon University, 2002.

[76] A. Dobra, "Markov bases for decomposable graphical models," *Bernoulli*, vol. 9, no. 6, pp. 1093–1108, 2003.

[77] A. Dobra and S. E. Fienberg, *Assessing the Risk of Disclosure of Confidential Categorical Data*. Bayesian Statistics 7, Oxford University Press, 2000.

[78] A. Dobra and S. E. Fienberg, “Bounding entries in multi-way contingency tables given a set of marginal totals,” in *Foundations of Statistical Inference: Proceedings of the Shoresh Conference 2000*, Springer Verlag, 2003.

[79] J. Domingo-Ferrer and J. M. Mateo-Sanz, “Practical data-oriented microaggregation for statistical disclosure control,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 4, no. 1, 2002.

[80] J. Domingo-Ferrer, F. Sebe, and J. Castella-Roca, “On the security of noise addition for privacy in statistical databases,” in *Privacy in Statistical Databases*, 2004.

[81] J. Domingo-Ferrer and V. Torra, “A quantitative comparison of disclosure control methods for microdata,” in *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, 2001.

[82] W. Du, Z. Teng, and Z. Zhu, “Privacy-maxent: Integrating background knowledge in privacy quantification,” in *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data (SIGMOD)*, 2008.

[83] G. T. Duncan and D. Lambert, “The risk of disclosure for microdata,” *Journal of Business and Economic Statistics*, vol. 7, no. 2, 1989.

[84] Q. Duong, K. LeFevre, and M. Wellman, “Strategic modeling of information sharing among data privacy attackers,” in *Quantitative Risk Analysis for Security Applications Workshop*, 2009.

[85] C. Dwork, “Differential privacy,” in *ICALP*, 2006.

[86] C. Dwork, K. Kenthapadi, F. McSherry, and I. Mironov, “Our data, ourselves: Privacy via distributed noise generation,” in *EUROCRYPT*, 2006.

[87] C. Dwork, F. McSherry, K. Nissim, and A. Smith, “Calibrating noise to sensitivity in private data analysis,” in *Theory of Cryptography Conference*, pp. 265–284, 2006.

[88] C. Dwork, F. McSherry, and K. Talwar, “The price of privacy and the limits of LP decoding,” in *Proceedings of the Thirty-ninth Annual ACM Symposium on Theory of Computing (STOC)*, 2007.

[89] J. Edmonds, “Maximum matching and a polyhedron with 0–1 vertices,” *Journal of Research of the National Bureau of Standards, Section B, Mathematical Sciences*, vol. 69, 1965.

[90] M. Elliot and A. Dale, “Scenarios of attack: The data intruder’s perspective on statistical disclosure risk,” *Netherlands Official Statistics*, vol. 14, pp. 6–10, 1999.

[91] A. Evfimievski, R. Fagin, and D. P. Woodruff, “Epistemic privacy,” in *PODS*, 2008.

[92] A. Evfimievski, J. Gehrke, and R. Srikant, “Limiting privacy breaches in privacy-preserving data mining,” in *Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, 2003.

[93] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, “Privacy preserving mining of association rules,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002.

[94] C. Fang and E.-C. Chang, "Information leakage in optimal anonymized and diversified data," in *Information Hiding*, 2008.

[95] Federal Committee on Statistical Methodology. Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology, December 2005.

[96] I. Fellegi and A. Sunter, "A theory for record linkage," *Journal of the American Statistical Association*, vol. 64, no. 328, pp. 1183–1210, 1969.

[97] I. P. Fellegi, "On the question of statistical confidentiality," *Journal of the American Statistical Association*, vol. 67, no. 337, pp. 7–18, 1972.

[98] S. E. Fienberg, "Confidentiality and disclosure limitation methodology: Challenges for national statistics and statistical research," Technical Report 668, Carnegie Mellon University, 1997.

[99] S. E. Fienberg, U. E. Makov, and A. P. Sanil, "A bayesian approach to data disclosure: Optimal intruder behavior for continuous data," *Journal of Official Statistics*, vol. 13, no. 1, pp. 75–89, 1997.

[100] S. E. Fienberg, U. E. Makov, and R. J. Steele, "Disclosure limitation using perturbation and related methods for categorical data," *Journal of Official Statistics*, vol. 14, no. 4, pp. 485–502, 1998.

[101] S. E. Fienberg and J. McIntyre, "Data swapping: Variations on a theme by Dalenius and Reiss," in *Privacy in Statistical Databases*, pp. 14–29, 2004.

[102] S. E. Fienberg and A. B. Slavkovic, "Preserving the confidentiality of categorical statistical data bases when releasing information for association rules," *Data Mining Knowledge Discovery*, vol. 11, no. 2, pp. 155–180, 2005.

[103] L. Franconi and S. Polettini, "Individual risk estimation in μ-argus: A review," in *Privacy in Statistical Databases*, 2004.

[104] L. Franconi and J. Stander, "A model-based method for disclosure limitation of business microdata," *The Statistician*, vol. 51, no. 1, pp. 51–61, 2002.

[105] D. Frankowski, D. Cosley, S. Sen, L. Terveen, and J. Riedl, "You are what you say: Privacy risks of public mentions," in *In Proceedings of the 29th SIGIR*, 2006.

[106] W. A. Fuller, "Masking procedures for microdata disclosure limitation," *Journal of Official Statistics*, vol. 9, no. 2, pp. 383–406, 1993.

[107] B. C. M. Fung, Ke. Wang, and P. S. Yu, "Top-down specialization for information and privacy preservation," in *Proceedings of the 21st International Conference on Data Engineering (ICDE)*, 2005.

[108] S. R. Ganta, S. P. Kasiviswanathan, and A. Smith, "Composition attacks and auxiliary information in data privacy," in *The 14th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2008.

[109] B. Gedik and L. Liu, "Location privacy in mobile systems: A personalized approach," in *Proceedings of the 25th IEEE International Conference on Distributed Computing Systems*, 2005.

[110] B. Gedik and L. Liu, "Protecting location privacy with personalized k-anonymity: Architecture and algorithms," *IEEE Transactions on Mobile Computing*, vol. 7, no. 1, 2008.

[111] G. Ghinita, P. Kalnis, A. Khoshgozaran, C. Shahabi, and K.-L. Tan, "Private queries in location based services: Anonymizers are not necessary," in

Proceedings of the ACM SIGMOD International Conference on Management of Data, 2008.

[112] G. Ghinita, P. Kalnis, and S. Skiadopoulis, "PRIVE: Anonymous location-based queries in distributed mobile systems," in *Proceedings of the 16th International World Wide Web Conference*, 2007.

[113] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, "Fast data anonymization with low information loss," in *Proceedings of the 34th International Conference on Very Large Databases*, 2007.

[114] G. Ghinita, Y. Tao, and P. Kalnis, "On the anonymization of sparse high-dimensional data," *IEEE 24th International Conference on Data Engineering (ICDE)*, 2008.

[115] A. Gionis and T. Tassa, "k-anonymization with minimal loss of information," in *TKDE*, 2008.

[116] S. Gomatam and A. F. Karr, "Distortion measures for categorical data swapping," Technical Report, National Institute of Statistical Sciences, Technical Report Number 131, 2003.

[117] J. Gouweleeuw, P. Kooiman, L. C. R. J. Willenborg, and P.-P. de Wolf, "Post randomisation for statistical disclosure control: Theory and implementation," *Journal of Official Statistics*, vol. 14, no. 4, 1998.

[118] A. J. Grove, J. Y. Halpern, and D. Koller, "Random worlds and maximum entropy," in *Logic in Computer Science*, pp. 22–33, 1992.

[119] M. Gruteser and D. Grunwald, "Anonymous usage of location-based services through spatial and temporal cloaking," in *Proceedings of the First International Conference on Mobile Systems, Applications and Services*, 2003.

[120] M. Gruteser and B. Hoh, "On the anonymity of periodic location samples," in *Proceedings of the Second International Conference on Security in Pervasive Computing*, 2005.

[121] J. Hajek, Z. Sidak, and P. K. Sen, *Theory of Rank Tests.* Academic Press, 2nd ed., 1999.

[122] S. Hawla, L. Zayatz, and S. Rowland, "American factfinder: Disclosure limitation for the advanced query system," *Journal of Official Statistics*, vol. 20, no. 1, pp. 115–124, 2004.

[123] M. Hay, G. Miklau, D. Jensen, D. Towsley, and P. Weis, "Resisting structural re-identification in anonymized social networks," in *Proceedings of the 34th International Conference on Very Large Data Bases (VLDB)*, 2008.

[124] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady, "Preserving privacy in GPS traces via uncertainty-aware path cloaking," in *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*, 2007.

[125] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, Dietrick A. Stephan John V. Pearson, S. F. Nelson, and D. W. Craig, "Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays," *Plos Genetics*, vol. 4, no. 8, 2008.

[126] Z. Huang, W. Du, and B. Chen, "Deriving private information from randomized data," in *Proceedings of the 23th ACM SIGMOD Conference on Management of Data*, June 2004.

[127] A. Hundepool, "The ARGUS software in the CASC-project," in *Privacy in Statistical Databases*, 2004.

[128] A. Hundepool and L. Willenborg, "μ- and τ-ARGUS: Software for statistical disclosure control," in *Proceedings of the Third International Seminar on Statistical Confidentiality*, 1996.

[129] J. T. Hwang, "Multiplicative errors-in-variables models with applications to recent data released by the U.S. department of energy," *Journal of the American Statistical Association*, vol. 81, no. 395, pp. 680–688, 1986.

[130] S. Im, Z. W. Ras, and L.-S. Tsay, "Multi-granularity classification rule discovery using ERID," in *Proceedings of the 3rd International Conference on Rough Sets and Knowledge Technology (RSKT)*, pp. 491–499, 2008.

[131] T. Iwuchukwu and J. Naughton, "K-anonymization as spatial indexing: Toward scalable and incremental anonymization," in *Proceedings of the 33rd International Conference on Very Large Databases (VLDB)*, 2007.

[132] V. S. Iyengar, "Transforming data to satisfy privacy constraints," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002.

[133] T. B. Jabine, "Statistical disclosure limitation practices at United States statistical agencies," *Journal of Official Statistics*, vol. 9, no. 2, pp. 427–454, 1993.

[134] W. Jin, K. LeFevre, and J. Patel, "An Online Framework for Publishing Dynamic Privacy-Sensitive GPS Traces," University of Michigan Technical Report CSE-TR-554-09, 2009.

[135] T. Johnsten and V. V. Raghavan, "Impact of decision-region based classification mining algorithms on database security," in *Proceedings of the IFIP WG 11.3 Thirteenth International Conference on Database Security*, pp. 177–191, Deventer, The Netherlands, The Netherlands, Kluwer, B.V., 2000.

[136] R. Jones, R. Kumar, B. Pang, and A. Tomkins, "I know what you did last summer — query logs and user privacy," in *Proceedings of the 16th Conference on Information and Knowledge Management (CIKM)*, 2007.

[137] P. Kalnis, G. Ghinita, K. Mouratidis, and D. Papadias, "Preventing Location-based identity inference on anonymous spatial queries," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 12, 2007.

[138] H. Kargupta, S. Datta, Qi. Wang, and K. Sivakumar, "On the privacy preserving properties of random data perturbation techniques," in *Proceedings of the International Conference on Data Mining*, 2003.

[139] A. F. Karr, C. N. Kohnen, A. Oganian, J. P. Reiter, and A. P. Sanil, "A framework for evaluating the utility of data altered to protect confidentiality," *The American Statistician*, vol. 60, pp. 224–232, 2006.

[140] A. B. Kennickell, "Multiple imputation and disclosure protection: The case of the 1995 survey of consumer finances," *Record Linkage Techniques*, pp. 248–267, 1997.

[141] D. Kifer, "Attacks on privacy and de Finetti's theorem," in *SIGMOD*, 2009.

[142] D. Kifer and J. Gehrke, "Injecting utility into anonymized datasets," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2006.

[143] J. Kim, "A method for limiting disclosure in microdata based on random noise and transformation," in *Proceedings of the Section on Survey Research Methods*, pp. 303–308, American Statistical Association, 1986.

[144] A. Korolova, K. Kenthapadi, N. Mishra, and A. Ntoulas, "Releasing search queries and clicks privately," in *18th International World Wide Web Conference*, pp. 171–171, 2009.

[145] J. Krumm, "Inference attacks on location tracks," in *Proceedings of the 5th International Conference on Pervasive Computing*, 2007.

[146] R. Kumar, J. Novak, Bo. Pang, and A. Tomkins, "On anonymizing query logs via token-based hashing," in *Proceedings of the 16th International World Wide Web Conference (WWW)*, 2007.

[147] L. V. S. Lakshmanan, R. T. Ng, and G. Ramesh, "To do or not to do: The dilemma of disclosing anonymized data," in *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data (SIGMOD)*, 2005.

[148] D. Lambert, "Measures of disclosure risk and harm," *Journal of Official Statistics*, vol. 9, no. 2, pp. 313–331, 1993.

[149] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2005.

[150] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Mondrian Multidimensional k-Anonymity," in *Proceedings of the 22nd International Conference on Data Engineering (ICDE)*, 2006.

[151] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Workload-Aware Anonymization," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.

[152] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Workload-aware anonymization techniques for large-scale data sets," *ACM Transactions on Database Systems*, vol. 33, no. 3, 2008.

[153] R. Lenz, "Measuring the disclosure protection of micro aggregated business microdata. An analysis taking as an example the german structure of costs survey," *Journal of Official Statistics*, vol. 22, no. 4, pp. 681–710, 2006.

[154] F. Li, J. Sun, S. Papadimitriou, G. A. Mihaila, and I. Stanoi, "Hiding in the crowd: Privacy preservation on evolving streams through correlation tracking," in *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, pp. 686–695, 2007.

[155] J. Li, Y. Tao, and X. Xiao, "Preservation of proximity privacy in publishing numeric sensitive data," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2008.

[156] K. H. Li, T. E. Raghunathan, and D. B. Rubin, "Large-sample significance levels from multiply imputed data using moment-based statistics and an f reference distribution," *Journal of the American Statistical Association*, vol. 86, no. 416, pp. 1065–1073, 1991.

[157] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy beyond k-anonymity and l-diversity," in *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, 2007.

[158] T. Li and N. Li, "Injector: Mining background knowledge for data anonymization," in *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, 2008.

[159] C. K. Liew, U. J. Choi, and C. J. Liew, "A data distortion by probability distribution," *ACM Transactions on Database Systems*, vol. 10, no. 3, pp. 395–411, 1985.

[160] R. J. A. Little, "Statistical analysis of masked data," *Journal of Official Statistics*, vol. 9, no. 2, pp. 407–426, 1993.

[161] F. Liu and R. J. A. Little, "Selective multiple imputation of keys for statistical disclosure control in microdata," in *ASA Proceedings of the Joint Statistical Meetings*, 2002.

[162] K. Liu, C. Giannella, and H. Kargupta, *A Survey of Attack Techniques on Privacy-preserving Data Perturbation Methods.* chapter 15, pp. 357–380, Springer, 2008.

[163] K. Liu and E. Terzi, "Towards identity anonymization on graphs," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2008.

[164] A. Machanavajjhala, "Defining and Enforcing Privacy in Data Sharing," PhD thesis, Cornell University, 2008.

[165] A. Machanavajjhala and J. Gehrke, "On the efficiency of checking perfect privacy," in *Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, 2006.

[166] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "ℓ-Diversity: Privacy beyond k-anonymity," in *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, 2006.

[167] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber, "Privacy: From theory to practice on the map," in *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, 2008.

[168] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "ℓ-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, 2007.

[169] B. Malin, "Betrayed by my shadow: Learning data identity via trail matching," *Journal of Privacy Technology*, p. 20050609001, 2005.

[170] B. Malin, "An evaluation of the current state of genomic data privacy protection and a roadmap for the future," *Journal of the American Medical Informatics Association*, vol. 12, no. 1, 2005.

[171] B. Malin, "Re-identification of familial database records," in *Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium*, 2006.

[172] B. Malin and L. Sweeney, "How (Not) to protect genomic data privacy in a distributed network: Using train re-identification to evaluate and design anonymity protection systems," *Journal of Biomedical Informatics*, vol. 37, no. 3, pp. 179–192, 2004.

[173] D. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Halpern, "Worst case background knowledge for privacy preserving data publishing," in *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, 2007.

[174] J. M. Mateo-Sanz, A. Martínez-Ballesté, and J. Domingo-Ferrer, "Fast generation of accurate synthetic microdata," in *Privacy in Statistical Databases*, 2004.

[175] P. McCullagh and J. A. Nelder, *Generalized Linear Models*. Chapman & Hall/CRC, 2nd ed., 1989.

[176] X.-L. Meng and D. B. Rubin, "Performing likelihood ratio tests with multiply-imputed data sets," *Biometrika*, vol. 79, no. 1, pp. 103–111, 1992.

[177] M. M. Meyer and J. B. Kadane, "Evaluation of a reconstruction of the adjusted 1990 census for Florida," *Journal of Official Statistics*, vol. 13, no. 2, pp. 103–112, 1997.

[178] A. Meyerson and R. Williams, "On the complexity of optimal k-anonymity," in *Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, 2004.

[179] G. Miklau and D. Suciu, "A formal analysis of information disclosure in data exchange," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2004.

[180] M. Mokbel, C. Chow, and W. Aref, "The new casper: Query processing for location services without compromising privacy," in *Proceedings of the 32nd International Conference on Very Large Databases (VLDB)*, 2006.

[181] R. A. Moore, "Controlled data-swapping techniques for masking public use microdata sets," Technical Report, U.S. Bureau of the Census, 1996.

[182] K. Muralidhar and R. Sarathy, "A theoretical basis for perturbation methods," *Statistics and Computing*, vol. 13, no. 4, pp. 329–335, 2003.

[183] K. Muralidhar and R. Sarathy, "A comparison of multiple imputation and data perturbation for masking numerical variables," *Journal of Official Statistics*, vol. 22, no. 3, pp. 507–524, 2006.

[184] M. Murugesan and C. Clifton, "Providing privacy through plausibly deniable search," in *Proceedings of the 9th SIAM International Conference on Data Mining (SDM)*, pp. 768–779, 2009.

[185] D. R. Musicant, J. M. Christensen, and J. F. Olson, "Supervised learning by training on aggregate outputs," in *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM)*, pp. 252–261, 2007.

[186] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *IEEE Symposium on Security and Privacy*, 2008.

[187] A. Narayanan and V. Shmatikov, "De-anonymizing social networks," in *IEEE Symposium on Security and Privacy*, 2009.

[188] M. E. Nergiz, M. Atzori, and C. W. Clifton, "Hiding the presence of individuals from shared databases," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2007.

[189] M. E. Nergiz, M. Atzori, and Y. Saygin, "Toward trajectory anonymization: A generalization-based approach," in *Proceedings of the 2nd SIGSPATIAL ACM GIS International Workshop on Security and Privacy in GIS and LBS*, 2008.

[190] M. E. Nergiz and C. Clifton, "Thoughts on k-anonymization," *Data & Knowledge Engineering*, vol. 63, no. 3, pp. 622–645, 2007.

[191] Netflix. The Netflix Prize Rules: http://www.netflixprize.com//rules.

[192] K. Nissim, S. Raskhodnikova, and A. Smith, "Smooth sensitivity and sampling in private data analysis," in *39th ACM Symposium on Theory of Computing (STOC)*, 2007.

[193] J. Novak, P. Raghavan, and A. Tomkins, "Anti-aliasing on the web," in *Proceedings of the 13th International Conference on World Wide Web (WWW)*, 2004.

[194] A. Ohrn and L. Ohno-Machado, "Using Boolean reasoning to anonymize databases," *Artificial Intelligence in Medicine*, vol. 15, no. 3, pp. 235–254, 1999.

[195] G. Paass, "Disclosure risk and disclosure avoidance for microdata," *Journal of Business & Economic Statistics*, vol. 6, no. 4, pp. 487–500, October 1988.

[196] M. A. Palley and J. S. Simonoff, "The use of regression methodology for the compromise of confidential information in statistical databases," *ACM Transactions on Database Systems*, vol. 12, no. 4, pp. 593–608, 1987.

[197] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing (EMNLP)*, 2002.

[198] S. Papadimitriou, F. Li, G. Kollios, and P. S. Yu, "Time series compressibility and privacy," in *VLDB*, 2007.

[199] J. B. Paris, *The Uncertain Reasoner's Companion.* Cambridge University Press, 1994.

[200] H. Park and K. Shim, "Approximation algorithms for k-anonymity," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2007.

[201] J. Pei, J. Xu, Z. Wang, W. Wang, and Ke. Wang, "Maintaining k-anonymity against incremental updates," in *SSDBM*, 2007.

[202] B. Poblete, M. Spiliopoulou, and R. Baeza-Yates, "Website privacy preservation for query log publishing," in *Proceedings of the 1st ACM SIGKDD International Workshop on Privacy, Security, and Trust in KDD (PinKDD)*, 2007.

[203] S. Polettini, "Maximum entropy simulation for microdata protection," *Statistics and Computing*, vol. 13, pp. 307–320, 2003.

[204] S. Polettini and J. Stander, "A comment on "A theoretical basis for perturbation methods" by Krishnamurty Muralidhar and Rathindra Sarathy," *Statistics and Computing*, vol. 13, no. 4, pp. 337–338, 2003.

[205] N. Quadrianto, A. J. Smola, T. S. Caetano, and Q. V. Le, "Estimating labels from label proportions," in *Proceedings of the 25th International Conference Machine Learning (ICML)*, pp. 776–783, 2008.

[206] T. E. Raghunathan, J. M. Lepkowski, J. V. Hoewyk, and P. Solenberger, "A multivariate technique for multiply imputing missing values using a sequence of regression models," *Survey Methodology*, vol. 27, no. 1, pp. 85–95, 2001.

[207] T. E. Raghunathan, J. P. Reiter, and D. B. Rubin, "Multiple imputation for statistical disclosure limitation," *Journal of Official Statistics*, vol. 19, pp. 1–16, 2003.

[208] G. Ramesh, "Can attackers learn from samples?," in *2nd VLDB Workshop on Secure Data Management (SDM)*, 2005.

[209] J. N. K. Rao, "On variance estimation with imputed survey data," *Journal of the American Statistical Association*, vol. 91, no. 434, pp. 499–506, 1996.

[210] V. Rastogi, M. Hay, G. Miklau, and D. Suciu, "Relationship privacy: Output perturbation for queries with joins," in *PODS*, 2009.

[211] V. Rastogi, D. Suciu, and S. Hong, "The boundary between privacy and utility in data publishing," Technical Report, University of Washington, 2007.

[212] V. Rastogi, D. Suciu, and S. Hong, "The boundary between privacy and utility in data publishing," in *Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB)*, 2007.

[213] J. P. Reiter, "Inference for partially synthetic, public use microdata sets," *Survey Methodology*, vol. 29, no. 2, pp. 181–188, 2003.

[214] J. P. Reiter, "New approaches to data dissemination: A glimpse into the future (?)," *Chance*, vol. 17, no. 3, pp. 12–16, 2004.

[215] J. P. Reiter, "Simultaneous use of multiple imputation for missing data and disclosure limitation," *Survey Methodology*, vol. 30, pp. 235–242, 2004.

[216] J. P. Reiter, "Estimating risks of identification disclosure in microdata," *Journal of the American Statistical Association*, vol. 100, pp. 1103–1113, 2005.

[217] J. P. Reiter, "Significance tests for multi-component estimands from multiply-imputed, synthetic microdata," *Journal of Statistical Planning and Inference*, vol. 131, pp. 365–377, 2005.

[218] J. P. Reiter, "Using cart to generate partially synthetic public use microdata," *Journal of Official Statistics*, vol. 21, no. 3, pp. 441–462, 2005.

[219] Research Data Centres of the Federal Statistical Office and the Statistical Offices of the Lander http://www.forschungsdatenzentrum.de/en/index.asp, June 9, 2009.

[220] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*. New York: Springer-Verlag, 2005.

[221] D. B. Rubin, "Discussion statistical disclosure limitation," *Journal of Official Statistics*, vol. 9, no. 2, 1993.

[222] D. B. Rubin, "Multiple imputation after 18+ years," *Journal of the American Statistical Association*, vol. 91, pp. 473–489, 1996.

[223] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. Wiley-Interscience, 2004.

[224] D. B. Rubin and N. Schenker, "Multiple imputation for interval estimation from simple random samples with ignorable nonresponse," *Journal of the American Statistical Association*, vol. 81, pp. 366–374, 1986.

[225] S. Ruggles Secure Data Laboratories: The U.S. Census Bureau Model.

[226] P. Samarati, "Protecting respondents' identities in microdata release," in *Transactions on Knowledge and Data Engineering*, pp. 1010–1027, 2001.

[227] J. A. Sanchez, J. Urrutia, and E. Ripoll, "Trade-off between disclosure risk and information loss using multivariate microaggregation: A case study on business data," in *Privacy in Statistical Databases*, pp. 307–322, 2004.

[228] J. L. Schafer, "Multiple imputation: A primer," *Statistical Methods in Medical Research*, vol. 8, no. 1, pp. 3–15, 1999.

[229] M. J. Schervish, *Theory of Statistics*. Springer, 1995.

[230] J. Schlorer, "Identification and retrieval of personal records from a statistical bank," in *Methods of Information in Medicine*, 1975.

[231] G. A. F. Seber and A. J. Lee, *Linear Regression Analysis.* John Wiley and Son, 2003.

[232] C. E. Shannon, "Communication theory of secrecy systems," *The Bell System Technical Journal*, 1949.

[233] P. K. Shivaswamy, C. Bhattacharyya, and A. J. Smola, "Second order cone programming approaches for handling missing and uncertain data," *Journal of Machine Learning Research*, vol. 7, pp. 1283–1314, 2006.

[234] C. Skinner, C. Marsh, S. Openshaw, and C. Wymer, "Disclosure control for census microdata," *Journal of Official Statistics*, vol. 10, no. 1, pp. 31–51, 1994.

[235] A. Slavkovic and S. E. Fienberg, "Bounds for cell entries in two-way tables given conditional relative frequencies," in *Privacy in Statistical Databases*, 2004.

[236] N. L. Spruill, "Measures of confidentiality," *Statistics of Income and Related Administrative Record Research*, pp. 131–136, 1982.

[237] K. Stoffel and T. Studer, "Provable data privacy," in *Proceedings of the International Conference on Database and Expert Systems Applications (DEXA)*, pp. 324–332, 2005.

[238] L. Sweeney, "Datafly: A system for providing anonymity in medical data," in *DBSec*, pp. 356–381, 1997.

[239] L. Sweeney, "Guaranteeing anonymity when sharing medical data, the datafly system," *Journal of the American Medical Informatics Association*, 1997.

[240] L. Sweeney, "Uniqueness of simple demographics in the U.S. population," Technical Report, Carnegie Mellon University, 2000.

[241] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

[242] A. Takemura, "Local recoding and record swapping by maximum weight matching for disclosure control of microdata sets," *Journal of Official Statistics*, vol. 18, 2002.

[243] P. Tendick, "Optimal noise addition for preserving confidentiality in multivariate data," *Journal of Statistical Planning and Inference*, vol. 27, no. 3, pp. 341–353, March 1991.

[244] B. J. Tepping, "A model for optimum linkage of records," *Journal of the American Statistical Association*, vol. 63, no. 324, pp. 1321–1332, 1968.

[245] M. Terrovitis and M. Mamoulis, "Privacy preservation in the publication of trajectories," in *Proceedings of the 9th International Conference on Mobile Data Management (MDM)*, 2008.

[246] M. Terrovitis, N. Mamoulis, and P. Kalnis, "Privacy-preserving anonymization of set-valued data," in *Proceedings of the 34th International Conference on Very Large Databases (VLDB)*, 2008.

[247] The Avatar Project at IBM: http://www.almaden.ibm.com/cs/projects/avatar/.

[248] The Federal Death Penalty System A Statistical Survey (1988–2000), http://www.usdoj.gov/dag/pubdoc/dpsurvey.html, September 12, 2000.

[249] The MystiQ Project at University of Washington: http://www.cs.washington.edu/homes/suciu/project-mystiq.html.

[250] The ORION database system at Purdue University: http://orion.cs. purdue.edu/.

[251] The Trio at Stanford University: http://infolab.stanford.edu/trio/.

[252] J.-A. Ting, A. D'Souza, and S. Schaal, "Bayesian regression with input noise for high dimensional data," in *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pp. 937–944, 2006.

[253] V. Torra, "Microaggregation for categorical variables: A median based approach," *Privacy in Statistical Databases*, 2004.

[254] A. Torres, "Contribucions a la Microagregacio per a la Proteccio de Dades Estadistiques (Contributions to the Microaggregation for the Statistical Data Protection," PhD thesis, Universitat Politecnica de Catalunya, 2003.

[255] M. Trottini and S. E. Fienberg, "Modelling user uncertainty for disclosure risk and data utility," *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, no. 5, pp. 511–527, 2002.

[256] U.S. Census Bureau Center for Economic Studies http://www.ces.census.gov, June 9, 2009.

[257] U.S. Census Bureau Statement of Confidentiality http://factfinder.census. gov/jsp/saff/SAFFInfo.jsp?_pageId=su5_confidentiality, Retrieved June 9, 2009.

[258] S. van Buuren and K. Oudshoom, "Flexible multivariate imputation by mice," Technical Report, Netherlands Organization for Applied Scientific Research (TNO) TNO report PG/VGZ/99.054, 1999.

[259] K. Wang, B. Fung, and P. Yu, "Template-based privacy preservation in classification problems," in *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, November 2005.

[260] K. Wang, B. Fung, and P. Yu, "Anonymizing sequential releases," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.

[261] K. Wang, P. S. Yu, and S. Chakraborty, "Bottom-up generalization: A data mining solution to privacy protection," in *Proceedings of the 4th IEEE International Conference on Data Mining (ICDM)*, 2004.

[262] S. L. Warner, "Randomized Response: A survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, 1965.

[263] L. Willenborg and T. de Waal, *Statistical Disclosure Control in Practice*. Springer-Verlag, 1996.

[264] L. Willenborg and T. de Waal, *Elements of Statistical Disclosure Control*. Springer, 2000.

[265] W. E. Winkler, "The state of record linkage and current research problems," Technical Report, Statistical Research Division, U.S. Census Bureau, 1999.

[266] W. E. Winkler, "Single-ranking micro-aggregation and re-identification," Techncial Report, U.S. Bureau of the Census Statistical Research Division, 2002.

[267] W. E. Winkler, "Using simulated annealing for k-anonymity," *Research Report Series (Statistics #2002-7), U. S. Census Bureau*, 2002.

[268] W. E. Winkler, "Masking and re-identification methods for public-use microdata: Overview and research problems," Research Report #2004-06, U.S. Bureau of the Census, 2004.

[269] W. E. Winkler, "Re-identification methods for masked microdata," *Privacy in Statistical Databases*, 2004.

[270] W. E. Winkler, "Overview of record linkage and current research directions," Technical Report, U.S. Bureau of the Census, 2005.

[271] R. Wong, A. Fu, K. Wang, and J. Pei, "Minimality attack in privacy preserving data publishing," in *Proceedings of the 33rd International Conference on Very Large Databases (VLDB)*, 2007.

[272] Workshop on Data Confidentiality http://dcws.stat.cmu.edu/zayatz.htm, September 6–7, 2007.

[273] X. Xiao and Y. Tao, "Anatomy: Simple and effective privacy preservation," in *Proceedings of the 32nd International Conference on Very Large Databases (VLDB)*, 2006.

[274] X. Xiao and Y. Tao, "Personalized Privacy Preservation," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2006.

[275] X. Xiao and Y. Tao, "m-invariance: Towards privacy preserving re-publication of dynamic datasets," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2007.

[276] H. Xiong, M. Steinbach, and V. Kumar, "Privacy leakage in multi-relational databases: A semi-supervised learning perspective," *VLDB Journal*, vol. 15, no. 4, pp. 388–402, 2006.

[277] S. Xu and X. Ye, "Risk & distortion based k-anonymity," *Information Security Applications*, 2008.

[278] C. Yao, X. S. Wang, and S. Jajodia, "Checking for k-anonymity violation by views," in *Proceedings of the 31st International Conference on Very Large Databases (VLDB)*, pp. 910–921, 2005.

[279] J. Zhang and V. Honavar, "Learning decision tree classifiers from attribute value taxonomies and partially specified data," in *Proceedings of International Conference Machine Learning (ICML)*, 2003.

[280] J. Zhang, D.-K. Kang, A. Silvescu, and V. Honavar, "Learning accurate and concise naïve Bayes classifiers from attribute value taxonomies and data," *Knowledge Information Systems*, vol. 9, no. 2, pp. 157–179, 2006.

[281] L. Zhang, S. Jajodia, and A. Brodsky, "Information disclosure under realistic assumptions: Privacy versus optimality," in *Proceedings of the 14th ACM Conference on Computer and Communications Security*, 2007.

[282] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu, "Aggregate query answering on anonymized tables," in *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, 2007.

[283] E. Zheleva and L. Getoor, "Preserving the privacy of sensitive relationships in graph data," in *Proceedings of Privacy, Security and Trust in KDD Workshop*, 2007.

[284] E. Zheleva and L. Getoor, "To join or not to join: The illusion of privacy in social networks with mixed public and private user profiles," in *Proceedings of the World Wide Web Conference*, 2009.

[285] B. Zhou and J. Pei, "Preserving privacy in social networks against neighborhoos attacks," in *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, 2008.