IEEE Access
The journal for rapid open access publishing

Received September 21, 2014, accepted October 4, 2014, date of publication October 9, 2014, date of current version October 20, 2014.

Digital Object Identifier 10.1109/ACCESS.2014.2362522

Information Security in Big Data: Privacy and Data Mining

LEI XU, CHUNXIAO JIANG, (Member, IEEE), JIAN WANG, (Member, IEEE), JIAN YUAN, (Member, IEEE), AND YONG REN, (Member, IEEE)
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

Corresponding author: C. Jiang (chx.jiang@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61371079, Grant 61273214, Grant 61271267, and Grant 91338203, in part by the Research Fund for the Doctoral Program of Higher Education of China under Grant 20110002110060, in part by the National Basic Research Program of China under Grant 2013CB329105, and in part by the Post-Doctoral Science Foundation Project.

ABSTRACT The growing popularity and development of data mining technologies bring serious threat to the security of individual's sensitive information. An emerging research topic in data mining, known as privacy-preserving data mining (PPDM), has been extensively studied in recent years. The basic idea of PPDM is to modify the data in such a way so as to perform data mining algorithms effectively without compromising the security of sensitive information contained in the data. Current studies of PPDM mainly focus on how to reduce the privacy risk brought by data mining operations, while in fact, unwanted disclosure of sensitive information may also happen in the process of data collecting, data publishing, and information (i.e., the data mining results) delivering. In this paper, we view the privacy issues related to data mining from a wider perspective and investigate various approaches that can help to protect sensitive information. In particular, we identify four different types of users involved in data mining applications, namely, data provider, data collector, data miner, and decision maker. For each type of user, we discuss his privacy concerns and the methods that can be adopted to protect sensitive information. We briefly introduce the basics of related research topics, review state-of-the-art approaches, and present some preliminary thoughts on future research directions. Besides exploring the privacy-preserving approaches for each type of user, we also review the game theoretical approaches, which are proposed for analyzing the interactions among different users in a data mining scenario, each of whom has his own valuation on the sensitive information. By differentiating the responsibilities of different users with respect to security of sensitive information, we would like to provide some useful insights into the study of PPDM.

INDEX TERMS Data mining, sensitive information, privacy-preserving data mining, anonymization, provenance, game theory, privacy auction, anti-tracking.

I. INTRODUCTION

Data mining has attracted more and more attention in recent years, probably because of the popularity of the "big data" concept. Data mining is the process of discovering interesting patterns and knowledge from large amounts of data [1]. As a highly application-driven discipline, data mining has been successfully applied to many domains, such as business intelligence, Web search, scientific discovery, digital libraries, etc.

A. THE PROCESS OF KDD

The term "data mining" is often treated as a synonym for another term "knowledge discovery from data" (KDD) which highlights the goal of the mining process. To obtain useful knowledge from data, the following steps are performed in an iterative way (see Fig. 1):

- Step 1: Data preprocessing. Basic operations include data selection (to retrieve data relevant to the KDD task from the database), data cleaning (to remove noise and inconsistent data, to handle the missing data fields, etc.) and data integration (to combine data from multiple sources).
- Step 2: Data transformation. The goal is to transform data into forms appropriate for the mining task, that is, to find useful features to represent the data. Feature selection and feature transformation are basic operations.
- Step 3: Data mining. This is an essential process where intelligent methods are employed to extract data

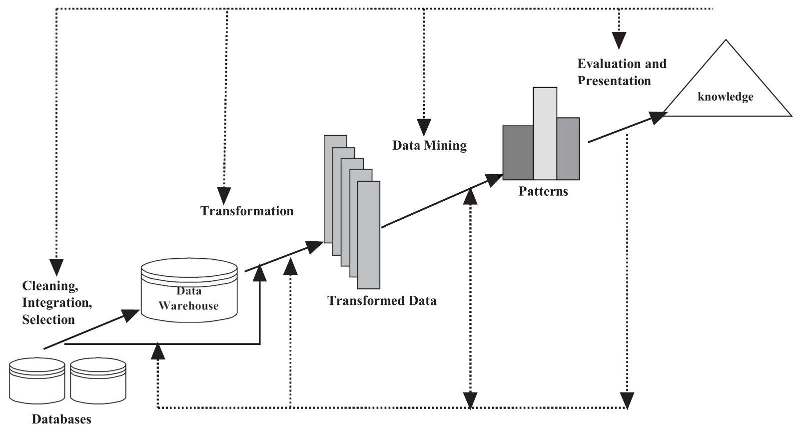


FIGURE 1. An overview of the KDD process.

patterns (e.g. association rules, clusters, classification rules, etc).

- Step 4: Pattern evaluation and presentation. Basic operations include identifying the truly interesting patterns which represent knowledge, and presenting the mined knowledge in an easy-to-understand fashion.

B. THE PRIVACY CONCERN AND PPDM

Despite that the information discovered by data mining can be very valuable to many applications, people have shown increasing concern about the other side of the coin, namely the privacy threats posed by data mining [2]. Individual's privacy may be violated due to the unauthorized access to personal data, the undesired discovery of one's embarrassing information, the use of personal data for purposes other than the one for which data has been collected, etc. For instance, the U.S. retailer Target once received complaints from a customer who was angry that Target sent coupons for baby clothes to his teenager daughter.[1] However, it was true that the daughter was pregnant at that time, and Target correctly inferred the fact by mining its customer data. From this story, we can see that the conflict between data mining and privacy security does exist.

To deal with the privacy issues in data mining, a subfield of data mining, referred to as *privacy preserving data mining* (PPDM) has gained a great development in recent years. The objective of PPDM is to safeguard sensitive information from unsolicited or unsanctioned disclosure, and meanwhile, preserve the utility of the data. The consideration of PPDM is two-fold. First, sensitive raw data, such as individual's ID card number and cell phone number, should not be directly used for mining. Second, sensitive mining results whose disclosure will result in privacy violation should be excluded. After the pioneering work of Agrawal et al. [3], [4], numerous studies on PPDM have been conducted [5]–[7].

[1] *http : //www.forbes.com/sites/kashmirhill/2012/02/16/how − target − figured − out − a − teen − girl − was − pregnant − before − her − father − did/*

C. USER ROLE-BASED METHODOLOGY

Current models and algorithms proposed for PPDM mainly focus on how to hide those sensitive information from certain mining operations. However, as depicted in Fig. 1, the whole KDD process involve multi-phase operations. Besides the mining phase, privacy issues may also arise in the phase of data collecting or data preprocessing, even in the delivery process of the mining results. In this paper, we investigate the privacy aspects of data mining by considering the whole knowledge-discovery process. We present an overview of the many approaches which can help to make proper use of sensitive data and protect the security of sensitive information discovered by data mining. We use the term "sensitive information" to refer to privileged or proprietary information that only certain people are allowed to see and that is therefore not accessible to everyone. If sensitive information is lost or used in any way other than intended, the result can be severe damage to the person or organization to which that information belongs. The term "sensitive data" refers to data from which sensitive information can be extracted. Throughout the paper, we consider the two terms "privacy" and "sensitive information" are interchangeable.

In this paper, we develop a user-role based methodology to conduct the review of related studies. Based on the stage division in KDD process (see Fig. 1), we can identify four different types of users, namely four *user roles*, in a typical data mining scenario (see Fig. 2):

- **Data Provider**: the user who owns some data that are desired by the data mining task.
- **Data Collector**: the user who collects data from data providers and then publish the data to the data miner.
- **Data Miner**: the user who performs data mining tasks on the data.
- **Decision Maker**: the user who makes decisions based on the data mining results in order to achieve certain goals.

In the data mining scenario depicted in Fig. 2, a user represents either a person or an organization. Also, one user can play multiple roles at once. For example, in the Target story we mentioned above, the customer plays the role of data

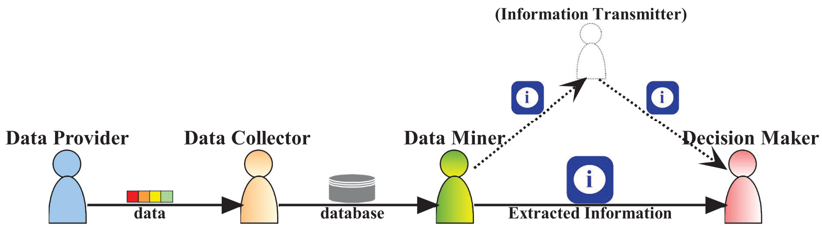


FIGURE 2. A simple illustration of the application scenario with data mining at the core.

provider, and the retailer plays the roles of data collector, data miner and decision maker.

By differentiating the four different user roles, we can explore the privacy issues in data mining in a principled way. All users care about the security of sensitive information, but each user role views the security issue from its own perspective. What we need to do is to identify the privacy problems that each user role is concerned about, and to find appropriate solutions the problems. Here we briefly describe the privacy concerns of each user role. Detailed discussions will be presented in following sections.

1) DATA PROVIDER

The major concern of a data provider is whether he can control the sensitivity of the data he provides to others. On one hand, the provider should be able to make his very private data, namely the data containing information that he does not want anyone else to know, inaccessible to the data collector. On the other hand, if the provider has to provide some data to the data collector, he wants to hide his sensitive information as much as possible and get enough compensations for the possible loss in privacy.

2) DATA COLLECTOR

The data collected from data providers may contain individuals' sensitive information. Directly releasing the data to the data miner will violate data providers' privacy, hence data modification is required. On the other hand, the data should still be useful after modification, otherwise collecting the data will be meaningless. Therefore, the major concern of data collector is to guarantee that the modified data contain no sensitive information but still preserve high utility.

3) DATA MINER

The data miner applies mining algorithms to the data provided by data collector, and he wishes to extract useful information from data in a privacy-preserving manner. As introduced in Section I-B, PPDM covers two types of protections, namely the protection of the sensitive data themselves and the protection of sensitive mining results. With the user role-based methodology proposed in this paper, we consider the data collector should take the major responsibility of protecting sensitive data, while data miner can focus on how to hide the sensitive mining results from untrusted parties.

4) DECISION MAKER

As shown in Fig. 2, a decision maker can get the data mining results directly from the data miner, or from some *Information Transmitter*. It is likely that the information transmitter changes the mining results intentionally or unintentionally, which may cause serious loss to the decision maker. Therefore, what the decision maker concerns is whether the mining results are credible.

In addition to investigate the privacy-protection approaches adopted by each user role, in this paper we emphasize a common type of approach, namely game theoretical approach, that can be applied to many problems involving privacy protection in data mining. The rationality is that, in the data mining scenario, each user pursues high self-interests in terms of privacy preservation or data utility, and the interests of different users are correlated. Hence the interactions among different users can be modeled as a game. By using methodologies from game theory [8], we can get useful implications on how each user role should behavior in an attempt to solve his privacy problems.

D. PAPER ORGANIZATION

The remainder of this paper is organized as follows: Section II to Section V discuss the privacy problems and approaches to these problems for data provider, data collector, data miner and decision maker, respectively. Studies of game theoretical approaches in the context of privacy-preserving data mining are reviewed in Section VI. Some non-technical issues related to sensitive information protection are discussed in Section VII. The paper is concluded in Section IX.

II. DATA PROVIDER

A. CONCERNS OF DATA PROVIDER

A data provider owns some data from which valuable information can be extracted. In the data mining scenario depicted in Fig. 2, there are actually two types of data providers: one refers to the data provider who provides data to data collector, and the other refers to the data collector who provides data to data miner. To differentiate the privacy protecting methods adopted by different user roles, here in this section, we restrict ourselves to the ordinary data provider, the one who owns a relatively small amount of data which contain only information about himself. Data reporting information about an individual are often referred to as "microdata" [9]. If a data provider reveals his microdata to the data collector, his privacy might be comprised due to the unexpected data breach or exposure of sensitive information. Hence, the privacy concern of a data provider is whether he can take control over what kind of and how much information other people can obtain from his data. To investigate the measures that the data provider can adopt to protect privacy, we consider the following three situations:

1) If the data provider considers his data to be very sensitive, that is, the data may reveal some information that he does not want anyone else to know, the provider can just refuse to provide such data. Effective access-control measures are desired by the data provider, so that he can prevent his sensitive data from being stolen by the data collector.
2) Realizing that his data are valuable to the data collector (as well as the data miner), the data provider may be willing to hand over some of his private data in exchange for certain benefit, such as better services or monetary rewards. The data provider needs to know how to negotiate with the data collector, so that he will get enough compensation for any possible loss in privacy.
3) If the data provider can neither prevent the access to his sensitive data nor make a lucrative deal with the data collector, the data provider can distort his data that will be fetched by the data collector, so that his true information cannot be easily disclosed.

B. APPROACHES TO PRIVACY PROTECTION

1) LIMIT THE ACCESS

A data provider provides his data to the collector in an active way or a passive way. By ''active'' we mean that the data provider voluntarily opts in a survey initiated by the data collector, or fill in some registration forms to create an account in a website. By ''passive'' we mean that the data, which are generated by the provider's routine activities, are recorded by the data collector, while the data provider may even have no awareness of the disclosure of his data. When the data provider provides his data actively, he can simply ignore the collector's demand for the information that he deems very sensitive. If his data are passively provided to the data collector, the data provider can take some measures to limit the collector's access to his sensitive data.

Suppose that the data provider is an Internet user who is afraid that his online activities may expose his privacy. To protect privacy, the user can try to erase the traces of his online activities by emptying browser's cache, deleting cookies, clearing usage records of applications, etc. Also, the provider can utilize various security tools that are developed for Internet environment to protect his data. Many of the security tools are designed as browser extensions for ease of use. Based on their basic functions, current security tools can be categorized into the following three types:

1) Anti-tracking extensions. Knowing that valuable information can be extracted from the data produced by users' online activities, Internet companies have a strong motivation to track the users' movements on the Internet. When browsing the Internet, a user can utilize an anti-tracking extension to block the trackers from collecting the cookies.[2] Popular anti-tracking extensions include Disconnect,[3] Do Not Track Me,[4] Ghostery,[5] etc. A major technology used for anti-tracking is called Do Not Track (DNT) [10], which enables users to opt out of tracking by websites they do not visit. A user's opt-out preference is signaled by an HTTP header field named *DNT*: if DNT=1, it means the user does not want to be tracked (opt out). Two U.S. researchers first created a prototype add-on supporting DNT header for the Firefox web browser in 2009. Later, many web browsers have added support for DNT. DNT is not only a technology but also a policy framework for how companies that receive the signal should respond. The W3C Tracking Protection Working Group [11] is now trying to standardize how websites should response to user's DNT request.
2) Advertisement and script blockers. This type of browser extensions can block advertisements on the sites, and kill scripts and widgets that send the user's data to some unknown third party. Example tools include AdBlock Plus,[6] NoScript,[7] FlashBlock,[8] etc.
3) Encryption tools. To make sure a private online communication between two parties cannot be intercepted by third parties, a user can utilize encryption tools, such as MailCloak[9] and TorChat,[10] to encrypt his emails, instant messages, or other types of web traffic. Also, a user can encrypt all of his internet traffic by using a VPN (virtual private network)[11] service.

In addition to the tools mentioned above, an Internet user should always use anti-virus and anti-malware tools to protect his data that are stored in digital equipment such as personal computer, cell phone and tablet. With the help of all these security tools, the data provider can limit other's access to his personal data. Though there is no guarantee that one's sensitive data can be completely kept out of the reach of untrustworthy data collectors, making it a habit of clearing online traces and using security tools does can help to reduce the risk of privacy disclosure.

2) TRADE PRIVACY FOR BENEFIT

In some cases, the data provider needs to make a trade-off between the loss of privacy and the benefits brought by

[2] *http://en.wikipedia.org/wiki/HTTP_cookie*
[3] *https://disconnect.me/*
[4] *https://www.abine.com/index.php*
[5] *https://www.ghostery.com/*
[6] *https://adblockplus.org/en/chrome*
[7] *http://noscript.net/*
[8] *http://flashblock.mozdev.org/*
[9] *http://www.gwebs.com/mailcloak.html*
[10] *http://code.google.com/p/torchat/*
[11] *http://en.wikipedia.org/wiki/Virtual_private_network*

participating in data mining. For example, by analyzing a user's demographic information and browsing history, a shopping website can offer personalized product recommendations to the user. The user's sensitive preference may be disclosed but he can enjoy a better shopping experience. Driven by some benefits, e.g. a personalized service or monetary incentives, the data provider may be willing to provide his sensitive data to a trustworthy data collector, who promises the provider's sensitive information will not be revealed to an unauthorized third-party. If the provider is able to predict how much benefit he can get, he can rationally decide what kind of and how many sensitive data to provide. For example, suppose a data collector asks the data provider to provide information about his age, gender, occupation and annual salary. And the data collector tells the data provider how much he would pay for each data item. If the data provider considers salary to be his sensitive information, then based on the prices offered by the collector, he chooses one of the following actions: i) not to report his salary, if he thinks the price is too low; ii) to report a fuzzy value of his salary, e.g. "less than 10,000 dollars", if he thinks the price is just acceptable; iii) to report an accurate value of his salary, if he thinks the price is high enough. For this example we can see that, both the privacy preference of data provider and the incentives offered by data collector will affect the data provider's decision on his sensitive data. On the other hand, the data collector can make profit from the data collected from data providers, and the profit heavily depends on the quantity and quality of the data. Hence, data providers' privacy preferences have great influence on data collector's profit. The profit plays an important role when data collector decides the incentives. That is to say, data collector's decision on incentives is related to data providers' privacy preferences. Therefore, if the data provider wants to obtain satisfying benefits by "selling" his data to the data collector, he needs to consider the effect of his decision on data collector's benefits (even the data miner's benefits), which will in turn affects the benefits he can get from the collector. In the data-selling scenario, both the seller (i.e. the data provider) and the buyer (i.e. the data collector) want to get more benefits, thus the interaction between data provider and data collector can be formally analyzed by using game theory [12]. Also, the sale of data can be treated as an auction, where mechanism design [13] theory can be applied. Considering that different user roles are involved in the sale, and the privacy-preserving methods adopted by data collector and data miner may have influence on data provider's decisions, we will review the applications of game theory and mechanism design in Section VI, after the discussions of other user roles.

3) PROVIDE FALSE DATA

As discussed above, a data provider can take some measures to prevent data collector from accessing his sensitive data. However, a disappointed fact that we have to admit is that no matter how hard they try, Internet users cannot completely stop the unwanted access to their personal information. So instead of trying to limit the access, the data provider can provide false information to those untrustworthy data collectors. The following three methods can help an Internet user to falsify his data:

1) Using "sockpuppets" to hide one's true activities. A sockpuppet[12] is a false online identity though which a member of an Internet community speaks while pretending to be another person, like a puppeteer manipulating a hand puppet. By using multiple sockpuppets, the data produced by one individual's activities will be deemed as data belonging to different individuals, assuming that the data collector does not have enough knowledge to relate different sockpuppets to one specific individual. As a result, the user's true activities are unknown to others and his sensitive information (e.g. political preference) cannot be easily discovered.
2) Using a fake identity to create phony information. In 2012, Apple Inc. was assigned a patient called "Techniques to pollute electronic profiling" [14] which can help to protect user's privacy. This patent discloses a method for polluting the information gathered by "network eavesdroppers" by making a false online identity of a principal agent, e.g. a service subscriber. The clone identity automatically carries out numerous online actions which are quite different from a user's true activities. When a network eavesdropper collects the data of a user who is utilizing this method, the eavesdropper will be interfered by the massive data created by the clone identity. Real information about of the user is buried under the manufactured phony information.
3) Using security tools to mask one's identity. When a user signs up for a web service or buys something online, he is often asked to provide information such as email address, credit card number, phone number, etc. A browser extension called MaskMe,[13] which was release by the online privacy company Abine, Inc. in 2013, can help the user to create and manage aliases (or *Masks*) of these personal information. Users can use these aliases just like they normally do when such information is required, while the websites cannot get the real information. In this way, user's privacy is protected.

C. SUMMARY

Once the data have been handed over to others, there is no guarantee that the provider's sensitive information will be safe. So it is important for data provider to make sure his sensitive data are out of reach for anyone untrustworthy at the beginning. The DNT technology seems to be a good solution to privacy problems, considering that it helps users to regain the control over "who sees what you are doing online". However, DNT cannot guarantee the safety of users' privacy, since all DNT does is making a request to the Web server,

[12] *http://en.wikipedia.org/wiki/Sockpuppet_(Internet)*
[13] *https://www.abine.com/maskme/*

saying that "please do not collect and store information about me". There is no compulsion for the server to look for the DNT header and honor the DNT request. Practical anti-tracking methods which are less dependent on data collectors' honesty are in urgent need.

In principle, the data provider can realize a perfect protection of his privacy by revealing no sensitive data to others, but this may kill the functionality of data mining. In order to enjoy the benefits brought by data mining, sometimes the data provider has to reveal some of his sensitive data. A clever data provider should know how to negotiate with the data collector in order to make every piece of the revealed sensitive information worth. Current mechanisms proposed for sensitive data auction usually incentivize the data providers to report their truthful valuation on privacy. However, from the point of view of data providers, mechanisms which allow them to put higher values on their privacy are desired, since the data providers always want to gain more benefits with less disclosure of sensitive information.

Another problem needs to be highlighted in future research is how the data provider can discover the unwanted disclosure of his sensitive information as early as possible. Studies in computer security and network security have developed various kinds of techniques for detecting attacks, intrusions and other types of security threats. However, in the context of data mining, the data provider usually has no awareness of how his data are used. Lacking of ways to monitor the behaviors of data collector and data miner, data providers learn about the invasion of their privacy mainly from media exposure. The U.S. telecommunications company, Verizon Communications Inc., has release a series of investigation reports on data breach since 2008. According to its 2013 report [15], about 62% of data breach incidents take months or even years to be discovered, and nearly 70% of the breaches are discovered by someone other than the data owners. This depressing statistic reminds us that it is in urgent need to develop effective methodologies to help ordinary user find misbehavior of data collectors and data miners in time.

III. DATA COLLECTOR

A. CONCERNS OF DATA COLLECTOR

As shown in Fig. 2, a data collector collects data from data providers in order to support the subsequent data mining operations. The original data collected from data providers usually contain sensitive information about individuals. If the data collector doesn't take sufficient precautions before releasing the data to public or data miners, those sensitive information may be disclosed, even though this is not the collector's original intention. For example, on October 2, 2006, the U.S. online movie rental service Netflix[14] released a data set containing movie ratings of 500,000 subscribers to the public for a challenging competition called "the Netflix Prize". The goal of the competition was to improve the accuracy of personalized movie recommendations. The released data set was supposed to be privacy-safe, since each data record only contained a subscriber ID (irrelevant with the subscriber's real identity), the movie info, the rating, and the date on which the subscriber rated the movie. However, soon after the release, two researchers [16] from University of Texas found that with a little bit of auxiliary information about an individual subscriber, e.g. 8 movie ratings (of which 2 may be completely wrong) and dates that may have a 14-day error, an adversary can easily identify the individual's record (if the record is present in the data set).

[14] *http://en.wikipedia.org/wiki/Netflix*

From above example we can see that, it is necessary for the data collector to modify the original data before releasing them to others, so that sensitive information about data providers can neither be found in the modified data nor be inferred by anyone with malicious intent. Generally, the modification will cause a loss in data utility. The data collector should also make sure that sufficient utility of the data can be retained after the modification, otherwise collecting the data will be a wasted effort. The data modification process adopted by data collector, with the goal of preserving privacy and utility simultaneously, is usually called *privacy preserving data publishing* (PPDP).

Extensive approaches to PPDP have been proposed in last decade. Fung et al. have systematically summarized and evaluated different approaches in their frequently cited survey [17]. Also, Wong and Fu have made a detailed review of studies on PPDP in their monograph [18]. To differentiate with their work, in this paper we mainly focus on how PPDP is realized in two emerging applications, namely social networks and location-based services. To make our review more self-contained, in next subsection we will first briefly introduce some basics of PPDP, e.g. the privacy model, typical anonymization operations, information metrics, etc, and then we will review studies on social networks and location-based services respectively.

B. APPROACHES TO PRIVACY PROTECTION

1) BASICS OF PPDP

PPDP mainly studies anonymization approaches for publishing useful data while preserving privacy. The original data is assumed to be a private table consisting of multiple records. Each record consists of the following 4 types of attributes:

- Identifier (ID): Attributes that can directly and uniquely identify an individual, such as name, ID number and mobile number.
- Quasi-identifier (QID): Attributes that can be linked with external data to re-identify individual records, such as gender, age and zip code.
- Sensitive Attribute (SA): Attributes that an individual wants to conceal, such as disease and salary.
- Non-sensitive Attribute (NSA): Attributes other than ID, QID and SA.

Before being published to others, the table is anonymized, that is, identifiers are removed and quasi-identifiers are modified. As a result, individual's identity and sensitive attribute values can be hidden from adversaries.

How the data table should be anonymized mainly depends on how much privacy we want to preserve in the anonymized data. Different privacy models have been proposed to quantify the preservation of privacy. Based on the attack model which describes the ability of the adversary in terms of identifying a target individual, privacy models can be roughly classified into two categories. The first category considers that the adversary is able to identify the record of a target individual by linking the record to data from other sources, such as liking the record to a record in a published data table (called *record linkage*), to a sensitive attribute in a published data table (called *attribute linkage*), or to the published data table itself (called *table linkage*). The second category considers that the adversary has enough background knowledge to carry out a *probabilistic attack*, that is, the adversary is able to make a confident inference about whether the target's record exist in the table or which value the target's sensitive attribute would take. Typical privacy models [17] includes k-anonymity (for preventing record linkage), l-diversity (for preventing record linkage and attribute linkage), t-closeness (for preventing attribute linkage and probabilistic attack), *epsilon*-differential privacy (for preventing table linkage and probabilistic attack), etc.

Age	Sex	Zipcode	Disease
5	Female	12000	HIV
9	Male	14000	dyspepsia
6	Male	18000	dyspepsia
8	Male	19000	bronchitis
12	Female	21000	HIV
15	Female	22000	cancer
17	Female	26000	pneumonia
19	Male	27000	gastritis
21	Female	33000	flu
24	Female	37000	pneumonia

(a)

Age	Sex	Zipcode	Disease
[1, 10]	People	1****	HIV
[1, 10]	People	1****	dyspepsia
[1, 10]	People	1****	dyspepsia
[1, 10]	People	1****	bronchitis
[11, 20]	People	2****	HIV
[11, 20]	People	2****	cancer
[11, 20]	People	2****	pneumonia
[11, 20]	People	2****	gastritis
[21, 60]	People	3****	flu
[21, 60]	People	3****	pneumonia

(b)

FIGURE 3. An example of 2-anonymity, where QID= $\{Age, Sex, Zipcode\}$. (a) Original table. (b) 2-anonymous table.

Among the many privacy models, k-anonymity and its variants are most widely used. The idea of k-anonymity is to modify the values of quasi-identifiers in original data table, so that every tuple in the anonymized table is indistinguishable from at least $k-1$ other tuples along the quasi-identifiers. The anonymized table is called a k-anonymous table. Fig. 3 shows an example of 2-anonymity. Intuitionally, if a table satisfies k-anonymity and the adversary only knows the quasi-identifier values of the target individual, then the probability that the target's record being identified by the adversary will not exceed $1/k$.

To make the data table satisfy the requirement of a specified privacy model, one can apply the following anonymization operations [17]:

- Generalization. This operation replaces some values with a parent value in the taxonomy of an attribute. Typical generalization schemes including full-domain generalization, subtree generalization, multidimensional generalization, etc.
- Suppression. This operation replaces some values with a special value (e.g. a asterisk '*'), indicating that the replaced values are not disclosed. Typical suppression schemes include record suppression, value suppression, cell suppression, etc.
- Anatomization. This operation does not modify the quasi-identifier or the sensitive attribute, but de-associates the relationship between the two. Anatomization-based method releases the data on QID and the data on SA in two separate tables.
- Permutation. This operation de-associates the relationship between a quasi-identifier and a numerical sensitive attribute by partitioning a set of data records into groups and shuffling their sensitive values within each group.
- Perturbation. This operation replaces the original data values with some synthetic data values, so that the statistical information computed from the perturbed data does not differ significantly from the statistical information computed from the original data. Typical perturbation methods include adding noise, swapping data, and generating synthetic data.

The anonymization operations will reduce the utility of data. The reduction of data utility is usually represented by *information loss*: higher information loss means lower utility of the anonymized data. Various metrics for measuring information loss have been proposed, such as minimal distortion [19], discernibility metric [20], the normalized average equivalence class size metric [21], weighted certainty penalty [22], information-theoretic metrics [23], etc. A fundamental problem of PPDP is how to make a tradeoff between privacy and utility. Given the metrics of privacy preservation and information loss, current PPDP algorithms usually take a greedy approach to achieve a proper tradeoff: multiple tables, all of which satisfy the requirement of the specified privacy model, are generated during the anonymization process, and the algorithm outputs the one that minimizes the information loss.

2) PRIVACY-PRESERVING PUBLISHING OF SOCIAL NETWORK DATA

Social networks have gained great development in recent years. Aiming at discovering interesting social patterns, social network analysis becomes more and more important. To support the analysis, the company who runs a social network application sometimes needs to publish its data to a third party. However, even if the truthful identifiers of individuals are removed from the published data, which is referred to as naïve anonymized, publication of the network data may lead to exposures of sensitive information about individuals, such as one's intimate relationships with others. Therefore, the network data need to be properly anonymized before they are published.

A social network is usually modeled as a graph, where the vertex represents an entity and the edge represents the relationship between two entities. Thus, PPDP in the context of social networks mainly deals with anonymizing graph data, which is much more challenging than anonymizing relational table data. Zhou et al. [24] have identified the following three challenges in social network data anonymization:

First, modeling adversary's background knowledge about the network is much harder. For relational data tables, a small set of quasi-identifiers are used to define the attack models. While given the network data, various information, such as attributes of an entity and relationships between different entities, may be utilized by the adversary.

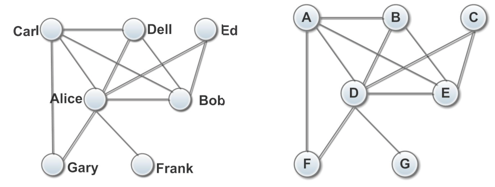
Second, measuring the information loss in anonymizing social network data is harder than that in anonymizing relational data. It is difficult to determine whether the original network and the anonymized network are different in certain properties of the network.

Third, devising anonymization methods for social network data is much harder than that for relational data. Anonymizing a group of tuples in a relational table does not affect other tuples. However, when modifying a network, changing one vertex or edge may affect the rest of the network. Therefore, "divide-and-conquer" methods, which are widely applied to relational data, cannot be applied to network data.

To deal with above challenges, many approaches have been proposed. According to [25], anonymization methods on simple graphs, where vertices are not associated with attributes and edges have no labels, can be classified into three categories, namely edge modification, edge randomization, and clustering-based generalization. Comprehensive surveys of approaches to on social network data anonymization can be found in [18], [25], and [26]. In this paper, we briefly review some of the very recent studies, with focus on the following three aspects: attack model, privacy model, and data utility.

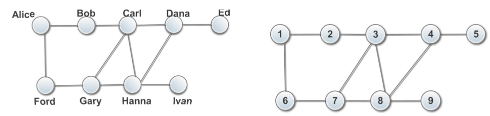
3) ATTACK MODEL

Given the anonymized network data, adversaries usually rely on background knowledge to de-anonymize individuals and learn relationships between de-anonymized individuals. Zhou et al. [24] identify six types of the background knowledge, i.e. attributes of vertices, vertex degrees, link relationship, neighborhoods, embedded subgraphs and graph metrics. Peng et al. [27] propose an algorithm called *Seed-and-Grow* to identify users from an anonymized social graph, based solely on graph structure. The algorithm first identifies a seed sub-graph which is either planted by an attacker or divulged by collusion of a small group of users, and then grows the seed larger based on the adversary's existing knowledge of users' social relations. Zhu et al. [28] design a *structural attack* to de-anonymize social graph data. The attack uses the cumulative degree of



(a) (b)

FIGURE 4. Example of mutual friend attack: (a) original network; (b) naïve anonymized network.



(a) (b)

FIGURE 5. Example of friend attack: (a) original network; (b) naïve anonymized network.

n-hop neighbors of a vertex as the regional feature, and combines it with the simulated annealing-based graph matching method to re-identify vertices in anonymous social graphs. Sun et al. [29] introduce a relationship attack model called *mutual friend attack*, which is based on the number of mutual friends of two connected individuals. Fig. 4 shows an example of the mutual friend attack. The original social network G with vertex identities is shown in Fig. 4(a), and Fig. 4(b) shows the corresponding anonymized network where all individuals' names are removed. In this network, only Alice and Bob have 4 mutual friends. If an adversary knows this information, then he can uniquely re-identify the edge (D, E) in Fig. 4(b) is $(Alice, Bob)$. In [30], Tai et al. investigate the *friendship attack* where an adversary utilizes the degrees of two vertices connected by an edge to re-identify related victims in a published social network data set. Fig. 5 shows an example of friendship attack. Suppose that each user's friend count (i.e. the degree of the vertex) is publicly available. If the adversary knows that Bob has 2 friends and Carl has 4 friends, and he also knows that Bob and Carl are friends, then he can uniquely identify that the edge (2, 3) in Fig. 5(b) corresponds to $(Bob, Carl)$. In [31], another type of attack, namely *degree attack*, is explored. The motivation is that each individual in a social network is inclined to associated with not only a vertex identity but also a community identity, and the community identity reflects some sensitive information about the individual. It has been shown that, based on some background knowledge about vertex degree, even if the adversary cannot precisely identify the vertex corresponding to an individual, community information and neighborhood information can still be inferred. For example, the network shown in Fig. 6 consists of two communities, and the community identity reveals sensitive information (i.e. disease status) about its members. Suppose that an adversary knows Jhon has 5 friends, then he can infer that Jhon has AIDS, even though he is not sure which of the two vertices

IEEE Access
The journal for rapid open access publishing

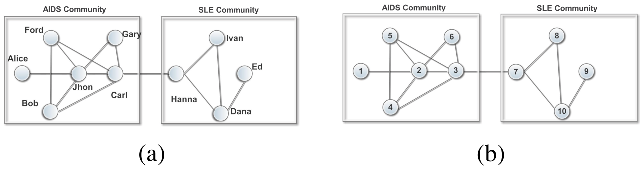


FIGURE 6. **Example of degree attack: (a) original network; (b) naïve anonymized network.**

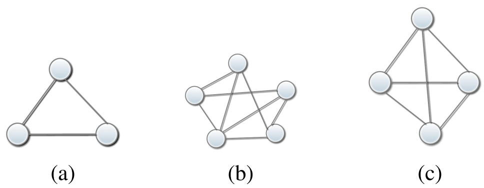


FIGURE 7. **Examples of *k-NMF* anonymity: (a) 3-NMF; (b) 4-NMF; (c) 6-NMF.**

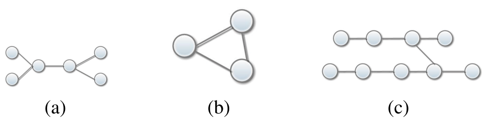


FIGURE 8. **Examples of k^2-degree anonymous graphs: (a) 2^2-degree; (b) 3^2-degree; (c) 2^2-degree.**

(vertex 2 and vertex 3) in the anonymized network (Fig. 6(b)) corresponds to Jhon. From above discussion we can see that, the graph data contain rich information that can be explored by the adversary to initiate an attack. Modeling the background knowledge of the adversary is difficult yet very important for deriving the privacy models.

a: PRIVACY MODEL

Based on the classic k-anonymity model, a number of privacy models have been proposed for graph data. Some of the models have been summarized in the survey [32], such as k-degree,k-neighborhood, k-automorphism, k-isomorphism, and k-symmetry. In order to protect the privacy of relationship from the mutual friend attack, Sun et al. [29] introduce a variant of k-anonymity, called *k-NMF* anonymity. *NMF* is a property defined for the edge in an undirected simple graph, representing the number of mutual friends between the two individuals linked by the edge. If a network satisfies *k-NMF* anonymity (see Fig. 7), then for each edge e, there will be at least $k - 1$ other edges with the same number of mutual friends as e. It can be guaranteed that the probability of an edge being identified is not greater than $1/k$. Tai et al. [30] introduce the concept of k^2-degree anonymity to prevent friendship attacks. A graph $\bar{G}$ is k^2-degree anonymous if, for every vertex with an incident edge of degree pair (d_1, d_2) in $\bar{G}$, there exist at least $k - 1$ other vertices, such that each of the $k - 1$ vertices also has an incident edge of the same degree pair (see Fig. 8). Intuitively, if a graph is k^2-degree anonymous, then the probability of a vertex being re-identified is not greater than $1/k$, even if an adversary knows a certain degree pair (d_A, d_B), where

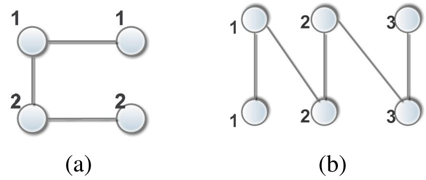


FIGURE 9. **Examples of 2-structurally diverse graphs, where the community ID is indicated beside each vertex.**

A and B are friends. To prevent degree attacks, Tai et al. [31] introduce the concept of *structural diversity*. A graph satisfies *k-structural diversity anonymization* (*k-SDA*), if for every vertex v in the graph, there are at least k communities, such that each of the communities contains at least one vertex with the same degree as v (see Fig. 9). In other words, for each vertex v, there are at least $k - 1$ other vertices located in at least $k - 1$ other communities.

b: DATA UTILITY

In the context of network data anonymization, the implication of data utility is: whether and to what extent properties of the graph are preserved. Wu et al. [25] summarize three types of properties considered in current studies. The first type is graph topological properties, which are defined for applications aiming at analyzing graph properties. Various measures have been proposed to indicate the structure characteristics of the network. The second type is graph spectral properties. The spectrum of a graph is usually defined as the set of eigenvalues of the graph's adjacency matrix or other derived matrices, which has close relations with many graph characteristics. The third type is aggregate network queries. An aggregate network query calculates the aggregate on some paths or subgraphs satisfying some query conditions. The accuracy of answering aggregate network queries can be considered as the measure of utility preservation. Most existing k-anonymization algorithms for network data publishing perform edge insertion and/or deletion operations, and they try to reduce the utility loss by minimizing the changes on the graph degree sequence. Wang et al. [33] consider that the degree sequence only captures limited structural properties of the graph and the derived anonymization methods may cause large utility loss. They propose utility loss measurements built on the community-based graph models, including both the flat community model and the hierarchical community model, to better capture the impact of anonymization on network topology.

One important characteristic of social networks is that they keep evolving over time. Sometimes the data collector needs to publish the network data periodically. The privacy issue in sequential publishing of dynamic social network data has recently attracted researchers' attention. Medforth and Wang [34] identify a new class of privacy attack, named *degree-trail attack*, arising from publishing a sequence of graph data. They demonstrate that even if each published graph is anonymized by strong privacy preserving techniques, an adversary with little background

knowledge can re-identify the vertex belonging to a known target
individual by comparing the degrees of vertices in the published graphs with the degree evolution of a target. In [35], Tai et al. adopt the same attack model used in [34], and propose a privacy model called dynamic k^w-*structural diversity anonymity* (k^w-*SDA*), for protecting the vertex and multi-community identities in sequential releases of a dynamic network. The parameter k has a similar implication as in the original k-anonymity model, and w denotes a time period that an adversary can monitor a target to collect the attack knowledge. They develop a heuristic algorithm for generating releases satisfying this privacy requirement.

4) PRIVACY-PRESERVING PUBLISHING OF TRAJECTORY DATA

Driven by the increased availability of mobile communication devices with embedded positioning capabilities, location-based services (LBS) have become very popular in recent years. By utilizing the location information of individuals, LBS can bring convenience to our daily life. For example, one can search for recommendations about restaurant that are close to his current position, or monitor congestion levels of vehicle traffic in certain places. However, the use of private location information may raise serious privacy problems. Among the many privacy issues in LBS [36], [37], here we focus on the privacy threat brought by publishing trajectory data of individuals. To provide location-based services, commercial entities (e.g. a telecommunication company) and public entities (e.g. a transportation company) collect large amount of individuals' trajectory data, i.e. sequences of consecutive location readings along with time stamps. If the data collector publish such spatio-temporal data to a third party (e.g. a data-mining company), sensitive information about individuals may be disclosed. For example, an advertiser may make inappropriate use of an individual's food preference which is inferred from his frequent visits to some restaurant. To realize a privacy-preserving publication, anonymization techniques can be applied to the trajectory data set, so that no sensitive location can be linked to a specific individual. Compared to relational data, spatio-temporal data have some unique characteristics, such as time dependence, location dependence and high dimensionality. Therefore, traditional anonymization approaches cannot be directly applied.

Terrovitis and Mamoulis [38] first investigate the privacy problem in the publication of location sequences. They study how to transform a database of trajectories to a format that would prevent adversaries, who hold a projection of the data, from inferring locations missing in their projections with high certainty. They propose a technique that iteratively suppresses selected locations from the original trajectories until a privacy constraint is satisfied. For example, as shown in Fig. 10, if an adversary Jhon knows that his target Mary consecutively visited two location a_1 and a_3, then he can knows for sure that the trajectory t_3 corresponds to Mary, since there is only

id	trajectory
t_1	$a_1 \rightarrow b_1 \rightarrow a_2$
t_2	$a_1 \rightarrow b_1 \rightarrow a_2 \rightarrow b_3$
t_3	$a_1 \rightarrow a_3 \rightarrow b_1$
t_4	$a_3 \rightarrow b_1$
t_5	$a_3 \rightarrow b_2$

(a)

id	trajectory
t_1	$a_1 \rightarrow b_1 \rightarrow a_2$
t_2	$a_1 \rightarrow b_1 \rightarrow a_2$
t_3	$a_3 \rightarrow b_1$
t_4	$a_3 \rightarrow b_1$
t_5	$a_3 \rightarrow b_2$

(b)

FIGURE 10. Anonymizing trajectory data by suppression [38]. (a) original data. (b) transformed data.

trajectory that goes through a_1 and a_3. While if some of the locations are suppressed, as shown in Fig. 10(a), Jhon cannot distinguish between t_3 and t_4, thus the trajectory of Mary is not disclosed. Based on Terrovitis and Mamoulis's work, researchers have now proposed many approaches to solve the privacy problems in trajectory data publishing. Considering that quantification of privacy plays a very important role in the study of PPDP, here we briefly review the privacy models adopted in these studies, especially those proposed in very recent literatures.

Nergiz et al. [39] redefine the notion of k-anonymity for trajectories and propose a heuristic method for achieving the anonymity. In their study, an individual's trajectory is represented by an ordered set of spatio-temporal points. Adversaries are assumed to know about all or some of the spatio-temporal points about an individual, thus the set of all points corresponding to a trajectory can be used as the quasi-identifiers. They define *trajectory k-anonymity* as follows: a trajectory data set T^* is k-anonymization of a trajectory data set T if for every trajectory in T^*, there are at least $k-1$ other trajectories with exactly the same set of points.

Abul et al. [40] propose a new concept of k-anonymity based on co-localization which exploits the inherent uncertainty of the moving object's whereabouts. The trajectory of a moving object is represented by a cylindrical volume instead of a polyline in a three-dimensional space. The proposed privacy model is called (k, δ)-anonymity, where the radius parameter δ represents the possible location imprecision (uncertainty). The basic idea is to modify the paths of trajectories so that k different trajectories co-exist in a cylinder of the radiusδ.

Yarovoy et al [41] consider it is inappropriate to use a set of particular locations or timestamps as the QID (quasi-identifier) for all individuals' trajectory data. Instead, various moving objects may have different QIDs. They define QID as a function mapping from a moving object database $D = \{O_1, O_2, \ldots, O_n\}$ that corresponds to n individuals, to a set of m discrete time points $T = \{t_1, \ldots, t_m\}$. Based on this definition of QID, k-anonymity can be redefined as follows: for every moving object O in D, there exist at least $k - 1$ other distinct moving objects $O_1, \ldots, O_{k-1}$, in the modified database D^*, such that $\forall t \in QID(O)$, O is indistinguishable from each of $O_1, \ldots, O_{k-1}$ at time t. One thing should be noted that to generate the k-anonymous database D^*, the data collector must be aware of the QIDs of all moving objects.

Chen et al. [42] assume that, in the context of trajectory data, an adversary's background knowledge on a target individual is bounded by at most L location-time pairs. They propose a privacy model called $(K, C)_L$-privacy for trajectory data anonymization, which considers not only identity linkage attacks on trajectory data, but also attribute linkage attacks via trajectory data. An adversary's background knowledge κ is assumed to be any non-empty subsequence q with $|q| \leq L$ of any trajectory in the trajectory database T. Intuitively, $(K, C)_L$-privacy requires that every subsequence q with $|q| \leq L$ in T is shared by at least a certain number of records,which means the confidence of inferring any sensitive value via q cannot be too high.

Ghasemzadeh et al. [43] propose a method for achieving anonymity in a trajectory database while preserving the information to support effective passenger flow analysis. A privacy model called LK-privacy is adopted in their method to prevent identity linkage attacks. The model assumes that an adversary knows at most L previously visited spatio-temporal pairs of any individual. The LK-privacy model requires every subsequence with length at most L in a trajectory database T to be shared by at least K records in T, where L and K are positive integer thresholds. This requirement is quite similar to the $(K, C)_L$-privacy proposed in [42].

Different from previous anonymization methods which try to achieve a privacy requirement by grouping the trajectories, Cicek et al. [44] group nodes in the underlying map to create obfuscation areas around sensitive locations. The sensitive nodes on the map are pre-specified by the data owner. Groups are generated around these sensitive nodes to form supernodes. Each supernode replaces nodes and edges in the corresponding group, therefore acts as an obfuscated region. They introduce a privacy metric called p-confidentiality with p measuring the level of privacy protection for the individuals. That is, given the path of a trajectory, p bounds the probability that the trajectory stops at a sensitive node in any group.

Poulis et al. [45] consider previous anonymization methods either produce inaccurate data, or are limited in their privacy specification component always. As a result, the cost of data utility is high. To overcome this shortcoming, they propose an approach which applies k^m-anonymity to trajectory data and performs generalization in a way that minimizes the distance between the original trajectory data and the anonymized one. A trajectory is represented by an ordered list of locations that are visited by a moving object. A subtrajectory is formed by removing some locations from the original trajectory, while maintaining the order of the remaining locations. A set of trajectories T satisfies k^m- anonymity if and only if every subtrajectory s of every trajectory $t \in T$, which contains m or fewer locations, is contained in at least k distinct trajectories of T. For example, as shown in Fig. 11, if an adversary knows that someone visited location c and then e, then he can infer that the individual corresponds to the trajectory t_1. While given the 2^2-anonymous data, the adversary cannot make a confident inference, since the subtrajectory (c, e) appears in four trajectories.

The privacy models introduced above can all be seen as variants of the classic k-anonymity model. Each model has its own assumptions about the adversary's background knowledge, hence each model has its limitations. A more detailed survey of adversary knowledge, privacy model, and anonymization algorithms proposed for trajectory data publication can be found in [46].

id	trajectory
t_1	(d, a, c, e)
t_2	(b, a, e, c)
t_3	(a, d, e)
t_4	(b, d, e, c)
t_5	(d, c)
t_6	(d, e)

(a)

id	trajectory
t_1	$(d, \{a, b, c\}, \{a, b, c\}, e)$
t_2	$(\{a, b, c\}, \{a, b, c\}, e, \{a, b, c\})$
t_3	$(\{a, b, c\}, d, e)$
t_4	$(\{a, b, c\}, d, e, \{a, b, c\})$
t_5	$(d, \{a, b, c\})$
t_6	(d, e)

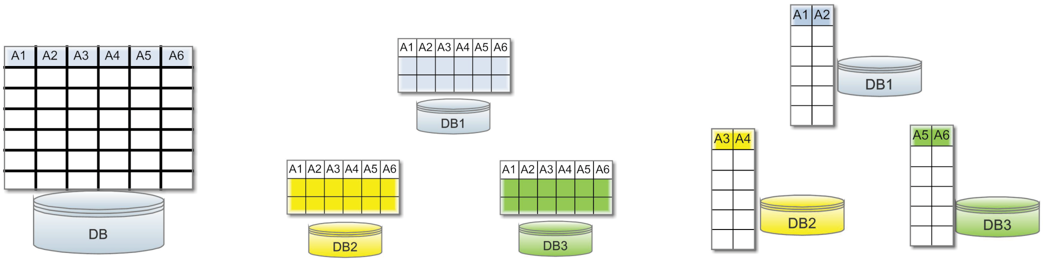
(b)

FIGURE 11. Anonymizing trajectory data by generalization [45]. (a) original data. (b) 2^2-anonymous data.

C. SUMMARY

Privacy-preserving data publishing provides methods to hide identity or sensitive attributes of original data owner. Despite the many advances in the study of data anonymization, there remain some research topics awaiting to be explored. Here we highlight two topics that are important for developing a practically effective anonymization method, namely personalized privacy preservation and modeling the background knowledge of adversaries.

Current studies on PPDP mainly manage to achieve privacy preserving in a statistical sense, that is, they focus on a universal approach that exerts the same amount of preservation for all individuals. While in practice, the implication of privacy varies from person to person. For example, someone considers salary to be sensitive information while someone doesn't; someone cares much about privacy while someone cares less. Therefore, the ''personality'' of privacy must be taken into account when anonymizing the data. Some researcher have already investigated the issue of personalized privacy preserving. In [47], Xiao and Tao present a generalization framework based on the concept of *personalized anonymity*, where an individual can specify the degree of privacy protection for his sensitive data. Some variants of k-anonymity have also been proposed to support personalized privacy preservation, such as (P, α, K)-anonymity [48], personalized (α, k)-anonymity [49], PK-anonymity [50], individualized (α, k)-anonymity [51], etc. In current studies, individual's personalized preference on privacy preserving is formulated through the parameters of the anonymity model (e.g. the value of k, or the degree of attention paid on certain sensitive value), or nodes in a domain generalization hierarchy. The data provider needs to declare his own privacy requirements when providing data to the collector. However, it is somewhat unrealistic to expect every data provider to define his privacy preference in such a formal way. As ''personalization'' becomes a trend in current data-driven applications, issues related to personalized data anonymization, such as how to formulate personalized privacy preference in a more flexible way and how to obtain such preference with less effort paid by data providers, need to be further investigated in future research.



(a) (b) (c)

FIGURE 12. Data distribution. (a) centralized data. (b) horizontally partitioned data. (c) vertically partitioned data.

The objective of data anonymization is to prevent the potential adversary from discovering information about a certain individual (i.e. the target). The adversary can utilize various kinds of knowledge to dig up the target's information from the published data. From previous discussions on social network data publishing and trajectory data publishing we can see that, if the data collector doesn't have a clear understanding of the capability of the adversary, i.e. the knowledge that the adversary can acquire from other resources, the knowledge which can be learned from the published data, and the way through which the knowledge can help to make an inference about target's information, it is very likely that the anonymized data will be de-anonymized by the adversary. Therefore, in order to design an effective privacy model for preventing various possible attacks, the data collector first needs to make a comprehensive analysis of the adversary's background knowledge and develop proper models to formalize the attacks. However, we are now in an open environment for information exchange, it is difficult to predict from which resources the adversary can retrieve information related to the published data. Besides, as the data type becomes more complex and more advanced data analysis techniques emerge, it is more difficult to determine what kind of knowledge the adversary can learn from the published data. Facing above difficulties, researches should explore more approaches to model adversary's background knowledge. Methodologies from data integration [52], information retrieval, graph data analysis, spatio-temporal data analysis, can be incorporated into this study.

IV. DATA MINER

A. CONCERNS OF DATA MINER

In order to discover useful knowledge which is desired by the decision maker, the data miner applies data mining algorithms to the data obtained from data collector. The privacy issues coming with the data mining operations are twofold. On one hand, if personal information can be directly observed in the data and data breach happens, privacy of the original data owner (i.e. the data provider) will be compromised. On the other hand, equipping with the many powerful data mining techniques, the data miner is able to find out various kinds of information underlying the data. Sometimes the data mining results may reveal sensitive information about the data owners. For example, in the Target story we mentioned in Section I-B, the information about the daughter's pregnancy, which is inferred by the retailer via mining customer data, is something that the daughter does not want others to know. To encourage data providers to participate in the data mining activity and provide more sensitive data, the data miner needs to make sure that the above two privacy threats are eliminated, or in other words, data providers' privacy must be well preserved. Different from existing surveys on privacy-preserving data mining (PPDM), in this paper, we consider it is the data collector's responsibility to ensure that sensitive raw data are modified or trimmed out from the published data(see Section III). The primary concern of data miner is how to prevent sensitive information from appearing in the mining results. To perform a privacy-preserving data mining, the data miner usually needs to modify the data he got from the data collector. As a result, the decline of data utility is inevitable. Similar to data collector, the data miner also faces the privacy-utility trade-off problem. But in the context of PPDM, quantifications of privacy and utility are closely related to the mining algorithm employed by the data miner.

B. APPROACHES TO PRIVACY PROTECTION

Extensive PPDM approaches have been proposed (see [5]–[7] for detailed surveys). These approaches can be classified by different criteria [53], such as data distribution, data modification method, data mining algorithm, etc. Based on the distribution of data, PPDM approaches can be classified into two categories, namely approaches for centralized data mining and approaches for distributed data mining. Distributed data mining can be further categorized into data mining over horizontally partitioned data and data mining over vertically partitioned data (see Fig. 12). Based on the technique adopted for data modification, PPDM can be classified into perturbation-based, blocking-based, swapping-based, etc. Since we define the privacy-preserving goal of data miner as preventing sensitive information from being revealed by the data mining results, in this section, we classify PPDM approaches according to the type of data mining tasks. Specifically, we review recent studies on privacy-preserving association rule mining, privacy-preserving classification, and privacy-preserving clustering, respectively.

Since many of the studies deal with distributed data mining where secure multi-party computation [54] is widely applied, here we make a brief introduction of secure multi-party computation (SMC). SMC is a subfield of cryptography. In general, SMC assumes a number of participants $P_1, P_2, \ldots, P_m$, each has a private data, $X_1, X_2, \ldots, X_m$. The participants want to compute the value of a public function f on m variables at the point $X_1, X_2, \ldots, X_m$. A SMC protocol is called *secure*, if at the end of the computation, no participant knows anything except his own data and the results of global calculation. We can view this by imagining that there is a trusted-third-party (TTP). Every participant give his input to the TTP, and the TTP performs the computation and sends the results to the participants. By employing a SMC protocol, the same result can be achieved without the TTP. In the context of distributed data mining, the goal of SMC is to make sure that each participant can get the correct data mining result without revealing his data to others.

1) PRIVACY-PRESERVING ASSOCIATION RULE MINING

Association rule mining is one of the most important data mining tasks, which aims at finding interesting associations and correlation relationships among large sets of data items [55]. A typical example of association rule mining is market basket analysis [1], which analyzes customer buying habits by finding associations between different items that customers place in their "shopping baskets". These associations can help retailers develop better marketing strategies. The problem of mining association rules can be formalized as follows [1]. Given a set of items $I = \{i_1, i_2, \ldots, i_m\}$, and a set of transactions $T = \{t_1, t_2, \ldots, t_n\}$, where each transaction consists of several items from I. An association rule is an implication of the form: $A \Rightarrow B$, where $A \subset I, B \subset I$, $A \neq \emptyset$, $B \neq \emptyset$, and $A \cap B \neq \emptyset$. The rule $A \Rightarrow B$ holds in the transaction set T with support s, where s denotes the percentage of transactions in T that contain $A \cup B$. The rule $A \Rightarrow B$ has confidence c in the transaction set T, where c is the percentage of transactions in T containing A that also contain B. Generally, the process of association rule mining contains the following two steps:

- Step 1: Find all frequent itemsets. A set of items is referred to as an *itemset*. The occurrence frequency of an itemset is the number of transactions that contain the itemset. A frequent itemset is an itemset whose occurrence frequency is larger than a predetermined minimum support count.
- Step 2: Generate strong association rules from the frequent itemsets. Rules that satisfy both a minimum support threshold ($min_s up$) and a minimum confidence threshold ($min_c onf$) are called strong association rules.

Given the thresholds of *support* and *confidence*, the data miner can find a set of association rules from the transactional data set. Some of the rules are considered to be sensitive, either from the data provider's perspective or from the data miner's perspective. To hiding these rules, the data miner can modify the original data set to generate a *sanitized* data set from which sensitive rules cannot be mined, while those non-sensitive ones can still be discovered, at the same thresholds or higher.

Various kinds of approaches have been proposed to perform association rule hiding [56], [57]. These approaches can roughly be categorized into the following five groups:

- Heuristic distortion approaches, which resolve how to select the appropriate data sets for data modification.
- Heuristic blocking approaches, which reduce the degree of support and confidence of the sensitive association rules by replacing certain attributes of some data items with a specific symbol (e.g. '?').
- Probabilistic distortion approaches, which distort the data through random numbers generated from a predefined probability distribution function.
- Exact database distortion approaches, which formulate the solution of the hiding problem as a constraint satisfaction problem (CSP), and apply linear programming approaches to its solution.
- Reconstruction-based approaches, which generate a database from the scratch that is compatible with a given set of non-sensitive association rules.

The main idea behind association rule hiding is to modify the support and/or confidence of certain rules. Here we briefly review some of the modification approaches proposed in recent studies.

Transaction ID	Items	Modified Items
T1	ABC	AB
T2	ABC	ABC
T3	ABC	ABC
T4	AB	AB
T5	A	AC
T6	AC	AC

FIGURE 13. Altering the position of sensitive item (e.g. *C*) to hide sensitive association rules [58].

Jain et al. [58] propose a distortion-based approach for hiding sensitive rules, where the position of the sensitive item is altered so that the confidence of the sensitive rule can be reduced, but the support of the sensitive item is never changed and the size of the database remains the same. For example, given the transactional data set shown in Fig. 13, set the threshold of support at 33% and the threshold of confidence at 70%, then the following three rules can be mined from the data: $C \Rightarrow A$ (66.67%, 100%), $A, B \Rightarrow C$ (50%, 75%), $C, A \Rightarrow B$ (50%, 75%). If we consider the item C to be a sensitive item, then we can delete C from the transaction $T1$, and add C to the transaction $T5$. As a result, the above three rules cannot be mined from the modified data set.

Zhu et al. [59] employ *hybrid partial hiding* (HPH) algorithm to reconstruct the support of itemset, and then uses Apriori [1] algorithm to generate frequent itemsets based on which only non-sensitive rules can be obtained. Le et al. [60] propose a heuristic algorithm based on the intersection lattice of frequent itemsets for hiding sensitive rules. The algorithm first determines the *victim item* such that modifying this item causes the least impact on the set of frequent itemsets. Then, the minimum number of transactions that need to be modified are specified. After that, the victim item is removed from the specified transactions and the data set is sanitized. Dehkoridi [61] considers hiding sensitive rules and keeping the accuracy of transactions as two objectives of some fitness function, and applies genetic algorithm to find the best solution for sanitizing original data. Bonam et al. [62] treat the problem of reducing frequency of sensitive item as a non-linear and multidimensional optimization problem. They apply *particle swarm optimization* (PSO) technique to this problem, since PSO can find high-quality solutions efficiently while requiring negligible parametrization.

Modi et al. [63] propose a heuristic algorithm named DSRRC (decrease support of right hand side item of rule clusters) for hiding sensitive association rules. The algorithm clusters the sensitive rules based on certain criteria in order to hide as many as possible rules at one time. One shortcoming of this algorithm is that it cannot hide association rules with multiple items in antecedent (left hand side) and consequent (right hand side). To overcome this shortcoming, Radadiya et al. [64] propose an improved algorithm named ADSRRC (advance DSRRC), where the item with highest count in right hand side of sensitive rules are iteratively deleted during the data sanitization process. Pathak et al. [65] propose a hiding approach which uses the concept of *impact factor* to build clusters of association rules. The impact factor of a transaction is equal to number of itemsets that are present in those itemsets which represents sensitive association rule. Higher impact factor means higher sensitivity. Utilizing the impact factor to build clusters can help to reduce the number of modifications, so that the quality of data is less affected.

Among different types of approaches proposed for sensitive rule hiding, we are particularly interested in the reconstruction-based approaches, where a special kind of data mining algorithms, named *inverse frequent set mining* (*IFM*), can be utilized. The problem of IFM was first investigated by Mielikäinen in [66]. The IFM problem can be described as follows [67]: given a collection of frequent itemsets and their support, find a transactional data set such that the data set precisely agrees with the supports of the given frequent itemset collection while the supports of other itemsets would be less than the pre-determined threshold. Guo et al [68] propose a reconstruction-based approach for association rule hiding where data reconstruction is implemented by solving an IFM problem. Their approach consists of three steps (see Fig. 14):

- First, use frequent itemset mining algorithm to generate all frequent itemsets with their supports and support counts from original data set.
- Second, determine which itemsets are related to sensitive association rules and remove the sensitive itemsets.
- Third, use the rest itemsets to generate a new transactional data set via inverse frequent set mining.

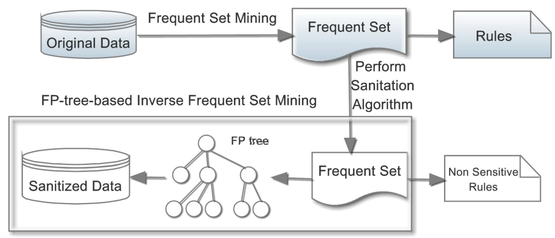


FIGURE 14. Reconstruction-based association rule hiding [68].

The idea of using IFM to reconstruct sanitized data set seems appealing. However, the IFM problem is difficult to solve. Mielikäinen [66] has proved that deciding whether there is a data set compatible with the given frequent sets is NP-complete. Researchers have made efforts towards reducing the computational cost of searching a compatible data set. Some representative algorithms include the vertical database generation algorithm [67], the linear program based algorithm [69], and the FP-tree-based method [70]. Despite the difficulty, the IFM problem does provide us some interesting insights on the privacy preserving issue. Inverse frequent set mining can be seen as the inverse problem of frequent set mining. Naturally, we may wonder whether we can define inverse problems for other types of data mining problems. If the inverse problem can be clearly defined and feasible algorithms for solving the problem can be found, then the data miner can use the inverse mining algorithms to *customize* the data to meet the requirements for data mining results, such as the support of certain association rules, or specific distributions of data categories. Therefore, we think it is worth exploring the inverse mining problems in future research.

2) PRIVACY-PRESERVING CLASSIFICATION

Classification [1] is a form of data analysis that extracts models describing important data classes. Data classification can be seen as a two-step process. In the first step, which is called *learning* step, a classification algorithm is employed to build a *classifier* (classification model) by analyzing a training set made up of tuples and their associated class labels. In the second step, the classifier is used for classification, i.e. predicting categorical class labels of new data. Typical classification model include decision tree, Bayesian model, support vector machine, etc.

a: DECISION TREE

A decision tree is a flowchart-like tree structure, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) represents a class label [1]. Given a tuple X, the attribute values of the tuple are tested against the decision tree. A path is traced from the root to a leaf node which holds the class prediction for the tuple. Decision trees can easily be converted to classification rules.

To realize privacy-preserving decision tree mining, Dowd et al. [71] propose a data perturbation technique based on *random substitutions*. Given a data tuple, the perturbation is done by replacing the value of an attribute by another value that is chosen randomly from the attribute domain according to a probabilistic model. They show that such perturbation is immune to *data-recovery attack* which aims at recovering the original data from the perturbed data, and *repeated-perturbation attack* where an adversary may repeatedly perturb the data with the hope to recover the original data. Brickell and Shmatikov [72] present a cryptographically secure protocol for privacy-preserving construction of decision trees. The protocol takes place between a user and a server. The user's input consists of the parameters of the decision tree that he wishes to construct, such as which attributes are treated as features and which attribute represents the class. The server's input is a relational database. The user's protocol output is a decision tree constructed from the server's data, while the server learns nothing about the constructed tree. Fong et al. [73] introduce a perturbation and randomization based approach to protect the data sets utilized in decision tree mining. Before being released to a third party for decision tree construction, the original data sets are converted into a group of unreal data sets, from which the original data cannot be reconstructed without the entire group of unreal data sets. Meanwhile, an accurate decision tree can be built directly from the unreal data sets. Sheela and Vijayalakshmi [74] propose a method based on *secure multi-party computation* (SMC) [75] to build a privacy-preserving decision tree over vertically partitioned data. The proposed method utilizes Shamir's secret sharing algorithm to securely compute the cardinality of scalar product, which is needed when computing information gain of attributes during the construction of the decision tree.

b: NAÏVE BAYESIAN CLASSIFICATION

Naïve Bayesian classification is based on Bayes' theorem of posterior probability. It assumes that the effect of an attribute value on a given class is independent of the values of other attributes. Given a tuple, a Bayesian classifier can predict the probability that the tuple belongs to a particular class.

Vaidya et al [76] study the privacy-preserving classification problem in a distributed scenario, where multi-parties collaborate to develop a classification model, but no one wants to disclose its data to others. Based on previous studies on secure multi-party computation, they propose different protocols to learn naïve Bayesian classification models from vertically partitioned or horizontally partitioned data. For horizontally partitioned data, all the attributes needed for classifying an instance are held by one site. Each party can directly get the classification result, therefore there is no need to hide the classification model. While for vertically partitioned data, since one party does not know all the attributes of the instance, he cannot learn the full model, which means sharing the classification model is required. In this case, protocols which can prevent the disclosure of sensitive information contained in the classification model (e.g. distributions of sensitive attributes) are desired. Skarkala et al. [77] also study the privacy-preserving classification problem for horizontally partitioned data. They propose a privacy-preserving version of the *tree augmented* naïve (TAN) Bayesian classifier [78] to extract global information from horizontally partitioned data. Compared to classical naïve Bayesian classifier, TAN classifier can produce better classification results, since it removes the assumption about conditional independence of attribute. Different from above work, Vaidya et al. [79] consider a centralized scenario, where the data miner has centralized access to a data set. The miner would like to release a classifier on the premise that sensitive information about the original data owners cannot be inferred from the classification model. They utilize differential privacy model [80] to construct a privacy-preserving Naïve Bayesian classifier. The basic idea is to derive the sensitivity for each attribute and to use the sensitivity to compute Laplacian noise. By adding noise to the parameters of the classifier, the data miner can get a classifier which is guaranteed to be differentially private.

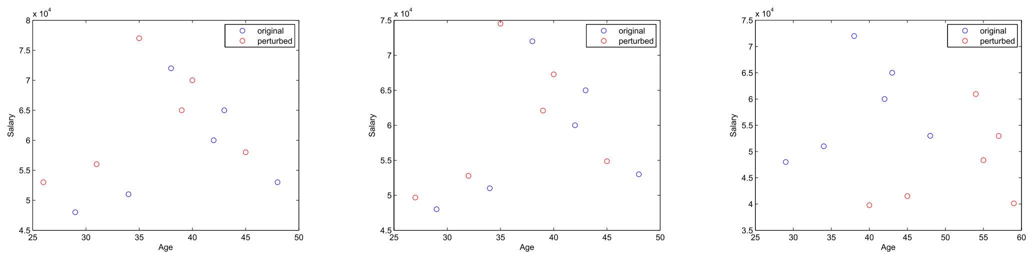
c: SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) is widely used in classification [1]. SVM uses a nonlinear mapping to transform the original training data into a higher dimension. Within this new dimension, SVM searches for a linear optimal separating hyperplane (i.e. a "decision boundary" separating tuples of one class from another), by using *support vectors* and *margins* (defined by the support vectors).

Vaidya et al. [81] propose a solution for constructing a global SVM classification model from data distributed at multiple parties, without disclosing the data of each party. They consider the *kernel matrix*, which is the central structure in a SVM, to be an intermediate profile that does not disclose any information on local data but can generate the global model. They propose a method based on gram matrix computation to securely compute the kernel matrix from the distributed data. Xia et al. [82] consider that the privacy threat of SVM-based classification comes from the support vectors in the learned classifier. The support vectors are intact instances taken from training data, hence the release of the SVM classifier may disclose sensitive information about the original owner of the training data. They develop a privacy-preserving SVM classifier based on hyperbolic tangent kernel. The kernel function in the classifier is an approximation of the original one. The degree of the approximation, which is determined by the number of support vectors, represents the level of privacy preserving. Lin and Chen [83] also think the release of support vectors will violate individual's privacy. They design a privacy-preserving SVM classifier based on Gaussian kernel function. Privacy-preserving is realized by transforming the original decision function, which is determined by support vectors, to an infinite series of linear combinations of monomial feature mapped support vectors. The sensitive content of support vectors are destroyed by the linear combination, while the decision function can precisely approximate the original one.

TABLE 1. Approaches to privacy-preserving classification.

	Data Mining Algorithm	Data Distribution	Privacy Concerns	Method Description	Performance Measurement
Dowd et al. [71]	C4.5 decision tree learning	centralized	prevent data-recovery attack and repeated-perturbation attack	random substitution-based data perturbation; data reconstruction	estimation error of data distribution; classification accuracy
Brickell and Shmatikov [72]	recursive decision-tree learning (CART algorithm)	asymmetrically distributed (user provides parameters, server provides data)	Server: reveal information about its data as little as possible User: the selected feature attributes and class attribute are not revealed to the server	SMC-based protocol; build the tree "one tier at a time"	online time required by the protocol
Fong et al. [73]	ID3 decision tree learning	centralized	decrease the privacy loss incurred by the match between the sanitized data set and the original data set	data set complementation approach where an extra perturbed data set is utilized	classification accuracy; storage complexity; privacy loss
Sheela and Vijayalakshmi [74]	ID3 decision tree learning	distributed (vertically partitioned)	the original data of each party cannot be revealed to others	SMC-based protocol; Using Shamir's secret sharing to find the cardinality of the scalar product	effect of collusion on security, communication cost, computation cost
Vaidya et al. [76]	naïve Bayesian	distributed (vertically/horizontally partitioned)	horizontally partitioned: learn the classifier without revealing each party's data; vertically partitioned: the model parameters also needed to be hidden	several secure computation protocols, e.g. secure sum, scalar product protocol, square computation, etc.	effect of collusion on security, communication cost, computation cost
Skarkala et al. [77]	tree augmented naïve Bayesian	Distributed (horizontally partitioned)	confidentiality of data exchanged among one party and the miner; anonymity and un-linkability of each party's identity	SMC-based protocol; Paillier cryptosystem	computation time, classification accuracy
Vaidya et al. [79]	naïve Bayesian	centralized	the classifier should be differentially private	adding noise to classifier's parameters	classification accuracy
Vaidya et al. [81]	SVM	distributed (vertically/horizontally/arbitrarily partitioned	each party's data should not be revealed	using gram matrix to compute the kernel matrix; secure computation protocols	effect of collusion on security, computation cost, communication cost
Xia et al. [82]	SVM	centralized	support vectors in the learned classifier should be hidden	using hyperbolic tangent kernel to approximate the original decision function	classification accuracy
Lin and Chen [83]	SVM	centralized	support vectors in the learned classifier should be hidden	approximating the original decision function by using an infinite series of linear combinations of monomial feature mapped support vectors	security against attacks on support vectors; approximating Precision



(a) (b) (c)

FIGURE 15. Examples of geometric data transformation [84]. Red circles represent original data and blue circles represent perturbed data. Data are perturbed in 3 ways: (a) translation; (b) scaling; (c) rotation.

In above discussions we briefly reviewed the privacy-preserving approaches proposed for different classification models. To provide a clear view of these studies, we summarize the main points of some representative approaches in Table 1.

3) PRIVACY-PRESERVING CLUSTERING

Cluster analysis [1] is the process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters. Dissimilarities and similarities are assessed based on the attribute values describing the objects and often involve distance measures. Clustering methods can be categorized into partitioning methods, hierarchical methods, density-based methods, etc.

Current studies on privacy-preserving clustering can be roughly categorized into two types, namely approaches based on perturbation and approaches based on secure multi-party computation (SMC).

Perturbation-based approach modifies the data before performing clustering. Oliveira and Zaiane [84] introduce a family of geometric data transformation methods for privacy-preserving clustering. The proposed transformation methods distort confidential data attributes by translation, scaling, or rotation (see Fig. 15), while general features for cluster analysis are preserved. Oliveira and Zaiane have demonstrated that the transformation methods can well balance privacy and effectiveness, where privacy is evaluated by computing the variance between actual and perturbed values, and effectiveness is evaluated by comparing the number of legitimate points grouped in the original and the distorted databases. The methods proposed in [84] deal with numerical attributes, while in [84], Rajalaxmi and Natarajan propose a set of hybrid data transformations for categorical attributes. Recently, Lakshmi and Rani [85] propose two hybrid methods to hide the sensitive numerical attributes. The methods utilize three different techniques, namely singular value decomposition (SVD), rotation data perturbation and independent component analysis. SVD can identify information that is not important for data mining, while ICA can identify those important information. Rotation data perturbation can retains the statistical properties of the data set. Compared to method solely based on perturbation, the hybrid methods can better protect sensitive data and retain the important information for cluster analysis.

The SMC-based approaches make use of primitives from secure multi-party computation to design a formal model for preserving privacy during the execution of a clustering algorithm. Two pioneer studies on SMC-based clustering are presented in [86] and [87]. Vaidya and Clifton [86] present a privacy-preserving method for k-means clustering over vertically partitioned data, where multiple data sites, each having different attributes for the same set of data points, wish to conduct k-means clustering on their joint data. At each iteration of the clustering process, each site can securely find the cluster with the minimum distance for each point, and can independently compute the components of the cluster means corresponding to its attributes. A *checkThreshold* algorithm is proposed to determine whether the stopping criterion is met. Jha et al. [87] design a privacy-preserving k-means clustering algorithm for horizontally partitioned data, where only the cluster means at various steps of the algorithm are revealed to the participating parties. They present two protocols for privacy-preserving computation of cluster means. The first protocol is based on oblivious polynomial evaluation and the second one uses homomorphic encryption. Based on above studies, many privacy-preserving approaches have been developed for k-means clustering. Meskine and Bahloul present an overview of these approaches in [88].

Most of the SMC-based approaches deal with *semi-honest* model, which assumes that participating parties always follow the protocol. In a recent study, Akhter et al. [88] consider the malicious model, where a party may substitute its local input or abort the protocol prematurely. They propose a protocol based on NIZK (non-interactive zero knowledge) proofs to conducting privacy-preserving k-means clustering between two parties in a malicious model.

In [89], Yi and Zhang identify another shortcoming of previous protocols, that is, each party does not equally contribute to k-means clustering. As a result, a party, who learns the outcome prior to other parties, may tell a lie of the outcome to other parties. To prevent this perfidious attack, they propose a k-means clustering protocol for vertically partitioned data, in which each party equally contributes to the clustering. The basic idea is that, at each iteration of k-means clustering, multi-parties cooperate to encrypt k values (each corresponds to a distance between a data point and a cluster center) with a common public key, and then securely compare the k values in order to assign the point to the closest cluster. Based on the assignment, each party can update the means corresponding to his own attributes. Intermediate information during the clustering process, such as the aforementioned k values, are not revealed to any party. Under this protocol, no party can learn the outcome prior to other parties.

Different from previous studies which focus on k-means clustering, De and Tripathy [90] recently develop a secure algorithm for hierarchical clustering over vertically partitioned data. There are two parties involved in the computation. In the proposed algorithm, each party first computes k clusters on their own private data set. Then, both parties compute the distance between each data point and each of the k cluster centers. The resulting distance matrices along with the randomized cluster centers are exchanged between the two parties. Based on the information provided by the other party, each party can compute the final clustering result.

C. SUMMARY

From above discussions we can see that, for a data miner, the privacy trouble may come from the discovery of sensitive knowledge(e.g. sensitive association rules), the release of the learned model (e.g. the SVM classifier), or the collaboration with other data miners. To fight against different privacy threats, the data miner needs to take different measures:

1) To prevent sensitive information from appearing in the mining results, the data miner can modify the original data via randomization,blocking, geometric transformation, or reconstruction. The modification often has a negative effect on the utility of the data. To make sure that those non-sensitive information can still be mined from the modified data, the data miner needs to make a balance between privacy and utility. The implications of privacy and data utility vary with the characteristics of data and the purpose of the mining task. As data types become more complex and new types of data mining applications emerge, finding appropriate ways to quantify privacy and utility becomes a challenging task, which is of high priority in future study of PPDM.
2) If the data miner needs to release the model learned (e.g. the decision function of a SVM classifier) from the data to others, the data miner should consider the possibility that some attackers may be able to infer sensitive information from the released model. Compared to privacy-preserving data publishing where attack models and corresponding privacy models have been clearly defined, current studies on PPDM pay less attention to the privacy attacks towards the data mining model. For different data mining algorithms, what kind of sensitive information can be inferred from the parameters of the model, what kind of background knowledge can be utilized by the attacker, and how to modify the model built from data to prevent the disclosure of sensitive information, these problems needs to be explored in future study.
3) When participating in a distributed data mining task, the data miner treats all his data as sensitive data, and his objective is to get the correct mining results without reveal his data to other participators.

Various SMC-based approaches have been proposed for privacy-preserving distributed data mining. What kind of information can be exchanged between different participators and how to exchange the information are formally defined by a protocol. However, it is no guarantee that every participator will follow the protocol or truthfully share his data. Interactions among different participators need to be further investigated. Considering the selfish nature of the data miner, game theory may be a proper tool for such problems. Some game theoretical approaches have been proposed for distributed data mining. We will discuss these approaches in Section VI.

4) The data miner has the ability to discover valuable information hidden in the data. Unwanted disclosure of such information may cause more serious problems than the leakage/breach/disclosure of original data. Studies on PPDM aim at developing algorithms that can preserve privacy without bringing too much side/negative effect to the mining results. But also, the data miner can utilize the PPDM approaches to punish the one who has made improper use of the mining results, so that the misbehaviors can be reduced.

V. DECISION MAKER

A. CONCERNS OF DECISION MAKER

The ultimate goal of data mining is to provide useful information to the decision maker, so that the decision maker can choose a better way to achieve his objective, such as increasing sales of products or making correct diagnoses of diseases. At a first glance, it seems that the decision maker has no responsibility for protecting privacy, since we usually interpret privacy as sensitive information about the original data owners (i.e. data providers). Generally, the data miner, the data collector and the data provider himself are considered to be responsible for the safety of privacy. However, if we look at the privacy issue from a wider perspective, we can see that the decision maker also has his own privacy concerns. The data mining results provided by the data miner are of high importance to the decision maker. If the results are disclosed to someone else, e.g. a competing company, the decision maker may suffer a loss. That is to say, from the perspective of decision maker, the data mining results are sensitive information. On the other hand, if the decision maker does not get the data mining results directly from the data miner, but from someone else which we called *information transmitter*, the decision maker should be skeptical about the credibility of the results, in case that the results have been distorted. Therefore, the privacy concerns of the decision maker are twofold: how to prevent unwanted disclosure of sensitive mining results, and how to evaluate the credibility of the received mining results.

B. APPROACHES TO PRIVACY PROTECTION

To deal with the first privacy issue proposed above, i.e. to prevent unwanted disclosure of sensitive mining results, usually the decision maker has to resort to legal measures. For example, making a contract with the data miner to forbid the miner from disclosing the mining results to a third party. To handle the second issue, i.e. to determine whether the received information can be trusted, the decision maker can utilize methodologies from data provenance, credibility analysis of web information, or other related research fields. In the rest part of this section, we will first briefly review the studies on data provenance and web information credibility, and then present a preliminary discussion about how these studies can help to analyze the credibility of data mining results.

1) DATA PROVENANCE

If the decision maker does not get the data mining results directly from the data miner, he would want to know how the results are delivered to him and what kind of modification may have been applied to the results, so that he can determine whether the results can be trusted. This is why ''provenance'' is needed. The term *provenance* originally refers to the chronology of the ownership, custody or location of a historical object. In information science, a piece of data is treated as the historical object, and *data provenance* refers to the information that helps determine the derivation history of the data, starting from the original source [91]. Two kinds of information can be found in the provenance of the data: the ancestral data from which current data evolved, and the transformations applied to ancestral data that helped to produce current data. With such information, people can better understand the data and judge the credibility of the data.

Since 1990s, data provenance has been extensively studied in the fields of databases and workflows. Several surveys are now available. In [91], Simmhan et al. present a taxonomy of data provenance techniques. The following five aspects are used to capture the characteristics of a provenance system:

- Application of provenance. Provenance systems may be constructed to support a number of uses, such as estimate data quality and data reliability, trace the audit trail of data, repeat the derivation of data, etc.
- Subject of provenance. Provenance information can be collected about different resources present in the data processing system and at various levels of detail.
- Representation of provenance. There are mainly two types of methods to represent provenance information, one is annotation and the other is inversion. The annotation method uses metadata, which comprise of the derivation history of the data, as annotations and descriptions about sources data and processes. The inversion method uses the property by which some derivations can be inverted to find the input data supplied to derive the output data.
- Provenance storage. Provenance can be tightly coupled to the data it describes and located in the same data storage system or even be embedded within the data file. Alternatively, provenance can be stored separately with other metadata or simply by itself.

- Provenance dissemination. A provenance system can use different ways to disseminate the provenance information, such as providing a derivation graph that users can browse and inspect.

In [92], Glavic et al. present another categorization scheme for provenance system. The proposed scheme consists of three main categorizes: provenance model, query and manipulation functionality, storage model and recording strategy. Davidson and Freire [93] review studies on provenance for scientific workflows. They summarize the key components of a provenance management solution, discuss applications for workflow provenance, and outline a few open problems for database-related research.

As Internet becomes a major platform for information sharing, provenance of Internet information has attracted some attention. Researchers have developed approaches for information provenance in semantic web [94], [95] and social media [96]. Hartig [94] proposes a provenance model that captures both the information about web-based data access and information about the creation of data. In this model, an ontology-based vocabulary is developed to describe the provenance information. Moreau [95] reviews research issues related to tracking provenance in semantic web from the following four aspects: publishing provenance on the web; using semantic web technologies to facilitate provenance acquisition, representation, and reasoning; tracking the provenance of RDF (resource description framework)-based information; tracking the provenance of inferred knowledge. Barbier and Liu [96] study the information provenance problem in social media. They model the social network as a directed graph $G(V, E, p)$, where V is the node set and E is the edge set. Each node in the graph represents an entity and each directed edge represents the direction of information propagation. An information propagation probability p is attached to each edge. Based on the model, they define *the information provenance problem* as follows: given a directed graph $G(V, E, p)$, with known terminals $T \subseteq V$, and a positive integer constant $k \in Z^+$, identify the sources $S \subseteq V$, such that $|S| \leq k$, and $U(S, T)$ is maximized. The function $U(S, T)$ estimates the utility of information propagation which starts from the sources S and stops at the terminals T. To solve this provenance problem, one can leverage the unique features of social networks, e.g. user profiles, user interactions, spatial or temporal information, etc. Two approaches are developed to seek the provenance of information. One approach utilizes the network information to directly seek the provenance of information, and the other approach aims at finding the reverse flows of information propagation.

The special characteristics of Internet, such as openness, freedom and anonymity, pose great challenges for seeking provenance of information. Compared to the approaches developed in the context of databases and workflows, current solutions proposed for supporting provenance in Internet environment are less mature. There are still many problems to be explored in future study.

2) WEB INFORMATION CREDIBILITY

Because of the lack of publishing barriers, the low cost of dissemination, and the lax control of quality, credibility of web information has become a serious issue. Tudjman et al. [97] identify the following five criteria that can be employed by Internet users to differentiate false information from the truth:

- Authority: the real author of false information is usually unclear.
- Accuracy: false information dose not contain accurate data or approved facts.
- Objectivity: false information is often prejudicial.
- Currency: for false information, the data about its source, time and place of its origin is incomplete, out of date, or missing.
- Coverage: false information usually contains no effective links to other information online.

In [98], Metzger summarizes the skills that can help users to assess the credibility of online information.

With the rapid growth of online social media, false information breeds more easily and spreads more widely than before, which further increases the difficulty of judging information credibility. Identifying rumors and their sources in microblogging networks has recently become a hot research topic [99]–[102]. Current research usually treats rumor identification as a classification problem, thus the following two issues are involved:

- Preparation of training data set. Current studies usually take rumors that have been confirmed by authorities as positive training samples. Considering the huge amount of messages in microblogging networks, such training samples are far from enough to train a good classifier. Building a large benchmark data set of rumors is in urgent need.
- Feature selection. Various kinds of features can be used to characterize the microblogging messages. In current literature, the following three types of features are often used: content-based features, such as word unigram/bigram, part-of-speech unigram/bigram, text length, number of sentiment word (positive/negative), number of URL, and number of hashtag; user-related features, such as registration time, registration location, number of friends, number of followers, and number of messages posted by the user; network features, such as number of comments and number of retweets.

So far, it is still quite difficult to automatically identifying false information on the Internet. It is necessary to incorporate methodologies from multiple disciplines, such as nature language processing, data mining, machine learning, social networking analysis, and information provenance, into the identification procedure.

C. SUMMARY

Provenance, which describes where the data came from and how the data evolved over time, can help people evaluate the credibility of data. For a decision maker, if he can acquire complete provenance of the data mining results, then he can easily determine whether the mining results are trustworthy. However, in most cases, provenance of the data mining results is not available. If the mining results are not directly delivered to the decision maker, it is very likely that they are propagated in a less controlled environment. As we introduced earlier, a major approach to represent the provenance information is adding annotations to data. While the reality is that the information transmitter has no motivation to make such annotations, especially when he attempts to alter the original mining results for his own interests. In other words, the possible transformation process of the mining results is non-transparent to the decision maker. In order to support provenance of the data mining results, setting up protocols, which explicitly demand the data miner and information transmitters to append provenance annotations to the data they delivered, is quite necessary. Also, standards which define the essential elements of the annotations should be created, so that the decision maker clearly knows how to interpret the provenance. In addition, techniques that help to automatically create the annotations are desired, with the purpose of reducing the cost of recording provenance information. Above issues should be further investigated in future research, not only because they can help the decision maker judge the credibility of data mining results, but also because they may induce constraints on transmitters' behaviors thus reduce the likelihood of distorted mining results.

Besides provenance, studies on identifying false Internet information also can provide some implications for decision makers. Inspired by the study on rumor identification, we consider it is reasonable to formalize the problem of evaluating credibility of data mining results as a classification problem. If the decision maker has accumulated some credible information from past interactions with the data miner or other reliable sources, a classifier, aiming at distinguishing between fake mining results and truthful results, can be built upon these information. Similar to the studies on microblogs, the decision maker needs to delicately choose the features to characterize the data mining results.

We have presented some preliminary thought on the credibility issue in above discussions. Detailed implementations of the provenance-based approach or the classification-based approach need to be further explored in future study.

VI. GAME THEORY IN DATA PRIVACY

A. GAME THEORY PRELIMINARIES

In above sections, we have discussed the privacy issues related to data provider, data collector, data miner and decision maker, respectively. Here in this section, we focus on the iterations among different users. When participating in a data mining activity, each user has his own consideration about the benefit he may obtain and the (privacy) cost he has to pay. For example, a company can make profit from the knowledge mined from customers' data, but he may need to pay high price for data containing sensitive information; a customer can get monetary incentives or better services by providing personal data to the company, but meanwhile he has to consider the potential privacy risks. Generally, the user would act in the way that can bring him more benefits, and one user's action may have effect on other users' interests. Therefore, it is natural to treat the data mining activity as a *game* played by multiple users, and apply game theoretical approaches to analyze the iterations among different users.

Game theory provides a formal approach to model situations where a group of agents have to choose optimum actions considering the mutual effects of other agents' decisions. The essential elements of a game are *players*, *actions*, *payoffs*, and *information* [8]. Players have actions that they can perform at designated times in the game. As a result of the performed actions, players receive payoffs. The payoff to each player depends on both the player's action and other players' actions. Information is modelled using the concept of *information set* which represents a player's knowledge about the values of different variables in the game. The outcome of the game is a set of elements picked from the values of actions, payoffs, and other variables after the game is played out. A player is called *rational* if he acts in such a way as to maximize his payoff. A player's strategy is a rule that tells him which action to choose at each instant of the game, given his information set. A strategy profile is an ordered set consisting of one strategy for each of the players in the game. An *equilibrium* is a strategy profile consisting of a best strategy for each of the players in the game. The most important equilibrium concept for the majority of games is *Nash equilibrium*. A strategy profile is a Nash equilibrium if no player has incentive to deviate from his strategy, given that other players do not deviate.

Game theory has been successfully applied to various fields, such as economics, political science, computer science, etc. Researchers have also employed game theory to deal with the privacy issues related to data mining. In following three subsections we will review some representative game theoretical approaches that are developed for data collection, distributed data mining and data anonymization.

B. PRIVATE DATA COLLECTION AND PUBLICATION

If a data collector wants to collect data from data providers who place high value on their private data, the collector may need to negotiate with the providers about the "price" of the sensitive data and the level of privacy protection. In [103], Adl et al. build a sequential game model to analyze the private data collection process. In the proposed model, a data user, who wants to buy a data set from the data collector, makes a price offer to the collector at the beginning of the game. If the data collector accepts the offer, he then announces some incentives to data providers in order to collect private data from them. Before selling the collected data to the data user, the data collector applies anonymization technique to the data, in order to protect the privacy of data providers at certain level. Knowing that data will be anonymized, the data user asks for a privacy protection level that facilitates his most preferable balance between data quality and quantity when making his offer. The data collector also announces a specific privacy protection level to data providers. Based on the protection level and incentives offered by data collector, a data provider decides whether to provide his data. In this data collection game, the level of privacy protection has significant influence on each player's action and payoff. Usually, the data collector and data user have different expectations on the protection level. By solving the subgame perfect Nash equilibriums of the proposed game, a consensus on the level of privacy protection can be achieved. In their later work [104], Adl et al. propose a similar game theoretical approach for aggregate query applications. They show that stable combinations of revelation level (how specific data are revealed), retention period of the collected data, price of per data item, and the incentives offered to data providers, can be found by solving the game's equilibriums. The game analysis has some implications on how to set a privacy policy to achieve maximum revenue while respecting data providers' privacy preferences. And the proposed game model can be potentially used for comparing different privacy protection approaches.

C. PRIVACY PRESERVING DISTRIBUTED DATA MINING

1) SMC-BASED PRIVACY PRESERVING DISTRIBUTED DATA MINING

As mentioned in Section IV-B, secure multi-party computation(SMC) is widely used in privacy preserving distributed data mining. In a SMC scenario, a set of mutually distrustful parties, each with a private input, jointly compute a function over their inputs. Some protocol is established to ensure that each party can only get the computation result and his own data stay private. However, during the execution of the protocol, a party may take one of the following actions in order to get more benefits:

- Semi-honest adversary: one follows the established protocol and correctly performs the computation but attempts to analyze others' private inputs;
- Malicious adversary: one arbitrarily deviates from the established protocol which leads to the failure of computation.
- Collusion: one colludes with several other parties to expose the private input of another party who doesn't participate in the collusion.

Kargupta et al. [105] formalize the SMC problem as a static game with complete information. By analyzing the Nash equilibriums, they find that if nobody is penalized for dishonest behavior, parties tend to collude. They also propose a cheap-talk based protocol to implement a punishment mechanism which can lead to an equilibrium state corresponding to no collusion. Miyaji et al [106] propose a two-party secure set-intersection protocol in a game theoretic setting. They assume that parties are neither honest nor corrupt but acted only in their own self-interest. They show that the proposed protocol satisfied computational versions of strict Nash equilibrium and stability with respect to trembles. Ge et al. [107] propose a SMC-based algorithm for privacy preserving distributed association rule mining(PPDARM). The algorithm employs Shamir's secret sharing technique to prevent the collusion of parties. In [108], Nanvati and Jinwala model the secret sharing in PPDARM as a repeated game, where a Nash equilibrium is achieved when all parties send their shares and attain a non-collusive behavior. Based on the game model, they develop punishment policies which aim at getting the maximum possible participants involved in the game so that they can get maximum utilities.

2) RECOMMENDER SYSTEM

Personalized recommendation is a typical application of data mining. The recommendation system predicts users' preference by analyzing the item ratings provided by users, thus the user can protect his private preference by falsifying his ratings. However, false ratings will cause a decline of the quality of recommendation. Halkidi et al. [109] employ game theory to address the trade-off between privacy preservation and high-quality recommendation. In the proposed game model, users are treated as players , and the rating data provided to the recommender server are seen as users' strategies. It has been shown that the Nash equilibrium strategy for each user is to declare false rating only for one item, the one that is highly ranked in his private profile and less correlated with items for which he anticipates recommendation. To find the equilibrium strategy, data exchange between users and the recommender server is modeled as an iterative process. At each iteration, by using the ratings provided by other users at previous iteration, each user computes a rating vector that can maximize the preservation of his privacy, with respect to a constraint of the recommendation quality. Then the user declare this rating vector to the recommender server. After several iterations, the process converges to a Nash equilibrium.

3) LINEAR REGRESSION AS A NON-COOPERATIVE GAME

Ioannidis and Loiseau [110] study the privacy issue in linear regression modeling. They consider a setting where a data analyst collects private data from multiple individuals to build a linear regression model. In order to protect privacy, individuals add noise to their data, which affects the accuracy of the model. In [110], the interactions among individuals are modeled as a non-cooperative game, where each individual selects the variance level of the noise to minimize his cost. The cost relates to both the privacy loss incurred by the release of data and the accuracy of the estimated linear regression model. It is shown that under appropriate assumptions on privacy and estimation costs, there exists a unique pure Nash equilibrium at which each individual's cost is bounded.

D. DATA ANONYMIZATION

Chakravarthy et al. [111] present an interesting application of game theory. They propose a k-anonymity method which utilizes coalitional game theory to achieve a proper privacy level, given the threshold for information loss. The proposed method models each tuple in the data table as a player, and computes the payoff to each player according to a

TABLE 2. Add caption.

Application	Players	Actions	Payoffs	Implications of Equilibriums
private data collection and publication [103]	data provider	opt-in; opt-out	incentives from data collector (if opt-in)	both the data collector and the data user satisfy with some value of the privacy parameter
	data collector	choose an incentive paid for each data record	income: price paid by data user; expenditure: incentives paid to data providers, a fixed cost of data collection, anonymization, and storing.	
	data user	(p, δ) p: price for each dat record δ: privacy parameter	income: depends on the precision of the data analysis result; expenditure: price paid to data collector	
SMC-based PPDDM [105]	participating parties	$\left(I_i^{(M)}, I_i^{(R)}, I_i^{(S)}, I_i^{(G)}\right)$ $I_i^{(M)}$: perform a required computation or not; $I_i^{(R)}$: receive message from other parties or not $I_i^{(S)}$: send message to other parties or not $I_i^{(G)}$: collude with other parties or not	a linear or nonlinear function of $\left(I_i^{(M)}, I_i^{(R)}, I_i^{(S)}, I_i^{(G)}\right)$	parties tend to collude
privacy-preserving recommendation [109]	multiple users	declare a false rating vetor $\mathbf{q}$ rather than the true rating vector $\mathbf{p}$	income: depends on the quality of recommendation results; expenditure: privacy loss represented by the distance bewteen $\mathbf{q}$ and $\mathbf{p}$	declare false rating only for one specific item
linear regression [110]	participating individuals: each has a public feature vector $\mathbf{x}$ and a private variable y	choose a λ_i (λ_i is actually determined by the variance of the noise added to the private variable)	privacy cost: depends on the player's own action λ_i; estimation cost: depends on actions of all players $\lambda = (\lambda_1, \cdots, \lambda_N)$	each player can get a correct estimation with a small cost
k-anonymity [111]	tuples in a data table	form a coalitional group with some other tuples	determined by the level that a tuple stays at in a concept hierarchy tree	each coalitional group corresponds to an equivalent class in the anonymous table

concept hierarchy tree (CHT) of quasi-identifiers. The equivalent class in the anonymous table is formed by establishing a coalition among different tuples based on their payoffs. Given the affordable information loss, the proposed method can automatically find the most feasible value of k, while traditional methods need to fix up the value of k before the anonymization process.

E. ASSUMPTIONS OF THE GAME MODEL

In above discussions we have reviewed the game theoretical approaches to privacy issues in data mining. We present the basic elements of some proposed game models in Table 2. Most of the proposed approaches adopt the following research paradigm:

- define the elements of the game, namely the players, the actions and the payoffs;
- determine the type of the game: static or dynamic, complete information or incomplete information;
- solve the game to find equilibriums;
- analyze the equilibriums to obtain some implications for practice.

The above paradigm seems to be simple and clear, while problems in real world can be very complicated. Usually we have to make a few assumptions when developing the game model. Unreasonable assumptions or too many assumptions will hurt the applicability of the game model. For example, the game theoretical approach proposed in [109] assumes that there is an iterative process of data exchange between users and the recommender server. To find the best response to other users' strategies, each user is assumed to be able to get a aggregated version of ratings provided by other users for each item, and can calculate the recommendation result by himself. However, in practical recommendation system, it is unlikely that the user would repeatedly modify the ratings he has already reported to the recommender server. Also, there are so many items in the system, it is unrealistic that a user will collect the ratings of all items. Besides, the recommendation algorithm employed by the recommender server is unknown to the user, hence the user cannot calculate the recommendations by himself. With these improper assumptions, the proposed game analysis can hardly provide meaningful guidance to user's rating action. Therefor, we think that future study on game theoretical approaches should pay more attention to the *assumptions*. Real-world problem should be formalized in a more realistic way, so that the game theoretical analysis can have more practical implications.

F. MECHANISM DESIGN AND PRIVACY PROTECTION

Mechanism design is a sub-field of microeconomics and game theory. It considers how to implement good system-wide solutions to problems that involve multiple self-interested agents with private information about their preferences for different outcomes [13]. Incorporating mechanism design into the study of privacy protecting has recently attracted some attention. In a nutshell, a mechanism defines the strategies available and the method used to select the final outcome based on agents' strategies. Specifically, consider a group of n agents $\{i\}$, and each agent i has a privately known type $t_i \in T$. A mechanism $M : T^n \rightarrow O$ is a mapping between (reported) types of the n agents, and some outcome space O. The agent's type t_i determines its preferences over different outcomes. The utility that the agent i with type t_i can get from the outcome $o \in O$ is denoted by $u_i(o, t_i)$. Agents are assumed to be rational, that is, agent i prefers outcome o_1 over o_2 when $u_i(o_1, t_i) > u_i(o_2, t_i)$. The mechanism designer designs the rules of a game, so that the agents will participate in the game and the equilibrium strategies of agents can lead to the designer's desired outcome.

Mechanism design is mostly applied to auction design, where an auction mechanism defines how to determine the winning bidder(s) and how much the bidder should pay for the goods. In the context of data mining, the data collector, who often plays the role of data miner as well, acts as the mechanism designer, and data providers are agents with private information. The data collector wants data providers to participate in the data mining activity, i.e. hand over their private data, but the data providers may choose to opt-out because of the privacy concerns. In order to get useful data mining results, the data collector needs to design mechanisms to encourage data providers to opt-in.

1) MECHANISMS FOR TRUTHFUL DATA SHARING

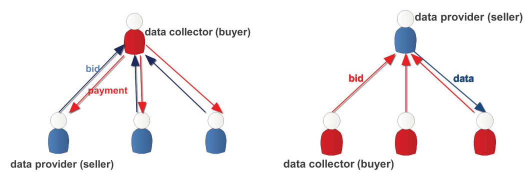
A mechanism requires agents to report their preferences over the outcomes. Since the preferences are private information and agents are self-interested, it is likely that the agent would report false preferences. In many cases, the mechanism is expected to be *incentive compatible* [13], that is, reporting one's true preferences should bring the agent larger utility than reporting false preferences. Such mechanism is also called *truthful mechanism*.

Researchers have investigated incentive compatible mechanisms for privacy preserving distributed data mining [112], [113]. In distributed data mining, data needed for the mining task are collected from multiple parties. Privacy-preserving methods such as secure multi-party computation protocols can guarantee that only the final result is disclosed. However, there is no guarantee that the data provided by participating parties are truthful. If the data mining function is reversible, that is, given two inputs, x and x' , and the result $f(x)$, a data provider is able to calculate $f\left(x'\right)$, then there is a motivation for the provider to provide false data in order to exclusively learn the correct mining result. To encourage truthful data sharing, Nix and Kantarciouglu [112] model the distributed data mining scenario as an incomplete information game and propose two incentive compatible mechanisms. The first mechanism, which is designed for non-cooperative game, is a Vickrey-Clarke-Groves (VCG) mechanism. The VCG mechanism can encourage truthful data sharing for the risk-averse data provider, and can give a close approximation that encourages minimal deviation from the true data for the risk-neutral data provider. The second mechanism, which is designed for cooperative game, is based on the Shapley value. When data providers form multiple coalitions, this mechanism can create incentives for entire groups of providers to truthfully reveal their data. The practical viability of these two mechanisms have been tested on three data mining models, namely naïve Bayesian classification, decision tree classification, and support vector machine classification. In their later work [113], Nix and Kantarciouglu investigate what kind of privacy-preserving data analysis (PPDA) techniques can be implemented in a way that participating parties have the incentive to provide their true private inputs upon engaging in the corresponding SMC protocols. Under the assumption that participating parties prefer to learn the data analysis result correctly and if possible exclusively, the study shows that several important PPDA tasks including privacy-preserving association rule mining, privacy-preserving naïve Bayesian classification and privacy-preserving decision tree classification are incentive driven. Based on Nix and Kantarcioglu's work, Panoui et al. [114] employ the VCG mechanism to achieve privacy preserving collaborative classification. They consider three types of strategies that a data provider can choose: providing true data, providing perturbed data, or providing randomized data. They show that the use of the VCG mechanism can lead to high accuracy of the data mining task, and meantime data providers are allowed to provide perturbed data, which means privacy of data providers can be preserved.

2) PRIVACY AUCTIONS

Aiming at providing support for some specific data mining task, the data collector may ask data providers to provide their sensitive data. The data provider will suffer a loss in privacy if he decides to hand over his sensitive data. In order to motivate data providers to participate in the task, the data collector needs to pay monetary incentives to data providers to compensate their privacy loss. Since different data providers assign different values to their privacy, it is natural for data collector to consider buying private data using an auction. In other words, the data provider can sell his privacy at auction. Ghosh and Roth [115] initiate the study of *privacy auction* in a setting where n individuals selling their binary data to a data analyst. Each individual possesses a private bit $b_i \in \{0, 1\}$ representing his sensitive information (e.g. whether the individual has some embarrassing disease), and reports a cost function c_i to the data analyst who wants to estimate the sum of bits $\sum_{i=1}^{n} b_i$. Differential privacy [80] is employed to quantify the privacy cost: $c_i(\varepsilon) = v_i \cdot \varepsilon$, where v_i is a privately known parameter representing individual's value for privacy, and ε is the parameter of differential privacy. The cost function determines the individual's privacy loss when his private bit b_i is used in an ε-differentially private manner. The compensation (i.e. payment) that an individual can get from the data analyst is determined by a mechanism which takes the cost parameters $\mathbf{v} = (v_1, \ldots, v_n)$ and a collection of private bit values $\mathbf{b} = (b_1, \ldots, b_n)$ as input. In an attempt to maximize his payment, an individual may misreport his value for privacy (i.e. v_i), thus the data collector needs to design truthful mechanisms that can incentivize individuals to report their true privacy cost. Ghosh and Roth study the mechanism design problem for two models, namely insensitive value model and sensitive value model. Insensitive value model considers the privacy cost only incurred by b_i and ignores the potential loss due to the implicit correlations between v_i and b_i. It is shown that truthful mechanism can be derived to help the data analyst achieve a desired trade-off between the accuracy of the estimate and the cost of payments. While the sensitive value model considers that the reported value for privacy also incurs a cost. The study shows that generally, it is impossible to derive truthful mechanisms that can compensate individuals for their privacy loss resulting from the unknown correlation between the private data b_i and the privacy valuation v_i.

To circumvent the impossibility of sensitive value model, Fleischer and Lyu [116] model the correlation between b_i and v_i by assuming that individual's private bit b_i determines a distribution from a set of accurate and publicly known distributions, and the privacy value v_i is drawn from that distribution. Based on this assumption, they design an approximately optimal truthful mechanisms that can produce accurate estimate and protect privacy of both the data (i.e. b_i) and cost (i.e. v_i), when priors of the aforementioned distributions are known. In [117], Ligett and Roth propose a different mechanism which makes no Bayesian assumptions about the distributions of the cost functions. Instead, they assume that the data analyst can randomly approach an individual and make a take-it-or-leave-it offer composing of the payment and differential privacy parameters. The proposed mechanism consists of two algorithms. The first algorithm makes an offer to an individual and receives a binary participation decision. The second algorithm computes an statistic over the private data provided by participating individuals. Nissim et al. [118] bypass the impossibility by assuming that individuals have monotonic privacy valuations, which captures common contexts where certain values for private data are expected to lead to higher valuations for privacy. They develop mechanisms that can incentivize individuals whose privacy valuations are not too large to report their truthful privacy valuations, and output accurate estimations of the sum of private bits, if there are not too many individuals with too-large privacy valuations. The main idea behind the proposed mechanism is to treat the private bit b_i as 0 for all individuals who value privacy too much.



(a) (b)

FIGURE 16. Privacy auction. (a) data provider makes a bid (privacy valuation v_i); (b) data collector makes a bid (price willing to pay for the data).

Above studies explore mechanisms for privacy auctions mainly from the perspective of the "buyer", that is, the data providers report their bids (privacy valuations) to the data analyst and the data analyst determine payments to data providers (see Fig. 16)(a). In [119], Riederer et al. study the mechanisms from the seller's perspective. They consider a setting where online users put up sales of their personal information, and information aggregators place bids to gain access to the corresponding user's information (see Fig. 16)(b). They propose a mechanism called Transactional Privacy (TP) that can help users decide what and how much information the aggregators should obtain. This mechanism is based on auction mechanism called the *exponential mechanism* which has been shown to be truthful and can bring approximate optimal revenue for the seller (users in this case). Riederer et al. show that TP can be efficiently implemented when there is a trusted third party. The third party runs an auction where aggregators bid for user's information, computes payments to users, and reports to the user about aggregators that received his information. With the proposed mechanism, users can take back control of their personal information.

VII. NON-TECHNICAL SOLUTIONS TO PRIVACY PROTECTION

In above sections, we mainly explore technical solutions to the privacy issues in data mining. However, the frequently happening information security incidents remind us that non-technical solutions, such as laws, regulations and industrial conventions, are also of great necessity for ensuring the security of sensitive information.

Legislation on privacy protection has always been a prime concern of people. Many countries have established laws to regulate the acts involving personal information. For example, in the U.S., people's right to privacy is regulated by *the Privacy Act of 1974*[15] and various states laws. The European Commission has released a proposal called *General Data Protection Regulation* in 2012, aiming at unifying data protection within the European Union. Despite the many laws and regulations, nowadays the definition of the right to privacy and the boundary of "legitimate" practice on personal data are still vague. For example, the exposure of the US surveillance data mining program *PRISM*[16] has triggered extensive discussions and debates in 2013. One thing we could learn from this incident is that there is an urgent need to improve current legislation to reconcile the conflict between individual's right to privacy and the government's need for accessing personal information for national security.

Besides laws and regulations, industry conventions are also required. Agreement between different organizations on how personal data should be collected, stored and analyzed, can help to build a privacy-safe environment for data mining applications. Also, it is necessary to enhance propaganda and education to increase public awareness of information security.

VIII. FUTURE RESEARCH DIRECTIONS

In previous sections, we have reviewed different approaches to privacy protection for different user roles. Although we have already pointed out some problems that need to be further investigated for each user role (see Section II-C, Section III-C, Section IV-C and Section V-C), here in this section, we highlight some of the problems and consider them to be the major directions of future research.

[15] *http://en.wikipedia.org/wiki/Privacy_Act_of_1974*
[16] *http://en.wikipedia.org/wiki/PRISM_(surveillance_program)*

A. PERSONALIZED PRIVACY PRESERVING

PPDP and PPDM provide methods to explore the utility of data while preserving privacy. However, most current studies only manage to achieve privacy preserving in a statistical sense. Considering that the definition of privacy is essentially personalized, developing methods that can support personalized privacy preserving is an important direction for the study of PPDP and PPDM. As mentioned in Section III-C, some researchers have already investigated the issue of personalized anonymization, but most current studies are still in the theoretical stage. Developing practical personalized anonymization methods is in urgent need. Besides, introducing personalized privacy into other types of PPDP/PPDM algorithms is also required. In addition, since complex socioeconomic and psychological factors are involved, quantifying individual's privacy preference is still an open question which expects more exploration.

B. DATA CUSTOMIZATION

In Section IV-B.1 we have discussed that in order to hiding sensitive mining results, we can employ inverse data mining such as inverse frequent set mining to generate data that cannot expose sensitive information. By inverse data mining, we can "customize" the data to get the desired mining result. Alexandra et al. [120] introduced a concept called *reverse data management* (RDM) which is similar to our specification for inverse data mining. RDM consists of problems where one needs to compute a database input, or modify an existing database input, in order to achieve a desired effect in the output. RDM covers many database problems such as inversion mappings, provenance, data generation, view update, constraint-based repair, etc. We may consider RDM to be a family of data customization methods by which we can get the desired data from which sensitive information cannot be discovered. In a word, data customization can be seen as the inverse process of ordinary data processing. Whenever we have explicit requirements for the outcome of data processing, we may resort to data customization. Exploring ways to solve the inverse problem is an important task for future study.

C. PROVENANCE FOR DATA MINING

The complete process of data mining consists of multiple phases such as data collection, data preprocessing, data mining, analyzing the extracted information to get knowledge, and applying the knowledge. This process can be seen as an evolvement of data. If the provenance information corresponding to every phase in the process, such as the ownership of data and how the data is processed, can be clearly recorded, it will be much easier to find the origins of security incidents such as sensitive data breach and the distortion of sensitive information. We may say that provenance provides us a way to monitor the process of data mining and the use of mining result. Therefore, techniques and mechanisms that can support provenance in data mining context should receive more attention in future study.

Glavic et al. [121] have discussed how traditional notions of provenance translated to data mining. They identified the need for new types of provenance that can be used to better interpret data mining results. In the context of privacy protection, we are more concerned with how to use provenance to better understand why and how "abnormal" mining result, e.g. result containing sensitive information or false result, appears. Different from provenance approaches that we have reviewed in Section V-B.1, approaches for data mining provenance are closely related to the mining algorithm. Therefore, it is necessary to develop new provenance models to specify what kind of provenance information is required and how to present, store, acquire and utilize the provenance information.

IX. CONCLUSION

How to protect sensitive information from the security threats brought by data mining has become a hot topic in recent years. In this paper we review the privacy issues related to data mining by using a user-role based methodology. We differentiate four different user roles that are commonly involved in data mining applications, i.e. data provider, data collector, data miner and decision maker. Each user role has its own privacy concerns, hence the privacy-preserving approaches adopted by one user role are generally different from those adopted by others:

- For data provider, his privacy-preserving objective is to effectively control the amount of sensitive data revealed to others. To achieve this goal, he can utilize security tools to limit other's access to his data, sell his data at auction to get enough compensations for privacy loss, or falsify his data to hide his true identity.
- For data collector, his privacy-preserving objective is to release useful data to data miners without disclosing data providers' identities and sensitive information about them. To achieve this goal, he needs to develop proper privacy models to quantify the possible loss of privacy under different attacks, and apply anonymization techniques to the data.
- For data miner, his privacy-preserving objective is to get correct data mining results while keep sensitive information undisclosed either in the process of data mining or in the mining results. To achieve this goal, he can choose a proper method to modify the data before certain mining algorithms are applied to, or utilize secure computation protocols to ensure the safety of private data and sensitive information contained in the learned model.
- For decision maker, his privacy-preserving objective is to make a correct judgement about the credibility of the data mining results he's got. To achieve this goal, he can utilize provenance techniques to trace back the history of the received information, or build classifier to discriminate true information from false information.

To achieve the privacy-preserving goals of different users roles, various methods from different research fields are required. We have reviewed recent progress in related studies, and discussed problems awaiting to be further investigated. We hope that the review presented in this paper can offer researchers different insights into the issue of privacy-preserving data mining, and promote the exploration of new solutions to the security of sensitive information.

REFERENCES

[1] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. San Mateo, CA, USA: Morgan Kaufmann, 2006.
[2] L. Brankovic and V. Estivill-Castro, "Privacy issues in knowledge discovery and data mining," in *Proc. Austral. Inst. Comput. Ethics Conf.*, 1999, pp. 89–99.
[3] R. Agrawal and R. Srikant, "Privacy-preserving data mining," *ACM SIGMOD Rec.*, vol. 29, no. 2, pp. 439–450, 2000.
[4] Y. Lindell and B. Pinkas, "Privacy preserving data mining," in *Advances in Cryptology*. Berlin, Germany: Springer-Verlag, 2000, pp. 36–54.
[5] C. C. Aggarwal and S. Y. Philip, *A General Survey of Privacy-Preserving Data Mining Models and Algorithms*. New York, NY, USA: Springer-Verlag, 2008.
[6] M. B. Malik, M. A. Ghazi, and R. Ali, "Privacy preserving data mining techniques: Current scenario and future prospects," in *Proc. 3rd Int. Conf. Comput. Commun. Technol. (ICCCT)*, Nov. 2012, pp. 26–32.
[7] S. Matwin, "Privacy-preserving data mining techniques: Survey and challenges," in *Discrimination and Privacy in the Information Society*. Berlin, Germany: Springer-Verlag, 2013, pp. 209–221.
[8] E. Rasmusen, *Games and Information: An Introduction to Game Theory*, vol. 2. Cambridge, MA, USA: Blackwell, 1994.
[9] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati, "Microdata protection," in *Secure Data Management in Decentralized Systems*. New York, NY, USA: Springer-Verlag, 2007, pp. 291–321.
[10] O. Tene and J. Polenetsky, "To track or 'do not track': Advancing transparency and individual control in online behavioral advertising," *Minnesota J. Law, Sci. Technol.*, no. 1, pp. 281–357, 2012.
[11] R. T. Fielding and D. Singer. (2014). *Tracking Preference Expression (DNT). W3C Working Draft*. [Online]. Available: http://www.w3.org/TR/2014/WD-tracking-dnt-20140128/
[12] R. Gibbons, *A Primer in Game Theory*. Hertfordshire, U.K.: Harvester Wheatsheaf, 1992.
[13] D. C. Parkes, "Iterative combinatorial auctions: Achieving economic and computational efficiency," Ph.D. dissertation, Univ. Pennsylvania, Philadelphia, PA, USA, 2001.
[14] S. Carter, "Techniques to pollute electronic profiling," U.S. Patent 11/257 614, Apr. 26, 2007. [Online]. Available: https://www.google.com/patents/US20070094738
[15] Verizon Communications Inc. (2013). *2013 Data Breach Investigations Report*. [Online]. Available: http://www.verizonenterprise.com/resources/reports/rp_data-breach-investigations-report-2013_en_xg.pdf
[16] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2008, pp. 111–125.
[17] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Comput. Surv.*, vol. 42, no. 4, Jun. 2010, Art. ID 14.
[18] R. C.-W. Wong and A. W.-C. Fu, "Privacy-preserving data publishing: An overview," *Synthesis Lectures Data Manage.*, vol. 2, no. 1, pp. 1–138, 2010.
[19] L. Sweeney, "k-anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, 2002.
[20] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in *Proc. 21st Int. Conf. Data Eng. (ICDE)*, Apr. 2005, pp. 217–228.
[21] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in *Proc. 22nd Int. Conf. Data Eng. (ICDE)*, Apr. 2006, p. 25.
[22] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. W.-C. Fu, "Utility-based anonymization for privacy preservation with less information loss," *ACM SIGKDD Explorations Newslett.*, vol. 8, no. 2, pp. 21–30, 2006.
[23] A. Gionis and T. Tassa, "k-anonymization with minimal loss of information," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 2, pp. 206–219, Feb. 2009.
[24] B. Zhou, J. Pei, and W. Luk, "A brief survey on anonymization techniques for privacy preserving publishing of social network data," *ACM SIGKDD Explorations Newslett.*, vol. 10, no. 2, pp. 12–22, 2008.
[25] X. Wu, X. Ying, K. Liu, and L. Chen, "A survey of privacy-preservation of graphs and social networks," in *Managing and Mining Graph Data*. New York, NY, USA: Springer-Verlag, 2010, pp. 421–453.
[26] S. Sharma, P. Gupta, and V. Bhatnagar, "Anonymisation in social network: A literature survey and classification," *Int. J. Soc. Netw. Mining*, vol. 1, no. 1, pp. 51–66, 2012.
[27] W. Peng, F. Li, X. Zou, and J. Wu, "A two-stage deanonymization attack against anonymized social networks," *IEEE Trans. Comput.*, vol, 63, no. 2, pp. 290–303, 2014.
[28] T. Zhu, S. Wang, X. Li, Z. Zhou, and R. Zhang, "Structural attack to anonymous graph of social networks," *Math. Problems Eng.*, vol. 2013, Oct. 2013, Art. ID 237024.
[29] C. Sun, P. S. Yu, X. Kong, and Y. Fu. (2013). "Privacy preserving social network publication against mutual friend attacks." [Online]. Available: http://arxiv.org/abs/1401.3201
[30] C.-H. Tai, P. S. Yu, D.-N. Yang, and M.-S. Chen, "Privacy-preserving social network publication against friendship attacks," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 1262–1270.
[31] C.-H. Tai, P. S. Yu, D.-N. Yang, and M.-S. Chen, "Structural diversity for resisting community identification in published social networks," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 235–252, Nov. 2013.
[32] M. I. Hafez Ninggal and J. Abawajy, "Attack vector analysis and privacy-preserving social network data publishing," in *Proc. IEEE 10th Int. Conf. Trust, Secur. Privacy Comput. Commun. (TrustCom)*, Nov. 2011, pp. 847–852.
[33] Y. Wang, L. Xie, B. Zheng, and K. C. K. Lee, "High utility k-anonymization for social network publishing," *Knowl. Inf. Syst.*, vol. 36, no. 1, pp. 1–29, 2013.
[34] N. Medforth and K. Wang, "Privacy risk in graph stream publishing for social network data," in *Proc. IEEE 11th Int. Conf. Data Mining (ICDM)*, Dec. 2011, pp. 437–446.
[35] C.-H. Tai, P.-J. Tseng, P. S. Yu, and M.-S. Chen, "Identity protection in sequential releases of dynamic networks," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 3, pp. 635–651, Mar. 2014.
[36] G. Ghinita, *Privacy for Location-Based Services* (Synthesis Lectures on Information Security, Privacy, and Trust). San Rafael, CA, USA: Morgan & Claypool, 2013.
[37] M. Wernke, P. Skvortsov, F. Dürr, and K. Rothermel, "A classification of location privacy attacks and approaches," *Pers. Ubiquitous Comput.*, vol. 18, no. 1, pp. 163–175, Jan. 2014.
[38] M. Terrovitis and N. Mamoulis, "Privacy preservation in the publication of trajectories," in *Proc. 9th Int. Conf. Mobile Data Manage. (MDM)*, 2008, pp. 65–72.
[39] M. E. Nergiz, M. Atzori, and Y. Saygin, "Towards trajectory anonymization: A generalization-based approach," in *Proc. SIGSPATIAL ACM GIS Int. Workshop Secur. Privacy GIS LBS*, 2008, pp. 52–61.
[40] O. Abul, F. Bonchi, and M. Nanni, "Never walk alone: Uncertainty for anonymity in moving objects databases," in *Proc. IEEE 24th Int. Conf. Data Eng. (ICDE)*, Apr. 2008, pp. 376–385.
[41] R. Yarovoy, F. Bonchi, L. V. S. Lakshmanan, and W. H. Wang, "Anonymizing moving objects: How to hide a MOB in a crowd?" in *Proc. 12th Int. Conf. Extending Database Technol., Adv. Database Technol.*, 2009, pp. 72–83.
[42] R. Chen, B. C. M. Fung, N. Mohammed, B. C. Desai, and K. Wang, "Privacy-preserving trajectory data publishing by local suppression," *Inf. Sci.*, vol. 231, pp. 83–97, May 2013.
[43] M. Ghasemzadeh, B. C. M. Fung, R. Chen, and A. Awasthi, "Anonymizing trajectory data for passenger flow analysis," *Transp. Res. C, Emerg. Technol.*, vol. 39, pp. 63–79, Feb. 2014.
[44] A. E. Cicek, M. E. Nergiz, and Y. Saygin, "Ensuring location diversity in privacy-preserving spatio-temporal data publishing," *VLDB J.*, vol. 23, no. 4, pp. 1–17, 2013.
[45] G. Poulis, S. Skiadopoulos, G. Loukides, and A. Gkoulalas-Divanis, "Distance-based kˆ m-anonymization of trajectory data," in *Proc. IEEE 14th Int. Conf. Mobile Data Manage. (MDM)*, vol. 2. Jun. 2013, pp. 57–62.
[46] F. Bonchi, L. V. S. Lakshmanan, and H. W. Wang, "Trajectory anonymity in publishing personal mobility data," *ACM SIGKDD Explorations Newslett.*, vol. 13, no. 1, pp. 30–42, Jun. 2011.
[47] X. Xiao and Y. Tao, "Personalized privacy preservation," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2006, pp. 229–240.
[48] K. Qing-Jiang, W. Xiao-Hao, and Z. Jun, "The (p, α, k) anonymity model for privacy protection of personal information in the social networks," in *Proc. 6th IEEE Joint Int. Inf. Technol. Artif. Intell. Conf. (ITAIC)*, vol. 2. Aug. 2011, pp. 420–423.

[49] B. Wang and J. Yang, "Personalized (α, k)-anonymity algorithm based on entropy classification," *J. Comput. Inf. Syst.*, vol. 8, no. 1, pp. 259–266, 2012.
[50] Y. Xua, X. Qin, Z. Yang, Y. Yang, and K. Li, "A personalized k-anonymity privacy preserving method," *J. Inf. Comput. Sci.*, vol. 10, no. 1, pp. 139–155, 2013.
[51] S. Yang, L. Lijie, Z. Jianpei, and Y. Jing, "Method for individualized privacy preservation," *Int. J. Secur. Appl.*, vol. 7, no. 6, p. 109, 2013.
[52] A. Halevy, A. Rajaraman, and J. Ordille, "Data integration: The teenage years," in *Proc. 32nd Int. Conf. Very Large Data Bases (VLDB)*, 2006, pp. 9–16.
[53] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis, "State-of-the-art in privacy preserving data mining," *ACM SIGMOD Rec.*, vol. 33, no. 1, pp. 50–57, 2004.
[54] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y. Zhu, "Tools for privacy preserving distributed data mining," *ACM SIGKDD Explorations Newslett.*, vol. 4, no. 2, pp. 28–34, 2002.
[55] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proc. ACM SIGMOD Rec.*, 1993, vol. 22, no. 2, pp. 207–216.
[56] V. S. Verykios, "Association rule hiding methods," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 3, no. 1, pp. 28–36, 2013.
[57] K. Sathiyapriya and G. S. Sadasivam, "A survey on privacy preserving association rule mining," *Int. J. Data Mining Knowl. Manage. Process*, vol. 3, no. 2, p. 119, 2013.
[58] D. Jain, P. Khatri, R. Soni, and B. K. Chaurasia, "Hiding sensitive association rules without altering the support of sensitive item(s)," in *Proc. 2nd Int. Conf. Adv. Comput. Sci. Inf. Technol. Netw. Commun.*, 2012, pp. 500–509.
[59] J.-M. Zhu, N. Zhang, and Z.-Y. Li, "A new privacy preserving association rule mining algorithm based on hybrid partial hiding strategy," *Cybern. Inf. Technol.*, vol. 13, pp. 41–50, Dec. 2013.
[60] H. Q. Le, S. Arch-Int, H. X. Nguyen, and N. Arch-Int, "Association rule hiding in risk management for retail supply chain collaboration," *Comput. Ind.*, vol. 64, no. 7, pp. 776–784, Sep. 2013.
[61] M. N. Dehkordi, "A novel association rule hiding approach in OLAP data cubes," *Indian J. Sci. Technol.*, vol. 6, no. 2, pp. 4063–4075, 2013.
[62] J. Bonam, A. R. Reddy, and G. Kalyani, "Privacy preserving in association rule mining by data distortion using PSO," in *Proc. ICT Critical Infrastruct., Proc. 48th Annu. Conv. Comput. Soc. India*, vol. 2. 2014, pp. 551–558.
[63] C. N. Modi, U. P. Rao, and D. R. Patel, "Maintaining privacy and data quality in privacy preserving association rule mining," in *Proc. Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT)*, Jul. 2010, pp. 1–6.
[64] N. R. Radadiya, N. B. Prajapati, and K. H. Shah, "Privacy preserving in association rule mining," *Int. J. Adv. Innovative Res.*, vol. 2, no. 4, pp. 203–213, 2013.
[65] K. Pathak, N. S. Chaudhari, and A. Tiwari, "Privacy preserving association rule mining by introducing concept of impact factor," in *Proc. 7th IEEE Conf. Ind. Electron. Appl. (ICIEA)*, Jul. 2012, pp. 1458–1461.
[66] T. Mielikäinen, "On inverse frequent set mining," in *Proc. 2nd Workshop Privacy Preserving Data Mining*, 2003, pp. 18–23.
[67] X. Chen and M. Orlowska, "A further study on inverse frequent set mining," in *Proc. 1st Int. Conf. Adv. Data Mining Appl.*, 2005, pp. 753–760.
[68] Y. Guo, "Reconstruction-based association rule hiding," in *Proc. SIGMOD Ph. D. Workshop Innovative Database Res.*, 2007, pp. 51–56.
[69] Y. Wang and X. Wu, "Approximate inverse frequent itemset mining: Privacy, complexity, and approximation," in *Proc. 5th IEEE Int. Conf. Data Mining*, Nov. 2005, p. 8.
[70] Y. Guo, Y. Tong, S. Tang, and D. Yang, "A FP-tree-based method for inverse frequent set mining," in *Proc. 23rd Brit. Nat. Conf. Flexible Efficient Inf. Handling*, 2006, pp. 152–163.
[71] J. Dowd, S. Xu, and W. Zhang, "Privacy-preserving decision tree mining based on random substitutions," in *Proc. Int. Conf. Emerg. Trends Inf. Commun. Security*, 2006, pp. 145–159.
[72] J. Brickell and V. Shmatikov, "Privacy-preserving classifier learning," in *Proc. 13th Int. Conf. Financial Cryptogr. Data Security*, 2009, pp. 128–147.
[73] P. K. Fong and J. H. Weber-Jahnke, "Privacy preserving decision tree learning using unrealized data sets," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 2, pp. 353–364, Feb. 2012.
[74] M. A. Sheela and K. Vijayalakshmi, "A novel privacy preserving decision tree induction," in *Proc. IEEE Conf. Inf. Commun. Technol. (ICT)*, Apr. 2013, pp. 1075–1079.
[75] O. Goldreich. (2002). *Secure Multi-Party Computation*. [Online]. Available: http://www.wisdom.weizmann.ac.il/~oded/PS/prot.ps
[76] J. Vaidya, M. Kantarcıoğlu, and C. Clifton, "Privacy-preserving Naïve Bayes classification," *Int. J. Very Large Data Bases*, vol. 17, no. 4, pp. 879–898, 2008.
[77] M. E. Skarkala, M. Maragoudakis, S. Gritzalis, and L. Mitrou, "Privacy preserving tree augmented Naïve Bayesian multi-party implementation on horizontally partitioned databases," in *Proc. 8th Int. Conf. Trust, Privacy, Security Digit. Bus.*, 2011, pp. 62–73.
[78] F. Zheng and G. I. Webb, "Tree augmented Naïve Bayes," in *Proc. Encyclopedia Mach. Learn.*, 2010, pp. 990–991.
[79] J. Vaidya, B. Shafiq, A. Basu, and Y. Hong, "Differentially private Naïve Bayes classification," in *Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. (WI) Intell. Agent Technol. (IAT)*, vol. 1. Nov. 2013, pp. 571–576.
[80] C. Dwork, "Differential privacy," in *Proc. 33rd Int. Conf. Autom., Lang., Program.*, 2006, pp. 1–12.
[81] J. Vaidya, H. Yu, and X. Jiang, "Privacy-preserving SVM classification," *Knowl. Inf. Syst.*, vol. 14, no. 2, pp. 161–178, 2008.
[82] H. Xia, Y. Fu, J. Zhou, and Y. Fang, "Privacy-preserving SVM classifier with hyperbolic tangent kernel," *J. Comput. Inf. Syst.*, vol. 6, no. 5, pp. 1415–1420, 2010.
[83] K.-P. Lin and M.-S. Chen, "On the design and analysis of the privacy-preserving SVM classifier," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 11, pp. 1704–1717, Nov. 2011.
[84] R. R. Rajalaxmi and A. M. Natarajan, "An effective data transformation approach for privacy preserving clustering," *J. Comput. Sci.*, vol. 4, no. 4, pp. 320–326, 2008.
[85] M. N. Lakshmi and K. S. Rani, "SVD based data transformation methods for privacy preserving clustering," *Int. J. Comput. Appl.*, vol. 78, no. 3, pp. 39–43, 2013.
[86] J. Vaidya and C. Clifton, "Privacy-preserving *k*-means clustering over vertically partitioned data," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2003, pp. 206–215.
[87] S. Jha, L. Kruger, and P. McDaniel, "Privacy preserving clustering," in *Proc. 10th Eur. Symp. Res. Comput. Security (ESORICS)*, 2005, pp. 397–417.
[88] R. Akhter, R. J. Chowdhury, K. Emura, T. Islam, M. S. Rahman, and N. Rubaiyat, "Privacy-preserving two-party *k*-means clustering in malicious model," in *Proc. IEEE 37th Annu. Comput. Softw. Appl. Conf. Workshops (COMPSACW)*, Jul. 2013, pp. 121–126.
[89] X. Yi and Y. Zhang, "Equally contributory privacy-preserving *k*-means clustering over vertically partitioned data," *Inf. Syst.*, vol. 38, no. 1, pp. 97–107, 2013.
[90] I. De and A. Tripathy, "A secure two party hierarchical clustering approach for vertically partitioned data set with accuracy measure," in *Proc. 2nd Int. Symp. Recent Adv. Intell. Informat.*, 2014, pp. 153–162.
[91] Y. L. Simmhan, B. Plale, and D. Gannon, "A survey of data provenance in e-science," *ACM Sigmod Rec.*, vol. 34, no. 3, pp. 31–36, 2005.
[92] B. Glavic and K. R. Dittrich, "Data provenance: A categorization of existing approaches," in *Proc. BTW*, 2007, vol. 7, no. 12, pp. 227–241.
[93] S. B. Davidson and J. Freire, "Provenance and scientific workflows: Challenges and opportunities," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2008, pp. 1345–1350.
[94] O. Hartig, "Provenance information in the web of data," in *Proc. LDOW*, 2009. [Online]. Available: http://ceur-ws.org/Vol-538/ldow2009_paper18.pdf
[95] L. Moreau, "The foundations for provenance on the web," *Found. Trends Web Sci.*, vol. 2, no. 2–3, pp. 99–241, 2010.
[96] G. Barbier, Z. Feng, P. Gundecha, and H. Liu, "Provenance data in social media," *Synth. Lectures Data Mining Knowl. Discovery*, vol. 4, no. 1, pp. 1–84, 2013.
[97] M. Tudjman and N. Mikelic, "Information science: Science about information, misinformation and disinformation," in *Proc. Inf. Sci.+Inf. Technol. Edu.*, 2003, pp. 1513–1527.
[98] M. J. Metzger, "Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 58, no. 13, pp. 2078–2091, 2007.
[99] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on Twitter," in *Proc. 20th Int. Conf. World Wide Web*, 2011, pp. 675–684.
[100] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei, "Rumor has it: Identifying misinformation in microblogs," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2011, pp. 1589–1599.
[101] F. Yang, Y. Liu, X. Yu, and M. Yang, "Automatic detection of rumor on Sina Weibo," in *Proc. ACM SIGKDD Workshop Mining Data Semantics*, 2012, Art. ID 13.
[102] S. Sun, H. Liu, J. He, and X. Du, "Detecting event rumors on Sina Weibo automatically," in *Proc. Web Technol. Appl.*, 2013, pp. 120–131.

[103] R. K. Adl, M. Askari, K. Barker, and R. Safavi-Naini, "Privacy consensus in anonymization systems via game theory," in *Proc. 26th Annu. Data Appl. Security Privacy*, 2012, pp. 74–89.
[104] R. Karimi Adl, K. Barker, and J. Denzinger, "A negotiation game: Establishing stable privacy policies for aggregate reasoning," Dept. Comput. Sci., Univ. Calgary, Calgary, AB, Canada, Tech. Rep., Oct. 2012. [Online]. Available: The paper is available at http://dspace.ucalgary.ca/jspui/bitstream/1880/49282/1/2012-1023-06.pdf
[105] H. Kargupta, K. Das, and K. Liu, "Multi-party, privacy-preserving distributed data mining using a game theoretic framework," in *Proc. 11th Eur. Conf. Principles Pract. Knowl. Discovery Databases (PKDD)*, 2007, pp. 523–531.
[106] A. Miyaji and M. S. Rahman, "Privacy-preserving data mining: A game-theoretic approach," in *Proc. 25th Data Appl. Security Privacy*, 2011, pp. 186–200.
[107] X. Ge, L. Yan, J. Zhu, and W. Shi, "Privacy-preserving distributed association rule mining based on the secret sharing technique," in *Proc. 2nd Int. Conf. Softw. Eng. Data Mining (SEDM)*, Jun. 2010, pp. 345–350.
[108] N. R. Nanavati and D. C. Jinwala, "A novel privacy preserving game theoretic repeated rational secret sharing scheme for distributed data mining," vol. 91, 2013. [Online]. Available: http://www.researchgate.net/ publication/256765823_ A_NOVEL_PRIVACY_PRESERVING_ GAME_THEORETIC_ REPEATED_RATIONAL_ SECRET_SHARING_SCHEME_ FOR_DISTRIBUTED_DATA_MINING
[109] M. Halkidi and I. Koutsopoulos, "A game theoretic framework for data privacy preservation in recommender systems," in *Proc. Mach. Learn. Knowl. Discovery Databases*, 2011, pp. 629–644.
[110] S. Ioannidis and P. Loiseau, "Linear regression as a non-cooperative game," in *Proc. Web Internet Econ.*, 2013, pp. 277–290.
[111] S. L. Chakravarthy, V. V. Kumari, and C. Sarojini, "A coalitional game theoretic mechanism for privacy preserving publishing based on k-anonymity," *Proc. Technol.*, vol. 6, pp. 889–896, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2212017312006536
[112] R. Nix and M. Kantarciouglu, "Incentive compatible privacy-preserving distributed classification," *IEEE Trans. Dependable Secure Comput.*, vol. 9, no. 4, pp. 451–462, Jul. 2012.
[113] M. Kantarcioglu and W. Jiang, "Incentive compatible privacy-preserving data analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 6, pp. 1323–1335, Jun. 2013.
[114] A. Panoui, S. Lambotharan, and R. C.-W. Phan, "Vickrey–Clarke–Groves for privacy-preserving collaborative classification," in *Proc. Fed. Conf. Comput. Sci. Inf. Syst. (FedCSIS)*, Sep. 2013, pp. 123–128.
[115] A. Ghosh and A. Roth, "Selling privacy at auction," in *Proc. 12th ACM Conf. Electron. Commerce*, 2011, pp. 199–208.
[116] L. K. Fleischer and Y.-H. Lyu, "Approximately optimal auctions for selling privacy when costs are correlated with data," in *Proc. 13th ACM Conf. Electron. Commerce*, 2012, pp. 568–585.
[117] K. Ligett and A. Roth, "Take it or leave it: Running a survey when privacy comes at a cost," in *Proc. 8th Internet Netw. Econ.*, 2012, pp. 378–391.
[118] K. Nissim, S. Vadhan, and D. Xiao, "Redrawing the boundaries on purchasing data from privacy-sensitive individuals," in *Proc. 5th Conf. Innov. Theoretical Comput. Sci.*, 2014, pp. 411–422.
[119] C. Riederer, V. Erramilli, A. Chaintreau, B. Krishnamurthy, and P. Rodriguez, "For sale: Your data: By: You," in *Proc. 10th ACM Workshop Hot Topics Netw.*, 2011, Art. ID 13.
[120] A. Meliou, W. Gatterbauer, and D. Suciu, "Reverse data management," in *Proc. VLDB Endowment*, 2011, vol. 4, no. 12. [Online]. Available: http://people.cs.umass.edu/~ameli/projects/reverse-data-management/papers/VLDB2011_vision.pdf
[121] B. Glavic, J. Siddique, P. Andritsos, and R. J. Miller, "Provenance for data mining," in *Proc. 5th USENIX Workshop Theory Pract. Provenance*, 2013, p. 5.

LEI XU received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2008, where she is currently pursuing the Ph.D. degree. Her research interests include privacy issues in data mining, text mining, and game theory.

CHUNXIAO JIANG (S'09–M'13) received the B.S. (Hons.) degree in information engineering from Beihang University, Beijing, China, in 2008, and the Ph.D. (Hons.) degree from Tsinghua University (THU), Beijing, in 2013. From 2011 to 2013, he visited the Signals and Information Group at the Department of Electrical and Computer Engineering, University of Maryland at College Park, College Park, MD, USA, with Prof. K. J. Ray Liu. He currently holds a post-doctoral position with the Department of Electrical Engineering, THU, with Prof. Y. Ren. His research interests include the applications of game theory and queuing theory in wireless communication and networking, and social networks.

Dr. Jiang was a recipient of the Best Paper Award from the IEEE Global Communications Conference in 2013, the Beijing Distinguished Graduated Student Award, the Chinese National Fellowship, and the Tsinghua Outstanding Distinguished Doctoral Dissertation in 2013.

JIAN WANG received the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2006. In 2006, he joined the faculty of Tsinghua University, where he is currently an Associate Professor with the Department of Electronic Engineering. His research interests are in the areas of information security, signal processing in the encrypted domain, and cognitive networks.

JIAN YUAN received the M.S. degree in signals and systems from Southeast University, Nanjing, China, in 1989, and the Ph.D. degree in electrical engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 1998. He is currently a Professor of Electronic Engineering with Tsinghua University, Beijing, China. His main interests are in dynamics of networked systems.

YONG REN received the B.S., M.S., and Ph.D. degrees in electronic engineering from the Harbin Institute of Technology, Harbin, China, in 1984, 1987, and 1994, respectively. He held a post-doctoral position with the Department of Electronics Engineering, Tsinghua University, Beijing, China, from 1995 to 1997, where he is currently a Professor with the Department of Electronics Engineering and the Director of the Complexity Engineered Systems Laboratory. He has authored or co-authored over 100 technical papers in the behavior of computer network, P2P network, and cognitive networks, and holds 12 patents. He has served as a reviewer of the *IEICE Transactions on Communications*, *Digital Signal Processing*, *Chinese Physics Letters*, *Chinese Journal of Electronics*, *Chinese Journal of Computer Science and Technology*, and *Chinese Journal of Aeronautics*. His current research interests include complex systems theory and its applications to the optimization and information sharing of the Internet, Internet of Things and ubiquitous network, cognitive networks, and cyber-physical systems.