IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 18, 2016, accepted May 18, 2016, date of publication June 6, 2016, date of current version June 24, 2016.

Digital Object Identifier 10.1109/ACCESS.2016.2577036

Big Privacy: Challenges and Opportunities of Privacy Study in the Age of Big Data

SHUI YU, (Senior Member, IEEE)
School of Information Technology, Deakin University, Waurn Ponds, VIC 3216, Australia (syu@deakin.edu.au)

This work was supported in part by the National 973 Program under Grant 2014CB340404, in part by the IBM Shared University Research Grant (2015), and in part by the National Natural Science Foundation of China under Grant 61379041.

ABSTRACT One of the biggest concerns of big data is privacy. However, the study on big data privacy is still at a very early stage. We believe the forthcoming solutions and theories of big data privacy root from the in place research output of the privacy discipline. Motivated by these factors, we extensively survey the existing research outputs and achievements of the privacy field in both application and theoretical angles, aiming to pave a solid starting ground for interested readers to address the challenges in the big data case. We first present an overview of the battle ground by defining the roles and operations of privacy systems. Second, we review the milestones of the current two major research categories of privacy: data clustering and privacy frameworks. Third, we discuss the effort of privacy study from the perspectives of different disciplines, respectively. Fourth, the mathematical description, measurement, and modeling on privacy are presented. We summarize the challenges and opportunities of this promising topic at the end of this paper, hoping to shed light on the exciting and almost uncharted land.

INDEX TERMS Big data, privacy, data clustering, differential privacy.

I. INTRODUCTION

Big data is a milestone in the information age, and brings deep impact on human society. Thanks to the dramatic development of information technology, especially the Internet and electronic storage techniques, we are embracing the age of big data, which involves many critical aspects of our society, such as climate [1], biology [2], health [3], and social science [4]. The available big data sets significantly advance our knowledge, services, and productivity across many sectors of our society. For example, a big medical data set can be used to find the best treatment plan for a given patient; a big traffic data set can improve the related traffic control and reduce congestion. The early version of the big data concept appeared in 2001 in the Gartner report by Laney [5], and big data was defined as large and complex data sets that current computing facilities were not able to handle. It is characterized by 3Vs (Volume, Velocity, and Variety). Today, almost every part of our society is expecting to improve itself using big data.

However, privacy protection has become one of the biggest problems with the progress of big data. Human privacy is usually challenged by the development of technology. The record of individuals for tax and draft purpose was a great threat to personal privacy in the 11th century in England, and photographs and yellow page services significantly threatened people's privacy in the late 19th century. Today, human beings can record extraordinary amount of information in various forms, such as photos, video clips, electronic documents, and footprints of web surfing. The easily available modern technologies and tools, e.g., search engines, social networks, hacking packages, especially the data mining and machine learning tools, pose a great challenge to individual privacy in the age of big data.

Recent research indicated that simply anonymized data sets can be easily attacked in terms of privacy. De Montjoye et al. [6] collected a 15 month mobility data set of 1.5 million people. After a simple anonymization operation (removing the obvious identifiers, such as name, home address, phone number, and staff ID), they obtained a data set where the location of an individual was specified hourly with a spatial resolution equal to that given by the carrier's antennas. From the processed data set, they were able to identify a person with 95% accuracy by only four spatial-temporal points. The weakness of simple anonymization was further confirmed by a recent major test [7], in which the authors studied a data set of 3 month credit card transactions of 1.1 million people, and found it again that four spatial-temporal points were sufficient to re-identify 90% of the individuals.

Furthermore, the failure of simple anonymization pushes us to think about the real meaning of identification. As IDs are pervasively used today. Therefore, a straightforward idea

for privacy protection is removing IDs. However, the reality has demonstrated that this method does not work. It is obvious that we were born without IDs (e.g., our names, driver licence numbers, student IDs, and staff IDs), which were given by different organizations. Essentially, an ID is a representation of a set of features of a described object. For example, we can differentiate Alice from Bob in a crowd based on their faces, build, and other physical features without knowing their IDs. In other words, we identify a person based on his or her various characteristics, rather than IDs. This may somehow explain why the simple anonymization processing does not work.

Recently, two world leading researchers in machine learning pointed out that big data is one of the drivers for the dramatic development of machine learning algorithms. However, the advance of the learning technology also greatly threatens privacy of individuals [8]. The mining community realized the privacy challenge decades ago, and extensive effort has been invested in privacy protection, such as privacy aware learning [9], [10], privacy preserving data publishing (PPDP) [11].

It is extraordinarily challenging on privacy protection in the age of big data. First of all, privacy is a subjective concept, it is hard to reach a clear and global definition or measurement on privacy. Secondly, the fast development of various technologies, especially the data mining and machine learning techniques, are desperate threat to privacy. A reliable privacy protection mechanism today may be easily breached tomorrow with the advancement of related technologies. We fully believe the solutions for big data privacy root from the existing research outputs. As a result, it is necessary to understand the outputs of privacy study in the non-big data circumstance, and design new solutions and algorithms to serve the challenges of the big data cases.

We can simply separate the privacy study into to two categories: *content privacy* and *interaction privacy*. In the former class, attackers may identify an individual from an anonymized or encrypted data set given some knowledge about the victims. For example, an attacker knows Alice went to a few shops at different time intervals, he may use this information to extract all the events of Alice from an annoymized data set of credit card records at a state or national level database. Another example is that we can identify the individuals using voice fingerprint from the record of a confidential meeting record under the condition that we have labeled voice records of the speakers from other sources. In the second category, we more care about privacy protection against eavesdroppers on user interactions on a given content, such as user behavior, habit, and other ''fingerprint'' in accessing services. For instance, by monitoring the victim's encrypted web traffic, an attacker can confirm whether the victim is accessing a sensitive web site or not [12]; by monitoring user behavior at application level, we can identify a user from a set of anonymized interactions [13].

There have been solid exploration in privacy protection in the past decades. The main stream of privacy protection covers various disciplines, such as cryptography, communication, information theory, and so on.

Cryptography is a matured and powerful tool for privacy protection. The major challenge of the current cryptography based privacy protection mechanism is how to deal with the extremely large scale of data in the big data cases. Further, more and more users are using mobile devices with limited computing power, which is a big disadvantage for computing intensive encryption and decryption algorithms.

Communication privacy is mainly explored by the communication and networking community. The fundamental work is Shannon's information theoretical work on perfect secrecy [14] in 1949, where the principle is to maximize the entropy to minimize the probability of recognition. An outstanding work in this direction was carried out by Chaum [15] in 1981, and a recent comprehensive survey on this topic appeared in 2009 [16]. Browsing privacy protection is obviously a hot topic in this Internet age, there are many proposals in web browsing attacks [17] and defence [12], [18]. There have been many mechanisms and systems in place, such as the onion routing mechanism [19], the Tor system [20], and the Crowds system [21].

Modern privacy study has been explored about two decades mainly in two classes, data clustering and privacy frameworks. The early work is the k-anonymity method [22] for privacy preserving proposed in 1998, which is the first method in the data clustering class, then its extension as ℓ-diversity [23] appeared in 2007, and then the t-closeness method developed in 2010 [24]. The data clustering methods are practical and feasible, but it lacks profound theoretical foundation. In 2006, Dwork et al. [25], [26] proposed the *differential privacy* framework, which is a strict mathematical model for privacy protection. A statistical interpretation of differential privacy [27] was developed in 2010. Following the research line of differential privacy, researchers proposed *differential identifiability* [28] and *membership privacy* [29] to cover the problems in the framework of differential privacy.

We note that the context is homogeneous for all of the existing data clustering strategies and the differential privacy framework we mentioned so far, namely, they suppose all the users share one given privacy standard, and ignore personal privacy differences. As a result, personalized privacy methods were developed, such as the work of Li et al. [24] and Jorgensen et al. [30]. Given the nature of privacy measurement, it is not surprise to see the employment of game theory in this field, such as mechanism design [31].

We have to note that legislation and regulation play a critical role in privacy protection. In order to protect cyber privacy, there have been some policies in place. In 2014, the European Court of Justice established a regulation that European citizens possess the right to ask search engines to remove items that are considered inaccurate, irrelevant, or excessive, which is called ''the right to be forgotten'' [32].

Privacy protection has been extensively applied in practice. Besides the aforementioned applications of privacy preserving data publishing and differential privacy, electronic voting

system is also a highly interested field in terms of privacy protection. Firstly, privacy measurement of voting systems has been explored [33], [34]. Bernhard et al. [35] conducted a comprehensive survey on privacy definitions of electronic voting systems, and classified all the definitions into three categories: purely cryptographic based, entropy based, and game based definitions.

Today, we are at the doorstep of the big data age, however, we have to address the privacy threat before we can extensively execute big data applications, and enjoy the benefit. In general, most of the available privacy solutions are ad hoc based, and lack of theoretical foundation. To date, we do not yet find a very good or suitable theory or a set of theories to model and analyze the problem. However, we should not be discouraged in this case by looking at the history of science: Shannon discovered the essence of information after nearly half century since the first appearance of electronic communication; Nash discovered the Nash equilibrium after centuries of human commercial activities. We humbly believe there will be an appropriate theoretical framework for privacy sooner or later based on the forthcoming efforts from every individual of the community. This is also the motivation of the article, in which we expect to review the latest development in privacy protection from both theoretical and practical perspectives, and pave a comfortable start point for passionate readers to explore further in this emerging, exciting and promising field. In order to serve our potential readers with different level of needs, we separate the idea parts of the surveyed papers from the related theoretical parts.

The rest of the paper is structured as follows. We discuss the different roles and operations of privacy systems in Section II. The major developments of modern privacy study are presented in Section III. In Section IV, we survey the privacy study from different disciplines. Mathematical description of privacy study, privacy measurement, and privacy models are surveyed and discussed in Section V, VI, and VII, respectively. We discuss the challenges and opportunities in privacy study in Section VIII. Finally, we summarize the paper in Section IX.

II. PRELIMINARY OF PRIVACY STUDY

In this section, we present an overview of privacy systems, including different participation roles, anonymization operations, and data status. We also introduce the terms and definitions of the system. We demonstrate these as shown in Figure 1.

In terms of participants, we can see four different roles in privacy study.

1) Data generator. Individuals or organizations who generate the original raw data (e.g., medical records of patients, bank transactions of customers), and offer the data to others in a way either actively (e.g. posting photos to social networks to public) or passively (leaving records of credit card transactions in commercial systems).

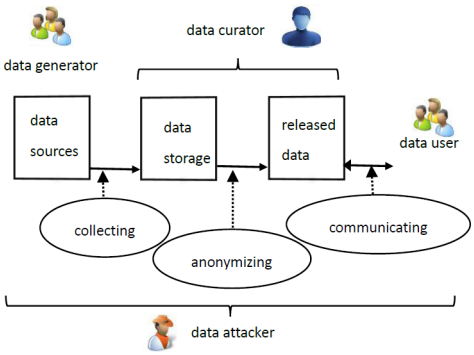


FIGURE 1. The roles and operations of a privacy system.

2) Data curator. The persons or organizations who collect, store, hold, and release the data. Of course, the released data sets are usually anonymized before publishing.
3) Data user. The people who access the released data sets for various purposes.
4) Data attacker. The people who try to gain more information from the released data sets with a benign or malicious purpose. We can see that a data attacker is a special kind of data user.

There are three major data operations in a privacy system.

1) Collecting. Data curators collect data from different data sources.
2) Anonymizing. Data curators anonymize the collected data sets in order to release it to public.
3) Communicating. Data users performan information retrieval on the released data sets.

Furthermore, a data set of the system possesses one of the following three different statuses.

1) Raw. The original format of data.
2) Collected. The data has been received and processed (such as de-noising, transforming), and stored in the storage space of the data curators.
3) Anonymized. The data has been processed by an anonymization operation.

We can see that an attacker could achieve his goals by attacking any of the roles and the operations.

In general, we can divide a given record into four categories according to its attributes.

1) Explicit identifiers. The unique attributes that clearly identify an individual, such as drive licence numbers.
2) Quasi-identifiers. The attributes that have the potential to re-identify individuals when we gather them together with the assistance of other information, such as age, career, postcode, and so on.
3) Sensitive information. The expected information interested by an adversary.
4) Other. The information not in the previous three categories.

We present an example as shown in Table 1. In this example, name is the explicit identifier, while the job, gender,

TABLE 1. A table of patients in a medical database.

Name	Job	Gender	Age	Disease	Other

and age form the quasi-identifier, disease belongs to sensitive information.

We call the quasi-identifiers of a record as a *qid* group, which is also called *equivalence class* in literature.

III. THE MILESTONES OF PRIVACY STUDY

To date, the majority work on privacy protection is conducted in the context of databases. There are mainly two categories: data clustering and theoretical frameworks of privacy. The data clustering direction developed from the initial k-anonymity method, then the ℓ-diversity method, and then the t-closeness (interested readers are encouraged to find the detailed information from [11]). The second category mainly includes the framework of differential privacy and its further developments. We show the journey of privacy study so far in Figure 2, and the details are going to be presented in the following of this section.

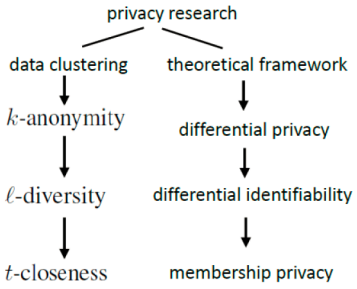


FIGURE 2. The categories of privacy study.

We use Table 1 as an example to quickly demonstrate the journey of the data clustering methods for privacy protection. In Table 1, it is obvious that we need to remove the explicit identifers before we release the data to public. As a result, we have Table 2 as follows.

TABLE 2. A table of patients in a medical database without explicit identifiers.

Job	Gender	Age	Disease	Other

However, the quasi-identifiers of Table 2 can be used to re-identify patients through various techniques, such as linking them to other publicly available data sets.

Around year 2000, Samarati and Sweeny [22], [36], [37] introduced the *k-anonymity* method to protect the content of Table 2. The basic strategy is to make sure each *qid* group has at least k entries in the table in order to decrease the probability of re-identification. For example, merging job types of dancer, singer, and painter to artist, engineer and lawyer to professional, the exact age (e.g., 38) can be represented in a range (e.g., [35-40)). A sample of the anynomized output using k-anonymity is shown in Table 3, where $k = 2$.

TABLE 3. An anonymized table of patients obeying k-anonymity ($k = 2$).

Job	Gender	Age	Disease	Other
Artist	F	[35-40)	Flu	NA
Artist	F	[35-40)	Cancer	NA
Professional	M	[30-35)	Flu	NA
Professional	M	[30-35)	Hepatitis	NA

In this way, an attacker can re-identify a victim with a maximum probability of $\frac{1}{k}$. With a decently large k, the privacy can be nicely protected. However, we have to be aware that a larger k meaning a large information loss.

Based on the k-anonymity mechanism, Tao and Xiao [38] further introduced personalized privacy for the k-anonymity method.

Outside of privacy preserving data publishing, k-anonymity is also widely used in mobile networks, such as location privacy of mobile users [39], and location based service privacy [40].

We can see that the k-anonymity data clustering method tries to work on attributes of the quasi-identifers, and invests no effort on the sensitive attributes. This exposes the k-anonymity method to some subtle but effective attacks, such as the homogeneity attack due to lack of diversity in sensitive attributes, and the background knowledge attack based an adversary's knowledge of the victims. For example, an attacker knows that Alice is in Table 3, and she has cancer (the sensitive information). Due to the fact that the number of the specific sensitive value is unique (or very limited), then he can identify that Alice is the second record in the table.

In order to overcome the disadvantages of the k-anonymity method, Machanavajjhala et al. [23] proposed the *ℓ-diversity* method in 2006, which requires the sensitive attributes to be well represented in anonymized data sets. A formal description of this method is "guarantee there are at least ℓ distinct values for the sensitive attributes in each *qid* group." A sample output of ℓ-diversity ($\ell = 2$) under k-anonymity ($k = 2$) is shown in Table 4.

TABLE 4. An anonymized table of patients obeying ℓ-diversity ($\ell = 2$) under k-anonymity ($k = 2$).

Job	Gender	Age	Disease	Other
Artist	F	[35-40)	Flu	NA
Artist	F	[35-40)	Flu	NA
Artist	F	[35-40)	Cancer	NA
Artist	F	[35-40)	Cancer	NA

We note that defenders have to prepare a sufficiently large ℓ to beat the attacks on sensitive attributes. For example, if an attacker is sure that Alice is in Table 4, then she has a possibility of 0.5 of suffering cancer. In particular, ℓ-diversity degraded to k-anonymity when $\ell = 1$. The term "well represented" can be measured by probability, entropy, and so on.

We can treat the ℓ-diversity mechanism as an extension of the k-anonymity principle by including the sensitive attributes.

To implement ℓ-diversity, we should increase the granularity of sensitive attributes (e.g., representing the diseases as common or deadly) or adding noise. We loss information of the original records to buy privacy protection.

In some specific cases, the ℓ-diversity method may release more privacy to attackers. For example, for a given test results of a virus, the probability of negative is 0.99. If we know Alice is in the data set, then she is positive with a probability of 0.01. However, after a ℓ-diversity operation, if we know Alice is in one *qid* group, then we can conclude that her positive probability is 0.5. In other words, ℓ-diversity operation offers more information gain to attackers in some specific cases.

In order to fix this vulnerability, Li et al. [24] proposed *t-closeness* in 2010. The idea is like this: for a given *qid* group, guarantee its distribution is bounded by t against its corresponding distribution on the whole data set. A further work of *t-closeness-like* was proposed by Rebollo-Monedero and colleagues in year 2000 [41].

Different from the data clustering strategy, the differential privacy framework [25] was proposed in 2006, which offers a strong privacy protection in sense of information theory. The basic background is that an attacker may obtain expected information by multiple queries to a statistical database on top of his background knowledge of victims. The defence strategy is: for two data sets with a minimum difference, the difference between the queries on the two data sets is very limited, therefore limiting the information gain for attackers. One popular methods to achieve this is adding noise to outputs. Lee and Clifton [28] found that differential privacy does not match the legal definition of privacy, which is required to protect individually identifiable data, rather than the how much one individual can affect an output as differential privacy provides. As a result, they proposed *differential identifiability* to provide the strong privacy guarantees of differential privacy, while letting policy makers set parameters based on the established privacy concept of individual identifiability. Following this research line, Li et al. [29] analyzed the pros and cons of differential privacy and differential identifiability, and proposed a framework called *membership privacy*. The proposed framework offers a principled approach to developing new privacy notions under which better utility can be achieved than what is possible under differential privacy.

As differential privacy is a global concept for all users of a given data set, namely the privacy protection granularity is the same to all protected users, therefore it is called *uniform privacy* or *homogenous differential privacy*. In order to offer customized privacy protection for individuals, *personalized differential privacy* (also named as *heterogenous differential privacy* or *non-uniform privacy*) were also extensively explored [30], [42].

IV. DISCIPLINES IN PRIVACY STUDY

Based on the content of the previous sections, we can see that privacy research just started, and privacy research in big data is almost untouched. In this section, we try to survey the major disciplines involving in privacy study. Of course, the list of disciplines is not exhaustive. We also note that Information Theory is extensively used as a theoretical foundation in various disciplines discussed here, and we therefore do not list it as an independent discipline.

A. CRYPTOGRAPHY

Based on the current situations in practice, we can conclude that encryption is still the dominant methodology for privacy protection although it is a bit away from the privacy protection theme we talking about here.

Cryptography can certainly be used in numerous fashions for privacy protection in the big data age. For example, a patient can use the public key of her doctor to encrypt her medical documents, and deposits the ciphertext into the doctor's online database for her treatment while her privacy is strictly preserved.

With the emergence of big data, clouds are built to serve many applications due to its economical nature and accessibility feature. For example, many medical data sets are outsourced to clouds, which triggers the privacy concerns from patients. The medical records of a patient can only be accessed by authorized persons, such as her doctors, rather than other doctors or people. The public key encryption is obviously not convenient if the number of the authorized persons is sufficiently large due to the key management issue. In this case, Attribute Based Encryption (ABE) is an appropriate tool [43], [44], which was invented in 2004 by Sahai and Waters [45]. In the ABE scheme, a set of descriptive attributes of the related parties, such as hospital ID, doctor ID, and so on, are used to generate a secret key to encrypt messages. The decryption of a ciphertext is possible only if the set of attributes of the user key matches the attributes of the ciphertext. The ABE scheme creatively integrates encryption and access control, and therefore no key exchange problem among the members of the authorized group.

The dilemma of encryption based privacy protection in big data is: on one hand, we need to offer sufficient privacy protection for users, at the same time, we have to make the encrypted data informative and meaningful for big data analysis and public usage. As a result, we face a number of challenges as follows.

One challenge is information retrieval on encrypted data. This research branch is also called *searchable encryption*, which boomed around year 2000 [46], [47]. The basic idea is as follows: a user indexes and encrypts her document collection, and sends the secure index together with the encrypted data to a server which may be malicious. To search for a given keyword, the user generates and submits a trapdoor for the keyword, which the server uses to run the search operation and recover pointers to the appropriate encrypted documents.

Another challenge here is operations on encrypted objects. This research branch is named as *homomorphic encryption* started in 1978 [48]. In this kind of encryptions, we expect to carry out computations on ciphertext, and obtain an encrypted

output. If we decrypt the output it should match the result of operations performed on the original plaintext. Mathematically, we can describe it as follows: given a message m, a key k, and an encryption algorithm E, we can obtain a ciphertext $E_k(m)$. Let f be a function, and its corresponding function is f', D_k be a decryption algorithm under key k, then an encryption scheme is homomorphic if $f(m) = D_k(f'(E_k(m)))$.

In 2009, Gentry kicked off a further development in this direction, *Fully Homomorphic Encryption* (FHE), which supports arbitrary computation on ciphertexts [49]. A survey on this branch can be found in [50]. The problem is that we do not have a feasible fully homomorphic encryption system in place yet due to the extraordinary inefficiency in computing. Compared to FHE, Multi-Party Computation (MPC), which was initiated by Yao in 1982 [51], has been used in practice by offering weaker security guarantees but much more efficient. The scenario of MPC is like this: multiple participants jointly compute a public function based on their private inputs while reserve their input privacy against the other participants, respectively.

We have to note that encryption can protect the privacy of a object itself, however, it is vulnerable against *side information attacks*, such as traffic analysis attack against anonymous communication systems. For example, we can encrypt web pages of a protected web site, however, the encryption cannot change the *fingerprints* of the web pages, which are represented by the size the HTML text, number of web objects, and the size of the web objects. An attacker can figure out which web pages or web sites a victim visited using traffic analysis methodology [52]–[54]. In terms of solutions, information theory based packet padding is the main player, including dummy packet padding [55] and predicted packet padding [12].

B. DATA MINING AND MACHINE LEARNING

Data mining and machine learning are the biggest threat to modern privacy protection. The essential purpose of mining and learning is to obtain new knowledge from data sets. However, these techniques are very damaging if they are in the evil hands. The data community realized the danger when they tried to release data sets to public. We note that in traditional data mining and machine learning, the data is usually stored in databases in a given venue, and the data environment is homogeneous, e.g., the studied objects are usually records or tables.

A comprehensive survey in this field was done in 2010 by Fung et al. [11] in terms of privacy preserving data publishing. They surveyed the data publishing issue with privacy protection: given a data set T, how to transform it to a publishable data set T' under the condition of privacy protection of the data generators in T. They classified the attacks in two categories.

- Linkage attack. Attackers combine the publicly released data set T' with other data sets they possess to re-identify the data generators at different granularities, such as attribute level, record level, or table level.
- Probabilistic attack. An attacker gains more new knowledge about a victim based on the released T' compared with his original background knowledge of the victim before the releasing.

Various privacy models, such as the k-anonymity, ℓ-diversity, t-closeness, and ε-differential privacy, were surveyed, and the privacy operation algorithms were also enumerated.

One thing we note here is that the authors "urge computer scientists in the privacy protection field to conduct cross-disciplinary research with social scientists in sociology, psychology, and public policy studies. Having a better understanding of the privacy problem from different perspectives can help realize successful applications of privacy-preserving technology."

In 2015, Xu et al. [56] conducted another extensive survey on privacy disclosure and protection from the data mining perspective. They surveyed the privacy concerns and privacy techniques from the perspectives of different roles of data mining applications: data provider, collector, miner, and decision maker. Game theory was identified as a promising tool for privacy protection by the authors.

Due to these two excellent surveys, we suggest readers who are interested in this field to read them for more details.

It is worthwhile to notice Zhu et al. [57] argued that data are usually related in data sets, and proposed a concept of correlated differential privacy in PPDP.

C. BIOMETRIC PRIVACY

Biometric is a powerful tool for security, which aims to identify individuals based on their physical, behavioral, and physiological attributes, such as face, iris, fingerprint, voice, and gait. Biometrics has been widely used in access control, and the procedure includes two stages: enrollment and release. In the first stage, biometric features, such as fingerprints, are sampled, and the information is stored in a database either as a raw data or in a transformed form. In the second stage, the related biometric characteristics are sampled again on site, and compared with the stored one for authentication.

Compared with the conventional passwords, biometric methods enjoy a few advantages, including hard to be steal, not need to be remembered. However, it is a nightmare once a biometric is compromised, e.g., it can be used to impersonate the victims, to break the privacy of the victims. In order to protect privacy introduced by biometric based security systems, securer multi-party computation techniques are usually hired to execute the job [58], [59].

Lai et al. [60], [61] studied the privacy-security tradeoffs in biometric systems based on information theory framework for both single use case and multiple use case. They concluded that it was possible to achieve perfect privacy of biometric systems if and only if common randomness could be generated from two biometric measurements. Ignatenko and Willems [62] determined the fundamental tradeoffs between secret-key, identification, and privacy-leakage for two specific biometric settings.

A recent paper [63] summarized the techniques of privacy protection in biometric domain. The authors pointed out that *soft biometrics* (e.g., age, gender, height), rather than fingerprint, iris, can be easily collected as an ancillary information during the procedure of biometric data collection. The soft biometrics poses a great threat to privacy. This conclusion is not surprise to us if we treat soft biometrics as a kind of quasi identifiers.

Nowadays, surveillance cameras are widely used in many countries. Some video records are broadcasted on TV as part of news or uploaded to the Internet for public for specific purposes, e.g., looking for evidence. In this case, some parts of the video need to be protected, e.g., an unrelated person in the video by chance, therefore, a mask is usually employed to cover the face or body. This is called *privacy region protection*. However, with the technical development in image processing, especially the recently invented compressed sensing [64], [65], it is possible to recover the original image based on the partial image (non-privacy region) through the dependency and other features.

D. GAME THEORY

Game theory is a rich set of mathematical models to deal with conflict and cooperation between intelligent and rationale decision-makers. It has been widely used in economics, political science, psychology, computer science, and so on. Given the nature of privacy protection, game theory is obviously a powerful tool to be hired to motivate privacy investment, settle argument among different participants, and so on. There is a long list of the applicable scenarios that game theory can contribute. In this section, we only enumerate a few closely related branches to the theme of this paper.

Game theory has been widely applied in various security studies. For example, Macnshaei et al. [66] studied game theory from network security perspectives, and researchers of [40] and [67] also applied game theory to analyze privacy protection in mobile networks. However, we noticed that the application of game theory in privacy is far less compared with its popularity in security. With a great concern on privacy in big data, it is sure that game theory will be extensively used. For example, Krishnamurthy and Poor [68] combined social and game theoretical learning to offer an overview of social sensing, where privacy is a big factor.

After the birth of differential privacy, game theoretical approaches were immediately applied to this new and exciting field. For example, McSherry and Talwar [69] proposed mechanism design for privacy under the framework of differential privacy in 2007. Pai and Roth [31] presented a survey on mechanism design and privacy with a focus on differential privacy in 2013.

Due to the value of data, people nowadays start to sale their data to organizations for monetary reward, game theory is then naturally employed to serve the needs, such as pricing and data auction. Ghosh and Roth [70] initiated the study of privacy auction where privacy is treated as a type of goods. Following this direction, a few further development have been conducted in [71] and [72]. Recently, Xu et al. [73] noticed the impact of information asymmetry in privacy auction, and proposed a contract-based approach to balance privacy protection and data utility.

E. POLICY AND SOCIAL SCIENCE

Acquisti et al. [74] reviewed diverse streams of empirical research on privacy behavior from the perspective of social and behavior science. They explained the phenomenon through three themes. 1) Uncertainty, people are not sure about the consequences of privacy related behaviors and their own preferences over these consequences; 2) Context-dependence, depending on a given situation, an individual can exhibit anything ranging from extreme concern to apathy on privacy; 3) Malleability and influence, the degree to which privacy concerns are malleable and manipulable by commercial and governmental interests. They concluded that privacy depended on many factors, such as culture and timing. The authors argued the problem as a tradeoff between transparency and control.

From the perspective of law and legislation, there have been laws and regulations in place, such as the American *Privacy Act of 1974*, the European *General Data Protection Regulation* in 2012. In 2014, the European Court of Justice enacted that their citizens have the right to request search engines to delink their results of items that are considered inaccurate, irrelevant, or excessive, which is a new right as right-to-be-forgotten. Newman [32] pointed out that it was a trend to execute privacy protection in a way of distributed regulation, in which law enforcement relied on individuals and firms to monitor and implement regulations.

Horvitz and Mulligan [75] discussed the control of big data from the policy-making perspective, and believed that discussions on data and the threat from machine learning among policy-makers and the public will lead to insightful designs of policies and programs, which can balance the goals of protecting privacy and ensuring fairness with those of reaping the benefits to scientific research, and to individual and public health.

V. MATHEMATICAL DESCRIPTION OF PRIVACY STUDY

In this section, we present the mathematical models that we have mentioned in the previous sections, aim to help readers to deeply understand the existing privacy protection schemes, and pave a solid starting ground for further exploration on the uncharted land.

From a system viewpoint, our privacy study can be summarized as shown in Figure 3.

In Figure 3, $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$ is the original data, and the anonymization system is a mapping function F, which transforms $\mathcal{X}$ to $\mathcal{Y} = \{Y_1, Y_2, \ldots, Y_m\}$ as an expected output. The goals of function F are to reserve the usability of † as much as possible, and protect the privacy at a certain level. For an attacker, his goal is to obtain $\widehat{\mathcal{X}}$, an estimation of $\mathcal{X}$ based on the released data $\mathcal{Y}$ and possible background knowledge.

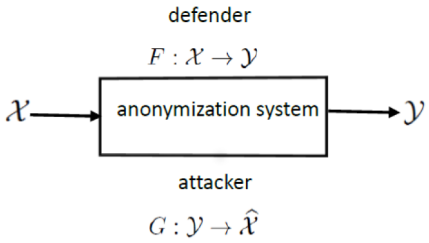


FIGURE 3. The defence and attack of an anonymization system.

The two goals of a privacy protection systems are usually termed as *utility* and *privacy*. In other words, utility and privacy are the two key constraints of function F. Utility is usually measured by *distortion* D, and privacy can be measured by *leakage* L.

Let λ be a Lebesgue measure (or simply an abstract measure), then we can represent data distortion as follows.

$$D = \lambda(\mathcal{X}; \mathcal{Y}). \tag{1}$$

Similarly, information leakage can be expressed as follows.

$$L = \lambda(\mathcal{X}; \widehat{\mathcal{X}}). \tag{2}$$

There are many different measurement metrics for D and L. For example, a simple measure of distortion [76] could be the average mean-square defined as follows.

$$D = \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}[(X_k - Y_k)^2], \tag{3}$$

where $\mathbb{E}[\cdot]$ is the expectation over the joint distribution of $\mathcal{X}$ and $\mathcal{Y}$.

A simple measure of L could be the mutual information $I(\cdot)$ defined in information theory.

$$L = I(X, \widehat{X}). \tag{4}$$

We will discuss related measurement metrics in privacy study in the next section for more details.

Given a distortion threshold D_0, and a information leakage threshold L_0, we can simply describe the studied object as an optimization problem as follows.

$$\begin{aligned} &optimize\ F \\ &s.t.\ D \le D_0 \\ &\quad L \le L_0 \end{aligned} \tag{5}$$

At the same time, from an attacker's viewpoint, he is interested about knowledge $\widehat{\mathcal{X}} = \{\widehat{X}_1, \widehat{X}_2, \ldots, \widehat{X}_k\}$. In general, we suppose $\widehat{X} \subseteq \mathcal{X}$, and he aim to learn knowledge of the corresponding to $\widehat{X}$ in $\mathcal{X}$ (in some occasions, he may be interested to know whether $\widehat{X} \subseteq \mathcal{X}$ or not).

Suppose an attacker possesses a background knowledge of the expected knowledge $\widehat{X}$ in $\mathcal{X}$. Then his learning is a mapping function $G : \mathcal{Y} \to \widehat{X}$ given the background knowledge. Similar to the defence operation, attack could also be represented as an optimization problem similar to Equation (5).

VI. PRIVACY MEASUREMENTS

In general, measurement is the foundation for scientific work. At one hand measurement is a must, just as indicated by Galilei "Measure what is measurable, and make measurable what is not so." At the other hand, some objects are not easy to be measured, such as the "madness of men" expressed by Sir Newton ("I can calculate the movement of stars, but not the madness of men.") To date, the measurement on privacy is not very clear. The good news is that researchers have invented some measurement tools in place from various research communities, we can borrow these metrics for privacy study with appropriate adjustments. We present a few commonly used measurement metrics below for readers.

A. RELATIVE MEASUREMENT

It is hard to obtain a direct measurement in some cases, then the relative measurements become an option, such as object A is bigger than B, Alice is closer to Bella than Ian. In other words, we have a benchmark, then we measure the distance between the studied object to the benchmark. We call this *relative measurement*.

One popular relative measurement is the Kullback-Leibler distance [77], which is used to measure the distance between two distributions, $p(x)$ and $q(x)$. It is represented as follows.

$$D(p, q) = \sum_{x\in\chi} p(x) \cdot \log\frac{p(x)}{q(x)}, \tag{6}$$

where χ is the sample space of x.

There are also many other measurements similar to Kullback-Leibler distance, such as Jeffrey's measure [78]. Interested readers are encouraged to find more information from reference [79]. These measures are the first order measures as they all base on first order statistics.

In order to measure accurately, second order metrics were also proposed. One popular second order metric is correntropy [80]. It works independently on measuring pair-wise arbitrary samples. For any two finite data sequences A and B, suppose we have sample $\{(A_j, B_j)\}_{j=1}^{m}, m \in \mathbb{N}$, then the similarity of the sequences is estimated as

$$\hat{V}_{m,\sigma}(A, B) = \frac{1}{m}\sum_{j=1}^{m} k_\sigma(A_j - B_j), \tag{7}$$

where $k_\sigma(\cdot)$ is the Gaussian kernel, which is usually defined as

$$k_\sigma(\cdot) \triangleq \exp\left(-\frac{x^2}{2\sigma^2}\right). \tag{8}$$

Correntropy metrics are symmetric, positive, and bounded.

One more example of relative measurement is the definition of *flash crowd* by Wendell and Freedman [81]. The phenomenon is that many web users will access a given web site due to some specific reasons, such as a breaking news. They gave the definition as follows: a flash crowd is a period over which request rates for a particular domain name are increasing exponentially.

In privacy studies, two measurement metrics were mentioned in [27], we also introduce them here for readers. Given P is the distribution of the data $X \in \chi$, and P_Z is the empirical distribution of the released data Z. Let $F(x)$ and $\hat{F}_Z(x)$ be the cumulative distribution function (cdf) corresponding to P and P_Z, respectively.

The Kolmogorov-Smirnov (KS) distance is defined as follows.

$$d = \sup_{x \in \chi} |F(x) - \hat{F}_Z(x)| \tag{9}$$

The squared L_2 distance is defined as follows.

$$d = \left[\sum_i \left(P(x_i) - \hat{P}_z(x_i) \right)^2 \right]^{\frac{1}{2}} \tag{10}$$

B. INFORMATION THEORETIC MEASUREMENT

From information theoretical perspective, Coney et al. [33] defined the privacy measurement for voting systems. Let a voter's vote be a random variable V, and S be the information (e.g., geographic information) through sources other than the voting system, E be the information that an adversary gaining from the voting system. They defined an election system is *perfectly private* if V is conditionally independent of E after conditioning on S, namely, $p_{V|S}(v; s) = p_{V|S,E}(v; s, e)$ for all v, s, e. In other words, the information from the voting system contributes no gain to the probability of V. They further defined the *amount of privacy loss*, L, as

$$L = \max \left(H(V|S) - H(V|S, E) \right), \tag{11}$$

where $H(X)$ is the entropy of random variable X.

The measurement was revisited by Bernhard et al. in year 2012 [34].

C. UNICITY MEASURE

De Mpntijoye [6] proposed a unicity ε_p as the measure of privacy risk of being re-identified from a simply anonymized data set, which does not contain obvious identifiers, such as name, home address, phone number, and staff ID. The detailed definition is as follows.

Given a $p(p \geq 1)$ point object O_p, and a simply anonymized data set D, unicity ε is the probability of extracting the subset of trajectories $S(O_p)$ from D that match the p points composing O_p. Namely

$$\varepsilon = \frac{|S(O_p)|}{|D|}, \tag{12}$$

where $|x|$ is the cardinality of set x.

An object is identified if $|S(O_{p=k})| = 1$ when k points are needed.

VII. MATHEMATICAL PRIVACY MODELS

A. k-ANONYMITY MODEL

For the k-anonymity model, Machanavajjhala et al. [23] mathematically described it as follows.

Let $T = \{t_1, t_2, \ldots, t_n\}$ be a table of a data set D, $A = \{A_1, A_2, \ldots, A_m\}$ be all the attributes of T, and $t_i[A_j]$ be the value of attribute A_j of tuple t_i. If $C = \{C_1, C_2, \ldots, C_k\} \subseteq A$, then we denote $T[C] = \{t[C_1], t[C_2], \ldots, t[C_k]\}$ as the projection of t onto the attributes in C.

The quasi-identifer is defined as a set of nonsensitive attributes of a table if these attributes can be linked with external data sets to uniquely identify at lest one individual in the data set D. We use QI to represent the set of all quasi-identifers.

A table T satisfies k-anonymity if for every tuple $t \in T$ there exist at least $k-1$ other tuples $t_{i_1}, t_{i_2}, \ldots, t_{i_{k-1}} \in T$, such that $t[C] = t_{i_1}[C] = t_{i_2}[C], \ldots, t_{i_{k-1}}[C]$, for all $C \in QI$.

B. ℓ-DIVERSITY MODEL

As aforementioned, *ℓ-diversity* [23] is an extension of the k-anonymity to "well represent" the sensitive attributes. There are four different interpretations of the term "well represented".

1) Distinct ℓ-diversity. Similar to k-anonymity, each sensitive attribute has to possess at least ℓ distinct values in each qid group.
2) Probabilistic ℓ-diversity. The frequency of a sensitive value in a qid group is at most $\frac{1}{\ell}$.
3) Entropy ℓ-diversity. For every qid group, its entropy is at least $\log \ell$.
4) (c, ℓ)-diversity. The frequency of sensitive values of a qid group is confined in the range defined by c (a real number) and ℓ (in integer).

C. t-CLOSENESS MODEL

t-Closeness [24] is a further development on ℓ-diversity. For a given table T, for a set and its superset of T, we limited the distance between the two sets not higher than a given threshold t, then the table follows t-closeness.

D. DIFFERENTIAL PRIVACY FRAMEWORK

Differential privacy [25] was proposed around 2006 about the measurement and standard of privacy for data tuples from databases. It is a stronger privacy protection framework compared with the data clustering methods. There are a set of definitions composing the framework.

Definition (ε-Differential Privacy): A randomized function $\mathcal{K}$ gives ε-differential privacy if for all data sets D_1 and D_2 differing on at most one element (the two data sets are called *neighboring data sets*), and all $S \in Range(\mathcal{K})$,

$$Pr[\mathcal{K}(D_1) \in S] \leq e^{\varepsilon} \times Pr[\mathcal{K}(D_2) \in S]. \tag{13}$$

In other words, with the minimum difference between two data sets, the difference after an anonymization operation is not greater than a given value (e^{ε}).

In the differential privacy framework, *global sensitivity* [26] is a key metric, which is defined as

$$\Delta_f = \max_{D_1 \sim D_2} ||f(D_1) - f(D_2)||_1, \tag{14}$$

where function $f : D \to \mathbb{R}^d$, $D_1 \sim D_2$ means they are neighboring data sets, and $||\cdot||_1$ is the L_1 norm.

The implementation of differential privacy is usually executed by the Laplace mechanism, which is defined as follows.

For a multidimensional real-valued query function $q : \mathcal{D} \to \mathbb{R}^d$ with sensitivity Δ_f, the Laplace mechanism will output

$$\begin{aligned}\mathcal{K}(D) &:= q(D) + Lap\left(\frac{\Delta_f}{\varepsilon}\right)^d \\ &= \left(q_1(D) + Lap\left(\frac{\Delta_f}{\varepsilon}\right), \ldots, q_d(D) + Lap\left(\frac{\Delta_f}{\varepsilon}\right)\right),\end{aligned} \tag{15}$$

where $Lap(\lambda)$ is a random variable with probability density function

$$f(x) = \frac{1}{2\lambda}\exp^{-|x|/\lambda}, \quad \forall x \in \mathbb{R}, \tag{16}$$

and all d Laplacian random variables are independent.

A number of features of differential privacy in distributed and parallel query scenario were derived by McSherry in 2009 [82].

A recent work [83] used the staircase mechanism to replace the Laplace mechanism in order to reduce the amount of noise needed for differential privacy.

Wasserman and Zhou interpreted the framework of differential privacy in the statistical domain in 2010 [27].

VIII. CHALLENGES AND OPPORTUNITIES

A. CHALLENGES OF PRIVACY STUDY

Big data is a new environment for computer science, and privacy is one of the critical problem, which has to be appropriately addressed before we can enjoy the pervasive applications of big data.

There are many problems and challenges ahead in terms of privacy study in big data. We summarize the major ones here for readers based on our current understanding.

1) Measurement of privacy. As privacy is a subjective concept, it various from person to person, from time to time even for the same person. It is hard to define it, and therefore, hard to measure. This problem is fundamental and challenging. It needs the effort not only from technical aspects, but more from social and psychological perspective.
2) Theoretical framework of privacy. We now have data clustering methods and the differential privacy framework for data privacy. However, we also see the limitations of various data clustering methods, and the needs to adapt the differential privacy in practice. Should we have new and better theoretical foundations for privacy study in big data era? we believe the answer is positive, and it takes time.
3) Scalability of privacy algorithms. We have some mechanisms and strategies in place to handle big databases, and the main strategy is divide and conquer. However, the scale of big data is far bigger than a database. Therefore, it is challenging to design scalable algorithms for privacy algorithms.
4) Heterogeneity of data source. The available privacy algorithms are almost all for homogeneous data sources, such as the records in databases. However, the data sources of forthcoming big data are heterogeneous with a high probability. It is challenging to deal with heterogenous data sources in an efficient way.
5) Efficiency of privacy algorithms. Given the volume of big data, efficiency becomes a very important element of privacy algorithms in the big data environment.

B. OPPORTUNITIES AND DIRECTIONS

We can conclude that we desperately need to improve the existing privacy protection methods to meet the unprecedented requirements of big data. Furthermore, new privacy frameworks and mechanisms are highly expected in the near future. Based on our understanding, we believe the followings are the promising directions for our investment.

1) Quantum computing for unconditional privacy preserving. We are getting closer and closer to the practical usage of quantum computers. One good news is that quantum computing can offer fantastic functions in security and privacy preserving. The majority advantage of the current encryption methods is time complexity. However, it is not an unconditional method for privacy protection or security. The recently proposed measurement based model of quantum computation [84] provide a promising diagram to achieve *blind computing*, meaning a client can delegate a computation to a quantum server, and the server can execute the task without gaining any knowledge about the input, output, and the client. Braz et al. [85] implemented the conceptual framework of the model and demonstrated the feasibility of blind quantum computing. It is time to start our exploration in quantum computing not only for privacy and security, but also the other aspects of computing.
2) Integrating computer techniques with social science. We have to accept that in terms of privacy study, computing techniques and strategies have to follow or serve the needs and findings of social science, which is the leading battle ground. This is supported by the leading researchers in the field, such as the authors of the highly cited survey paper [11], and the emerging discipline of Computational Psychophysiology [86].
3) Inventing new theoretical privacy frameworks. We have seen the practical usage of the various data clustering methods in privacy protection, and also the strictness of the differential privacy framework. The former suffers different vulnerability, and the later lacks some flexibility and feasibility in practice. Do we have new frameworks to fill the gap? the answer is yes. We note that given the complexity of privacy study, we may need a set of theories to serve the problem, rather

a single one. For example, fuzzy logic to deal with ambiguous concepts, and game theory to settle arguments from different parties.

IX. SUMMARY

With the approaching of the big data age, privacy protection is becoming a unavoidable hurdle in front of us. Motivated by this, we surveyed the major milestones in privacy study up to date from different perspectives, aiming to pave a reliable ground for interested readers to explore this exciting, emerging, and promising field. We summarized the outputs of privacy study in different research principles and communities. In particular, we presented the mathematical effort of the related privacy models and frameworks.

We have seen the great effort from different communities in privacy protection, especially from the theoretical aspect. However, these theoretical attempts are still insufficient to most of the incoming big data applications. We fully believe the theoretical effort in big data privacy is essential and highly demanded in problem solving in the big data age, and it is definitely worthwhile to invest our energy and passion in this direction without any reservation.

REFERENCES

[1] J. T. Overpeck, G. A. Meehl, S. Bony, and D. R. Easterling, "Climate data challenges in the 21st century," *Science*, vol. 331, no. 6018, pp. 700–702, 2011.

[2] V. Marx, "Biology: The big challenges of big data," *Nature*, vol. 498, no. 7453, pp. 255–260, 2013.

[3] L. Lang, "Advancing global health research through digital technology and sharing data," *Science*, vol. 331, no. 6018, pp. 714–717, 2011.

[4] G. King, "Ensuring the data-rich future of the social sciences," *Science*, vol. 331, no. 6018, pp. 719–721, 2011.

[5] D. Laney, "3D data management: Controlling data volume, velocity, and variety," Gartner, Stamford, CT, USA, Tech. Rep., 2011.

[6] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Sci. Rep.*, vol. 3, Mar. 2013, Art. no. 1376.

[7] Y.-A. de Montjoye, L. Radaelli, V. K. Singh, and A. Pentland, "Unique in the shopping mall: On the reidentifiability of credit card metadata," *Science*, vol. 347, no. 6221, pp. 536–539, 2015.

[8] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[9] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y. Zhu, "Tools for privacy preserving distributed data mining," *SIGKDD Explorations Newslett.*, vol. 4, no. 2, pp. 28–34, 2002. [Online]. Available: http://doi.acm.org/10.1145/772862.772867

[10] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Privacy aware learning," *J. ACM*, vol. 61, no. 6, 2014, Art. no. 38.

[11] 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, 2010, Art. no. 14.

[12] S. Yu, G. Zhao, W. Dou, and S. James, "Predicted packet padding for anonymous Web browsing against traffic analysis attacks," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 4, pp. 1381–1393, Aug. 2012.

[13] B. Shebaro, O. Oluwatimi, D. Midi, and E. Bertino, "IdentiDroid: Android can finally wear its anonymous suit," *Trans. Data Privacy*, vol. 7, no. 1, pp. 27–50, 2014. [Online]. Available: http://www.tdp.cat/issues11/abs.a150a13.php

[14] C. E. Shannon, "Communication theory of secrecy systems," *Bell Labs Tech. J.*, vol. 28, no. 4, pp. 656–715, Oct. 1949.

[15] D. Chaum, "Untraceable electronic mail, return addresses, and digital pseudonyms," *Commun. ACM*, vol. 24, no. 2, pp. 84–88, 1981.

[16] M. Edman and B. Yener, "On anonymity in an electronic society: A survey of anonymous communication systems," *ACM Comput. Surv.*, vol. 42, no. 1, 2009, Art. no. 5.

[17] C. V. Wright, F. Monrose, and G. M. Masson, "On inferring application protocol behaviors in encrypted network traffic," *J. Mach. Learn. Res.*, vol. 7, pp. 2745–2769, Dec. 2006.

[18] C. V. Wright, S. E. Coull, and F. Monrose, "Traffic morphing: An efficient defense against statistical traffic analysis," in *Proc. NDSS*, 2009, pp. 1–14.

[19] M. G. Reed, P. F. Syverson, and D. M. Goldschlag, "Anonymous connections and onion routing," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 4, pp. 482–494, May 1998.

[20] R. Dingledine, N. Mathewson, and P. F. Syverson, "Tor: The second-generation onion router," in *Proc. USENIX Secur. Symp.*, 2004, pp. 303–320.

[21] M. K. Reiter and A. D. Rubin, "Crowds: Anonymity for Web transactions," *ACM Trans. Inf. Syst. Secur.*, vol. 1, no. 1, pp. 66–92, 1998.

[22] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," in *Proc. IEEE Symp. Res. Secur. Privacy (S&P)*, May 1998, pp. 1–19.

[23] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," *ACM Trans. Knowl. Discovery Data*, vol. 1, no. 1, Mar. 2007, Art. no. 3.

[24] N. Li, T. Li, and S. Venkatasubramanian, "Closeness: A new privacy measure for data publishing," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 7, pp. 943–956, Jul. 2010.

[25] C. Dwork, "Differential privacy," in *Proc. ICALP*, 2006, pp. 1–12.

[26] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. 3rd Conf. Theory Cryptogr. (TCC)*, 2006, pp. 265–284.

[27] L. Wasserman and S. Zhou, "A statistical framework for differential privacy," *J. Amer. Statist. Assoc.*, vol. 105, no. 489, pp. 375–389, 2010.

[28] J. Lee and C. Clifton, "Differential identifiability," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, Aug. 2012, pp. 1041–1049.

[29] N. Li, W. H. Qardaji, D. Su, Y. Wu, and W. Yang, "Membership privacy: A unifying framework for privacy definitions," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, Nov. 2013, pp. 889–900.

[30] Z. Jorgensen, T. Yu, and G. Cormode, "Conservative or liberal? Personalized differential privacy," in *Proc. 31st IEEE Int. Conf. Data Eng. (ICDE)*, Seoul, South Korea, Apr. 2015, pp. 1023–1034.

[31] M. M. Pai and A. Roth, "Privacy and mechanism design," *ACM SIGecom Exchanges*, vol. 12, no. 1, pp. 8–29, 2013.

[32] A. L. Newman, "What the 'right to be forgotten' means for privacy in a digital age," *Science*, vol. 347, no. 6221, pp. 507–508, 2015.

[33] L. Coney, J. L. Hall, P. L. Vora, and D. Wagner, "Towards a privacy measurement criterion for voting systems," in *Proc. Nat. Conf. Digit. Government Res.*, Atlanta, GA, USA, May 2005, pp. 287–288.

[34] D. Bernhard, V. Cortier, O. Pereira, and B. Warinschi, "Measuring vote privacy, revisited," in *Proc. ACM Conf. Comput. Commun. Secur. (CCS)*, Raleigh, NC, USA, Oct. 2012, pp. 941–952.

[35] D. Bernhard, V. Cortier, D. Galindo, O. Pereira, and B. Warinschi, "SoK: A comprehensive analysis of game-based ballot privacy definitions," in *Proc. IEEE Symp. Secur. Privacy (SP)*, San Jose, CA, USA, May 2015, pp. 499–516.

[36] P. Samarati, "Protecting respondents' identities in microdata release," *IEEE Trans. Knowl. Data Eng.*, vol. 13, no. 6, pp. 1010–1027, Nov. 2001.

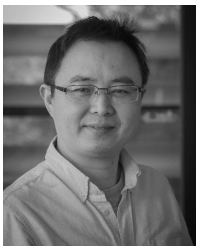
[37] L. Sweeney, "k-anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, 2002.

[38] Y. Tao and X. Xiao, "Personalized privacy preservation," in *Privacy-Preserving Data Mining: Models and Algorithms*. 2008, pp. 461–485. [Online]. Available: http://dblp.uni-trier.de/pers/hd/t/Tao:Yufei

[39] D. Yang, X. Fang, and G. Xue, "Truthful incentive mechanisms for k-anonymity location privacy," in *Proc. IEEE INFOCOM*, Turin, Italy, Apr. 2013, pp. 2994–3002.

[40] X. Liu, K. Liu, L. Guo, X. Li, and Y. Fang, "A game-theoretic approach for achieving k-anonymity in location based services," in *Proc. IEEE INFOCOM*, Turin, Italy, Apr. 2013, pp. 2985–2993.

[41] D. Rebollo-Monedero, J. Forné, and J. Domingo-Ferrer, "From t-closeness-like privacy to postrandomization via information theory," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 11, pp. 1623–1636, Nov. 2010.
[42] M. Alaggan, S. Gambs, and A. Kermarrec, "Heterogeneous differential privacy," *CoRR*, 2015. [Online]. Available: http://arxiv.org/abs/1504.06998
[43] V. Goyal, O. Pandey, A. Sahai, and B. Waters, "Attribute-based encryption for fine-grained access control of encrypted data," in *Proc. 13th ACM Conf. Comput. Commun. Secur. (CCS)*, Alexandria, VA, USA, Oct./Nov. 2006, pp. 89–98.
[44] A. B. Lewko and B. Waters, "Decentralizing attribute-based encryption," in *Proc. 30th Annu. Int. Conf. Theory Appl. Cryptogr. Techn.*, Tallinn, Estonia, May 2011, pp. 568–588.
[45] A. Sahai and B. Waters, "Fuzzy identity-based encryption," in *Proc. IACR Cryptol. ePrint Arch.*, 2004, p. 86.
[46] D. X. Song, D. Wagner, and A. Perrig, "Practical techniques for searches on encrypted data," in *Proc. IEEE Symp. Secur. Privacy*, Berkeley, CA, USA, May 2000, pp. 44–55.
[47] R. Curtmola, J. A. Garay, S. Kamara, and R. Ostrovsky, "Searchable symmetric encryption: Improved definitions and efficient constructions," in *Proc. 13th ACM Conf. Comput. Commun. Secur. (CCS)*, Alexandria, VA, USA, Oct./Nov. 2006, pp. 79–88.
[48] R. L. Rivest, L. Adleman, and M. L. Dertouzos, "On data banks and privacy homomorphisms," in *Foundations of Secure Computation*. San Diego, CA, USA: Academic, 1978, pp. 169–179.
[49] C. Gentry, "Fully homomorphic encryption using ideal lattices," in *Proc. 41st Annu. ACM Symp. Theory Comput.*, Bethesda, MD, USA, May/Jun. 2009, pp. 169–178.
[50] V. Vaikuntanathan, "Computing blindfolded: New developments in fully homomorphic encryption," in *Proc. IEEE 52nd Annu. Symp. Found. Comput. Sci.*, Palm Springs, CA, USA, Oct. 2011, pp. 5–16.
[51] A. C. Yao, "Protocols for secure computations," in *Proc. 23rd Annu. Symp. Found. Comput. Sci.*, 1982, pp. 160–164.
[52] Q. Sun, D. R. Simon, Y. Wang, W. Russell, V. N. Padmanabhan, and L. Qiu, "Statistical identification of encrypted Web browsing traffic," in *Proc. IEEE Symp. Secur. Privacy*, Berkeley, CA, USA, May 2002, pp. 19–30.
[53] M. Liberatore and B. N. Levine, "Inferring the source of encrypted HTTP connections," in *Proc. 13th ACM Conf. Comput. Commun. Secur.*, Alexandria, VA, USA, Oct./Nov. 2006, pp. 255–263.
[54] Y. Zhu, X. Fu, B. Graham, R. Bettati, and W. Zhao, "Correlation-based traffic analysis attacks on anonymity networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 21, no. 7, pp. 954–967, Jul. 2010.
[55] P. Venkitasubramaniam, T. He, and L. Tong, "Anonymous networking amidst eavesdroppers," *IEEE Trans. Inf. Theory*, vol. 54, no. 6, pp. 2770–2784, Jun. 2008.
[56] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren, "Information security in big data: Privacy and data mining," *IEEE Access*, vol. 2, pp. 1149–1176, Oct. 2014.
[57] T. Zhu, P. Xiong, G. Li, and W. Zhou, "Correlated differential privacy: Hiding information in non-IID data set," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 2, pp. 229–242, Feb. 2015.
[58] J. Bringer, H. Chabanne, and A. Patey, "Privacy-preserving biometric identification using secure multiparty computation: An overview and recent trends," *IEEE Signal Process. Mag.*, vol. 30, no. 2, pp. 42–52, Mar. 2013.
[59] M. Barni, G. Droandi, and R. Lazzeretti, "Privacy protection in biometric-based recognition systems: A marriage between cryptography and signal processing," *IEEE Signal Process. Mag.*, vol. 32, no. 5, pp. 66–76, Sep. 2015.
[60] L. Lai, S. Ho, and H. V. Poor, "Privacy-security trade-offs in biometric security systems—Part I: Single use case," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 1, pp. 122–139, Jan. 2011.
[61] L. Lai, S. Ho, and H. V. Poor, "Privacy-security trade-offs in biometric security systems—Part II: Multiple use case," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 1, pp. 140–151, Jan. 2011.
[62] T. Ignatenko and F. M. J. Willems, "Fundamental limits for privacy-preserving biometric identification systems that support authentication," *IEEE Trans. Inf. Theory*, vol. 61, no. 10, pp. 5583–5594, Oct. 2015.
[63] A. Dantcheva, P. Elia, and A. Ross, "What else does your biometric data reveal? A survey on soft biometrics," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 3, pp. 441–467, Mar. 2016.
[64] E. J. Candés, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, 2006.
[65] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.
[66] M. H. Manshaei, Q. Zhu, T. Alpcan, T. Basar, and J. Hubaux, "Game theory meets network security and privacy," *ACM Comput. Surv.*, vol. 45, no. 3, p. 25, 2013.
[67] J. Freudiger, M. H. Manshaei, J. Hubaux, and D. C. Parkes, "Non-cooperative location privacy," *IEEE Trans. Dependable Secure Comput.*, vol. 10, no. 2, pp. 84–98, Mar./Apr. 2013.
[68] V. Krishnamurthy and H. V. Poor, "A tutorial on interactive sensing in social networks," *IEEE Trans. Comput. Social Syst.*, vol. 1, no. 1, pp. 3–21, Mar. 2014.
[69] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *Proc. 48th Annu. IEEE Symp. Found. Comput. Sci.*, Oct. 2007, pp. 94–103.
[70] A. Ghosh and A. Roth, "Selling privacy at auction," in *Proc. 12th ACM Conf. Electron. Commerce*, San Jose, CA, USA, Jun. 2011, pp. 199–208.
[71] L. Fleischer and Y. Lyu, "Approximately optimal auctions for selling privacy when costs are correlated with data," in *Proc. ACM Conf. Electron. Commerce*, Valencia, Spain, Jun. 2012, pp. 568–585.
[72] K. Ligett and A. Roth, "Take it or leave it: Running a survey when privacy comes at a cost," in *Proc. 8th Int. Workshop Internet Netw. Econ.*, Liverpool, U.K., Dec. 2012, pp. 378–391.
[73] L. Xu, C. Jiang, Y. Chen, Y. Ren, and K. J. R. Liu, "Privacy or utility in data collection? A contract theoretic approach," *J. Sel. Topics Signal Process.*, vol. 9, no. 7, pp. 1256–1269, Oct. 2015.
[74] A. Acquisti, L. Brandimarte, and G. Loewenstein, "Privacy and human behavior in the age of information," *Science*, vol. 347, no. 6221, pp. 509–514, 2015.
[75] E. Horvitz and D. Mulligan, "Data, privacy, and the greater good," *Science*, vol. 349, no. 6245, pp. 253–255, 2015.
[76] L. Sankar, S. R. Rajagopalan, S. Mohajer, and H. V. Poor, "Smart meter privacy: A theoretical framework," *IEEE Trans. Smart Grid*, vol. 4, no. 2, pp. 837–846, Jun. 2013.
[77] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 2006.
[78] G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*. New York, NY, USA: Wiley, 1992.
[79] S. Yu, T. Thapngam, J. Liu, S. Wei, and W. Zhou, "Discriminating DDoS flows from flash crowds using information distance," in *Proc. 3rd Int. Conf. Netw. Syst. Secur.*, Gold Coast, QLD, Australia, Oct. 2009, pp. 351–356.
[80] W. Liu, P. P. Pokharel, and J. C. Principe, "Correntropy: Properties and applications in non-Gaussian signal processing," *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5286–5298, Nov. 2007.
[81] P. Wendell and M. J. Freedman, "Going viral: Flash crowds in an open CDN," in *Proc. 11th ACM SIGCOMM Internet Meas. Conf.*, Berlin, Germany, Nov. 2011, pp. 549–558.
[82] F. McSherry, "Privacy integrated queries: An extensible platform for privacy-preserving data analysis," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Providence, RI, USA, Jun./Jul. 2009, pp. 19–30.
[83] Q. Geng, P. Kairouz, S. Oh, and P. Viswanath, "The staircase mechanism in differential privacy," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 7, pp. 1176–1184, Oct. 2015.
[84] D. Gottesman and I. L. Chuang, "Demonstrating the viability of universal quantum computation using teleportation and single-qubit operations," *Nature*, vol. 402, no. 6760, pp. 390–393, 1999.
[85] S. Barz, E. Kashefi, A. Broadbent, J. F. Fitzsimons, A. Zeilinger, and P. Walther, "Demonstration of blind quantum computing," *Science*, vol. 335, no. 6066, pp. 303–308, 2012.
[86] *Advances in Computational Psychophysiology*, accessed on May 17, 2016. [Online]. Available: http://www.sciencemag.org/custom-publishing/collections/advances-computational-psychophysiology



SHUI YU (SM'12) is currently a Senior Lecturer with the School of Information Technology, Deakin University. He is a member of the Deakin University Academic Board (2015–2016), a member of AAAS and ACM, the Vice Chair of the Technical Subcommittee on Big Data Processing, Analytics, and Networking of the IEEE Communication Society, and a member of the IEEE Big Data Standardization Committee.

His research interest includes security and privacy in networking, big data, and cyberspace, and mathematical modeling. He has published two monographs and edited two books, over 150 technical papers, including top journals and top conferences, such as the IEEE TRANSACTIONS ON PARALLEL and DISTRIBUTED SYSTEMS, the IEEE TRANSACTIONS ON COMPUTERS, the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, the IEEE TRANSACTIONS ON MOBILE COMPUTING, the IEEE TRANSACTIONS ON KKNOWLEDGE AND DATA ENGINEERING, the IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTING, and the IEEE International Conference on Computer Communications.

Dr. Yu initiated the research field of networking for big data in 2013. He has an h-index of 22. He actively serves his research communities in various roles. He is serving on the Editorial Boards of the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS, the IEEE ACCESS, the IEEE INTERNET OF THINGS JOURNAL, and a number of other international journals. He has served over 70 international conferences as a member of Organizing Committee, such as the Publication Chair of the IEEE Globecom 2015 and the IEEE INFOCOM 2016, and the TPC Co-Chair of the IEEE BigDataService 2015 and the IEEE ITNAC 2015.