IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received October 6, 2020, accepted October 22, 2020, date of publication October 28, 2020, date of current version November 13, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3034398

A Semantic Inference Based Method for Privacy Measurement

BAOCUN CHEN[1], **NAFEI ZHU**[1], **JINGSHA HE**[1,2], **(Member, IEEE)**, **PENG HE**[2], **SHUTING JIN**[1], **AND SHIJIA PAN**[1]

[1]Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
[2]College of Computer and Information Technology, China Three Gorges University, Yichang 443002, China

Corresponding authors: Jingsha He (jhe@bjut.edu.cn) and Peng He (hpeng@ctgu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61602456.

ABSTRACT In the era of Internet and big data, an increasing number of intelligent applications have been developed. As the result, a lot of user data can be collected and stored by Internet companies as well as by ordinary users through various media platforms such as Facebook, WeChat, etc. that may contain information related to personal privacy. Even though privacy protection has been declared by Internet service providers, after collecting enough amount of seemingly less relevant data, an attacker can still infer user privacy via one means or another, e.g., by running a data mining algorithm. This can undoubtedly bring high risk of privacy disclosure to users under such an attack model. So, accurately measuring the leakage of privacy becomes an urgent issue. Although many privacy measurement and protection methods have been proposed in recent years, they mainly target at structured datasets and are thus inadequate to the measurement of the disclosure of specific privacy information. In addition, most of the methods have failed to consider the internal connections and relationships between privacy information and thus cannot be used to measure the implicit privacy disclosure risk on unstructured data. In this paper, we propose a semantic inference method based on the WordNet ontology to measure privacy disclosure in which we employ an information content (IC) based method to determine the weight of attributes to describe the inference preferences in the process of inferring privacy. Experiment was performed to verify the effectiveness of the IC based inference weight assignment method and to compare the proposed measurement method to some privacy disclosure behavior learned through a data mining algorithm and some existing privacy measurement methods to demonstrate the advantages of the proposed method for measuring privacy disclosure.

INDEX TERMS WordNet, semantic inference, privacy inference weight, privacy disclosure, privacy quantification.

I. INTRODUCTION

In recent years, with the advance of computer technology, high-performance computers have been widely applied to big data processing and analysis. Artificial intelligence technology has also been used in an increasing number of applications. Machine learning algorithms can be used to learn and mine valuable information from big data. As the result, Internet companies can provide users with more convenient services. In this process, the quantity and the quality of the training data is one of the most important factor to determine the final effect of the model. In ordinary Internet applications, in order to use some functions, users are usually required to provide some personal information during account setup, such as nickname, gender, age, occupation, etc. With the increase of demand for personalized services, applications begin to collect and store users' behavior data, such as commodity click, commodity collection, concerned bloggers, followers, comments and other data. With this scale of user data, Internet companies can use proper data mining algorithms to analyze users' interest, build user model and provide more personalized services. Recommendation system is a typical application of this kind that aims at providing more accurate and precise recommendations to the user, which can not only improve the application experience, but also enhance the competitiveness of services and bring huge economic

The associate editor coordinating the review of this manuscript and approving it for publication was Mamoun Alazab.

benefits to Internet companies and businesses. However, with the popularity of wide range of such intelligent applications, violation of user privacy starts to become a serious concern. The implicit strategy of data collection makes users less likely to be aware of the situation of the release of their own data. Neither do they likely know their privacy information has been leaked or is at the risk of being leaked. To better protect personal privacy, it is necessary to develop methods to predict personal information disclosure even if such information is not directly collected. In addition, Internet companies very often need to publish some user data to the public for legitimate purposes. In this process, in order to reduce the risk of disclosing personal privacy, they usually apply some algorithms to hide some key information in the data, making it less likely for attackers to uniquely identify specific users. The key questions then become how much privacy information is contained in the processed or published dataset and how much the risk of privacy disclosure is. Answers to these questions have a high guiding value for the development of privacy protection methods. Therefore, as a fundamental issue in the field of privacy protection, privacy measurement has received a great deal of attention in recent years.

Users may not realize that every time they publish or exchange some information through some Internet social media, they may disclose some personal information to the public. Such data can be easily obtained by anyone including malicious attackers through search engines and other ordinary ways. Users may think that only a small portion of personal information is disclosed each time. But the total amount of personal information they disclose intentionally and unintentionally on various applications and platforms may be a big surprise to them. Today, due to the imperfection of privacy protection systems, among other reasons, once users provide their personal data to service providers or to the public, they will lose control over the personal data. In addition, the superposition effect and inference between pieces of personal information make users face even a great risk of privacy disclosure. Many cases of privacy leakage from so-called trusted Internet companies such as Facebook show that the effective way of preventing personal privacy disclosure is to give users the control over the publication or disclosure of their personal data according to their own personal requirements and privacy needs. So, before providing data, being able to measure the amount of personal data contained in the data to be provided is the very important first step for the application of privacy protection methods to limit the disclosure of user information within what users specify, and is the main problem to be investigated and solved in our work.

In this paper, we propose a semantic inference based privacy measurement algorithm by applying the language knowledge map WordNet to find the semantic relationships between privacy information according to its organizational structure, establishing the rules for privacy transfer in different semantic relationships in which information theory is used to determine privacy inference preferences. We also present some experiment results to show that the proposed privacy measurement method can be used to evaluate the specific implicit privacy disclosure risk by considering information inference and superposition effect. It can also be used to find the source of disclosure which can guide the development of algorithms to prevent the leakage of user privacy during data publication or when making access control policy. Experiment results also show that the proposed IC based attribute weight assignment method for privacy inference preference is consistent with human subjective judgment. Comparative experiment results show that the proposed method can improve the accuracy of privacy measurement. Finally, privacy measurement results on unstructured data collected from the Internet for celebrities of different walks of life demonstrate that the proposed privacy measurement method has great practical value.

The remainder of this paper is organized as follows. Section 2 reviews some related work on privacy measurement, the basic structure of WordNet and the information content IC. Section 3 defines the attack model from the perspective of background knowledge that attackers may have acquired. Section 4 introduces the implementation logic of our proposed semantic inference based privacy measurement method and the IC based method for attribute weight assignment. Section 5 describes the experiments and analyzes the results. Finally, Section 6 concludes this paper in which some future research work will also be discussed.

II. RELATED WORK

A. PRIVACY MEASUREMENT

Privacy measurement originated from work on anonymity. Afterwards, various types of measurement methods have been proposed based on information entropy, set pair theory and differential privacy.

K-anonymity is one of the classical methods for privacy measurement. The value k in k-anonymity represents the anonymity degree of the quasi identifier attribute in the dataset. For each single data item, there are at least k-1 other data items that cannot be distinguished from it. So, the probability of recognizing anyone of the k anonymous data items is equal. Therefore, the probability that an attacker who has no prior background knowledge can recognize the privacy information of a user from the k data items is 1/k [1]. The larger the value of k in k-anonymity, the harder it is for the attacker to infer the privacy information and thus the lower the risk of privacy disclosure. However, k-anonymity can only anonymize the quasi identifier attribute in the data without any restriction on other sensitive attributes. Equipped with some background knowledge related to the privacy information, the attacker can still infer the corresponding relationship between the user and the values of sensitive attributes according to the distribution of the values of the sensitive attributes in the anonymous dataset [2]. Therefore, k-anonymity methods for privacy measurement are not comprehensive and accurate enough.

Under the condition of k-anonymity, some research tried to constrain the sensitive attribute values in the dataset by making them evenly distributed so that the anonymity of the dataset is enhanced. As the result, l-diversity emerged [2] which aimed at constraining the sensitive attribute values that appear frequently in anonymous datasets. Each equivalence class contains at least k anonymous quasi identifiers and at least l different sensitive attribute values, making k and l the indicators of the degree of privacy of anonymous data. Li *et al.* [3], [4] proposed a t-closeness privacy measurement method based on k-anonymity and l-diversity to determine the distribution of the sensitive attribute values. Under the condition of k-anonymity, EMD (earth mover's distance) method was developed to compute the difference between the global distribution of sensitive attribute values in the dataset and the distribution of the same sensitive attribute values in any equivalent class, further tightening the standard of privacy measurement. However, Zhang *et al.* pointed out that the EMD method failed to consider the stability of the distribution of sensitive attribute values between equivalent classes and data [5]. To solve the problem, a privacy measurement method called EKD (EMD and KL divergence distance) was proposed based on the EMD method and KL divergence in which the EMD method was used to compute the difference of distributions between sensitive attribute values and KL divergence was used to measure the difference of the stability between adjacent sensitive attribute values. Then, the higher the degree of anonymity of the insensitivity attribute in the anonymous dataset and the smaller the difference of the distribution of the sensitive attribute values in the equivalence class, the lower the possibility of privacy information leakage. In addition, inspired by k-anonymity, a method for measuring privacy information disclosure was proposed based on Bayesian inference [6]–[8]. In this method, binary trees are constructed based on the attacker's background knowledge and on the anonymous data, respectively, and then a binary tree is constructed based on information association deduced from Bayesian inference. The risk of privacy disclosure can be measured by analyzing and comparing the difference between conjectured information and privacy information.

Information entropy has been widely applied for information quantification, which has made great contribution in the field of communication. Since privacy is a special type of information, it can also be quantified by entropy. Díaz *et al.* were among the first who proposed to use information entropy to measure the anonymity of anonymous communication systems [9]. Clauß *et al.* used information entropy to describe the uncertainty of privacy information in datasets [10]. Hoh *et al.* proposed a new time confusion measure to represent the privacy of anonymous location traces based on the information entropy [11]. Ma *et al.* used information theory to quantify the level of location privacy for each user by quantifying the uncertainty of location information and specific personal contact [12], [13]. Shokri *et al.* proposed a privacy measurement method based on distortion, which reflects the user's privacy level by comparing the difference between the tracking user's trajectory observed by the attacker and the user's real trajectory [14]. In 2011, Chen *et al.* proposed to use conditional entropy to measure the query privacy in location-based services (LBS) to measure the query privacy of users in LBS [15]. Yang *et al.* proposed to use entropy to measure the threat of two types of attacks to network users in which the two types of attackers were defined based on the sensitive information of network access to identify personal identity [16]. In 2016, Peng *et al.* proposed several privacy protection information entropy models as well as a general privacy measurement method from a theoretical view after describing the privacy protection system as a communication model to make the measurement of information entropy more intuitive [17]. However, due to the spatiotemporal, subjective and fuzzy nature of privacy information, the design of a more suitable privacy information entropy model still remains a theoretical problem to be solved.

There is also a privacy measurement method based on set pair analysis theory. Set pair analysis theory [18] is a set pair of mutual relations, constraints and influences between two sets that have certain connections. By establishing the same, different and anti-association coefficients, it can depict the determination and uncertainty of the common attributes of things. In 2015, Yan *et al.* proposed a set pair analysis method of user privacy protection measurement [19]. Under three different application modes of database privacy protection, location privacy protection and trajectory privacy protection, the system standard and content of privacy measurement were established.

However, further research in privacy protection indicated that there are two main defects in privacy protection and measurement models. First, these models cannot provide enough security, they are always in need of continuous improvement due to the emergence of new attacks. Second, the early privacy measurement models could not provide an effective and strict method to prove the level of privacy protection. In 2006, Dwork proposed a new definition of privacy for database, i.e., differential privacy [20], based on a solid mathematical foundation. It can not only realize privacy protection, but also provide a quantifiable method to evaluate the risk of privacy information disclosure. Differential privacy was quickly recognized by the industry and provided a new research direction in privacy protection. Many measurement methods based on differential privacy have since emerged, such as the differential privacy measurement method based on multi-dataset association [21] and that based on mutual information [22].

B. WordNet AND INFORMATION CONTENT

WordNet is a research project [23] of Princeton University. It includes a large vocabulary and merges concepts with the same meaning into synonyms synset. These synonyms are represented by nodes in the WordNet network structure and these nodes are connected by corresponding semantic relationships. Since privacy information is concentrated on the concept of nouns, this paper only considers the nouns in the

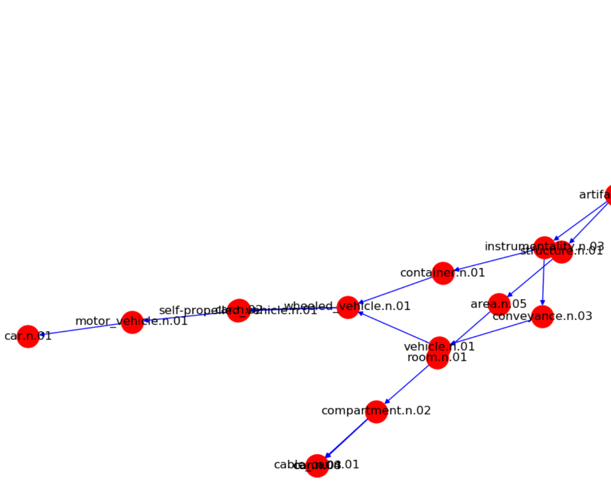

FIGURE 1. Example of the hierarchical structure.

WordNet structure in order to improve the efficiency of computation. The connections or relationships considered in this paper include the following 11 types: hypernym, hyponym, instance hypernym, instance hyponym, part holonym, part meronym, member holonym, member meronym, substance holonym, substance_meronym and attribute. Each relationship has its corresponding opposite semantic relationship except the attribute relationship. For example, hypernym and hyponym are reciprocal relationships, which are the most frequent connections in the WordNet. Through these two relationships, one can traverse all the noun nodes in the WordNet. Figure 1 shows the hierarchical structure of all nodes connected by hypernym and hyponym from the leaf node car. n.01.

Because it contains a lot of language knowledge, WordNet has been widely used in many semantic analysis tasks. Semantic similarity calculation of words is a representative application of the WordNet. In the early days of applying WordNet to determining semantic similarity, most methods proposed only focused on some simple graph features of the WordNet, such as the distance between two nodes, the closest common ancestor of two nodes or the depth of nodes in the hierarchical structure, and good performance was achieved. Later, information content IC of Shannon's information theory [24] was introduced into similarity calculation in which IC was defined using Formula (1):

$$\mathrm{IC}(x) = -\log_2 p(x) \tag{1}$$

The idea is that information is used to eliminate uncertainty. Therefore, the higher the uncertainty it can eliminate, the more information it contains. Resnik proposed an information content-based similarity measure in 1995 following information theoretic approach [25]. The idea is that for two given concepts, similarity depends on the extent to which they share common information. The more information two concepts share, the more similar they are. So, the WordNet structure could be used to find the common parent node c first, then IC(c) is used to calculate the degree of similarity between two concepts (c1 and c2) using Formula (2).

$$\mathrm{Sim}_{(\mathrm{Re}snik)}(c1, c2) = \underset{c \in common(c1,c2)}{Max}(IC(c)) \tag{2}$$

Although many IC and WordNet based word similarity algorithms have been proposed, there is still a problem in IC calculation, i.e., the calculation of p(c) that needs to be obtained from the statistical calculation of corpora. However, different or different amount of corpora will result in different calculation results. To solve the problem, Li regarded WordNet as a very stable corpus abstracted from a large number of corpora so that p(c) can be calculated by using only the WordNet itself [26]. We follow the same assumption in the development of our measurement method in which we improve the accuracy of the calculation by proposing a weight assignment method during privacy inference.

Intuitively, the higher the semantic similarity between two concepts, the higher the probability of information leakage of one concept when the other concept is leaked. However, Zhu *et al.* found that the correlation between calculation results and scores obtained through questionnaires was somewhat low [27], implying that it was not enough to describe the association and transmission probability of privacy information by only considering the similarity of words in the WordNet.

Although a lot of privacy measurement methods have been developed, most of them are applicable to structured data stored in relational databases. Neither do such methods adequately consider the content of personal information. Since today's data on the Internet is generally large-scale and mostly unstructured, a lot of information related to user privacy exists explicitly or implicitly and with superposition effects and inference relationships, which also poses a great risk of privacy disclosure that hasn't been paid much attention in previous research on privacy measurement. The development of effective methods to measure privacy disclosure has thus become a hot research. WordNet is a tool that may help in establishing the internal links between information and in providing the inference logic just like what attackers may use during attacks to deal with the issue of privacy measurement. Although some attempts have been made to evaluate the suitability of applying some existing WordNet based semantic methods to privacy measurement, few methods have been proposed for effective measurement. It is our goal in this paper to explore the semantic knowledge in the WordNet to conduct privacy measurement, leading to proposing an effective semantic based privacy measurement method and its accuracy promotion method.

III. THE ATTACK MODEL

The attacker model is very important in the design of privacy protection and privacy measurement method because the ability of the attacker has a direct impact on the risk of privacy information disclosure. In most current privacy measurement research, the attacker model is mainly constructed from the perspective of whether and how much the attacker has background knowledge [28], [29]. The more background knowledge the attacker has, the greater the risk of privacy disclosure. Therefore, description of the background knowledge the attacker has can help make the description of the risk of privacy disclosure more accurate.

The background knowledge $\mathbf{K_{attacker}}$ that the attacker may have can be divided into three categories. First is the prior knowledge $\mathbf{K_{prio}}$ which includes the public information already spread on the Internet, such as user name, gender, etc. This type of knowledge can be easily acquired by using search engines. Such knowledge could also include user data collected explicitly or implicitly by service providers. Second is the observed knowledge $\mathbf{K_{ob}}$ that the attacker can obtain or intercept by learning using language knowledge, query content, sensitive attributes, etc. Such knowledge can also include results of using prediction algorithms. Third is the posteriori knowledge $\mathbf{K_{post}}$, i.e., the new knowledge that the attacker can get through inference and prediction based on the prior and the observed knowledge.

Therefore, in the attack model to be used in this paper, the knowledge that can be obtained through successive attacks can be defined using Formula (3).

$$\mathbf{K_{attacker}} = \mathbf{K_{prio}} \cup \mathbf{K_{op}} \cup \mathbf{K_{post}},\ \mathbf{K_{post}} \infty \{\mathbf{K_{prio}}, \mathbf{K_{op}}\} \quad (3)$$

We will describe the process of privacy-oriented attacks by using the above attack model. First, we assume that the attacker has already collected a certain amount of prior knowledge $\mathbf{K_{prio}}$ via legal or illegal means. Then, the attacker evaluates whether $\mathbf{K_{prio}}$ is enough for conducting the privacy attack. If it is insufficient, the attacker can get more user information $\mathbf{K_{ob}}$ by launching network attacks or by impersonating legitimate users to request data services. In addition, data mining can also be applied to derive more background knowledge, making it another means of attack. Finally, the attacker will infer the posterior knowledge $\mathbf{K_{post}}$ based on $\mathbf{K_{prio}}$ and $\mathbf{K_{ob}}$ to derive user's privacy information $\mathbf{K_{post}}$, resulting in privacy leakage. After the process is complete, the sum of $\mathbf{K_{prio}}$, $\mathbf{K_{ob}}$ and $\mathbf{K_{post}}$ resulting from $\mathbf{K_{attacker}}$ will become the prior knowledge of the attacker for the next round of the attack.

IV. THE PROPOSED METHOD FOR PRIVACY MEASUREMENT

A. SEMANTIC INFERENCE BASED PRIVACY MEASUREMENT

The inference and superposition effect of personal information is an important cause of privacy disclosure. The inference of information may cause more information to be leaked based on one piece of leaked personal information. The superposition effect may also cause more information to be leaked than the two or more pieces of personal information that has already been leaked. For example, if the identity information of the privacy subject reveals that the subject is a father, then it can be inferred that the subject is a male, has family and has one or more child.

However, if two pieces of information connected by the hypernym relationship are published, the amount of information that becomes known is less than the sum of the information contained in the two original nodes, which is the result of the relationship between personal information. As it is well known, there is always some kind of connections, whether strong or weak, between two information points. Therefore, adequately quantifying the internal relationships between personal information becomes essential. Expert opinion may be feasible, but it suffers the problem of low efficiency and thus can hardly work for large-scale datasets. As an English dictionary, WordNet can provide semantic relationships between a large number of words, which can also express the logical relationship between information. Therefore, using the 11 semantic relationships contained in the WordNet to establish appropriate rules on the probability of privacy transfer, we can perform the quantification of information inference with the superposition effect consideration. Moreover, the graph structure and the access interface of the WordNet make it easy to traversing the nodes and performing automatic execution.

Regarding the probability of privacy inference, we define different inference probabilities for different semantic relationships. For instance, under the hyponym relationship, the child node not only inherits all the information of the parent node, but also contains some more information since the child node is more specific while the parent node is more abstract. If the information represented by the parent node is known, the probability that a child node also belongs to the subject is 1/n, i.e., $\mathrm{P}(\mathrm{node}_{\mathrm{child}}|\ \mathrm{node}_{\mathrm{father}}) = 1/\mathrm{n}$, where n is the number of children nodes. On the other hand, when the child node is known, the information represented by the parent node can be inferred with probability of 100%. The transfer probabilities of all 11 semantic relationships in the WordNet are shown in Table 1, where the privacy disclosure value $\mathbf{V_{disc}} \in [0, 1]$ is defined as follows.

TABLE 1. The transfer probability of all 11 kinds of semantic relationships in the WordNet.

Relationship	$\mathbf{V}_{\mathbf{disc}}$ (start node)	$\mathbf{V}_{\mathbf{disc}}$ (end node)	Explanation of n	Example
hypernym	1	1	—	bus.n.04 **is a kind of** car.n.01
hyponym	1	1/n	Number of children nodes	
instance_hypernym	1	1	—	shanghai.n.01 **is an instance of** city.n.01
instance_hyponym	1	1/n	Number of children nodes	
part_holonym	1	1/n	Number of father nodes	accelerator.n.01 **is a part of** car.n.01
part_meronym	1	1	—	
member_holonym	1	1/n	Number of father nodes	tree.n.01 **is a member of** forest.n.01
member_meronym	1	1	—	
substance_holonym	1	1/n	Number of father nodes	aluminum.n.01 **is a substance of** bauxite.n.01
substance_meronym	1	1	—	
attributes	1	1	—	heavy.a.01 **is an attribute of** weight.n.01

$\mathbf{V}_{\mathbf{disc}} = 0$ means that the privacy information has not been disclosed at all while $\mathbf{V}_{\mathbf{disc}} = 1$ means that the privacy information has been completely disclosed.

In order to measure the privacy disclosure of a specific node, we need to traverse all the simple acyclic paths between known information points and the target node, which are connected by the 11 semantic relationships shown in Table 1. Such paths represent the process of privacy disclosure from the known information nodes to the target node. Since the Python language used in our implementation puts restrictions on the maximum number of recursions, we use two stacks in the design and implementation of a non-recursive path search algorithm based on the depth-first principle of graph search. Figure 2 shows the main logic for traversing the nodes in the WordNet.

Note that in the length of any path traversed is set to a maximum of 14 to make our algorithm fairly inefficient. Our analysis indicated that since WordNet is a connected graph, for any two information nodes, there is always at least one path between them even if there is no obvious semantic relationship between the two. However, as the length of the path increases, more and more nodes bifurcate, leading to the exponential growth of the number of paths to be traversed and sharp decrease in the efficiency of the algorithm. Figure 3 shows the contribution of the four nodes age.n.01, bachelor's degree.n.01, married.n.01 and occupation.n.01 to the leakage of target node wage.n.01 when they are connected with paths of different lengths, indicating that the longer the path, the lower the contribution of a node to the leakage of the target node and that the leakage of the target node is mainly attributed to the semantic transfer path with the length within the range of 10-13. Therefore, to balance between the accuracy and efficiency of ontology-based privacy measurement, we chose to limit the maximum length of path in the traversal to 14 nodes.

Even if we limit the maximum length of path to 14, the process of path traversal is still time-consuming due to the sheer size of the WordNet. To facilitate our experiment and improve the efficiency of the measurement, we performed off-line path traversal calculation for all possible pairs of information nodes and saved the results in the form of [node, relationship, node] in a database for subsequent experiment on the measurement of privacy disclosure. In the process of measuring privacy leakage, we would extract all semantic transfer paths between known information points and the target node directly from the database and then calculate the results of privacy measurement based on the information transfer probability under different semantic relationships. If there have duplicate path between the paths that get from the traverse process between the target node and different known

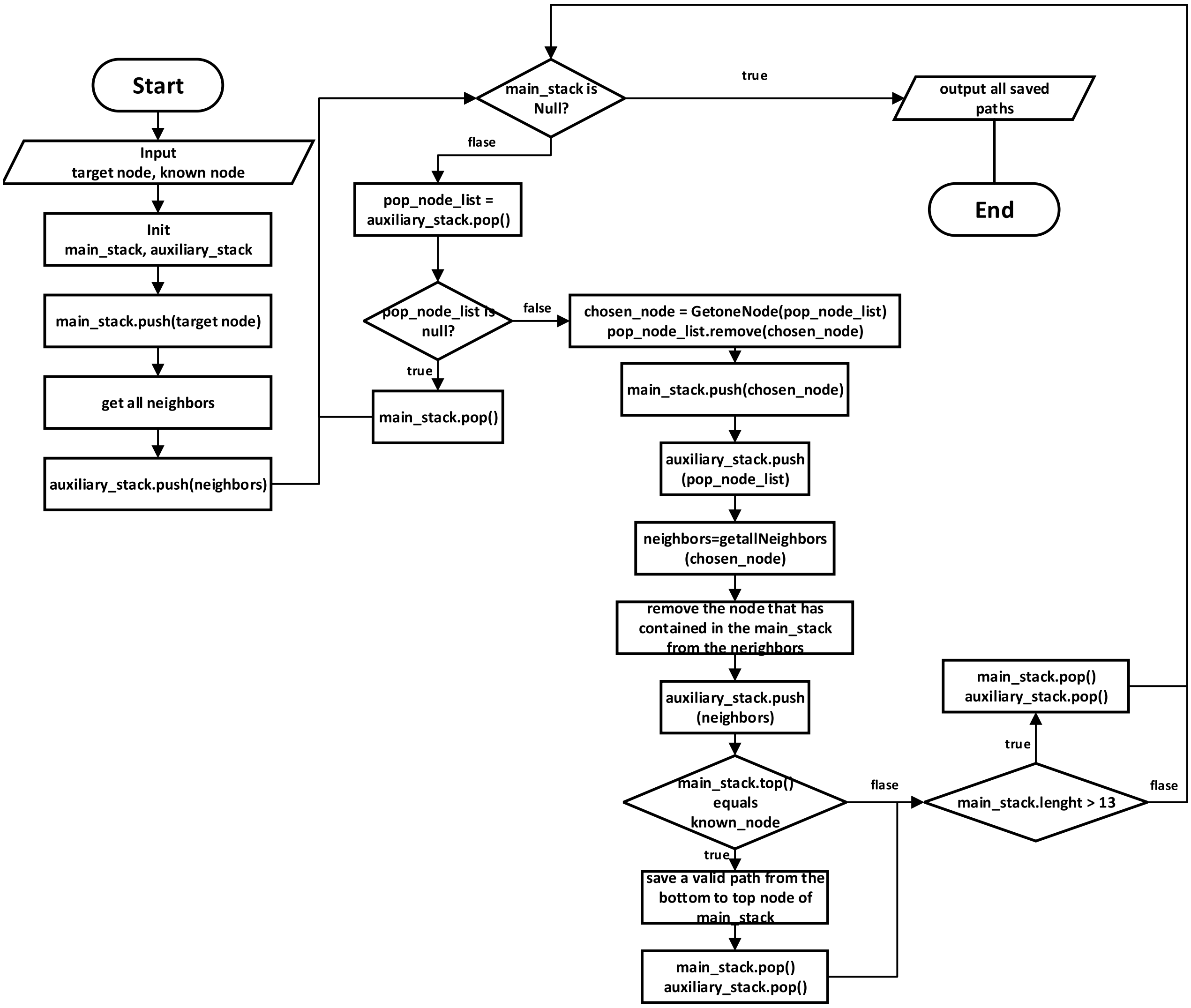
Start
Input
target node, known node
Init
main_stack, auxiliary_stack
main_stack.push(target node)
get all neighbors
auxiliary_stack.push(neighbors)
main_stack is Null?
true
output all saved paths
End
flase
pop_node_list = auxiliary_stack.pop()
pop_node_list is null?
false
true
main_stack.pop()
chosen_node = GetoneNode(pop_node_list)
pop_node_list.remove(chosen_node)
main_stack.push(chosen_node)
auxiliary_stack.push (pop_node_list)
neighbors=getallNeighbors (chosen_node)
remove the node that has contained in the main_stack from the nerighbors
auxiliary_stack.push (neighbors)
main_stack.top() equals known_node
flase
main_stack.lenght > 13
flase
true
main_stack.pop()
auxiliary_stack.pop()
true
save a valid path from the bottom to top node of main_stack
main_stack.pop()
auxiliary_stack.pop()

FIGURE 2. Procedure for finding the internal semantic relationships between the target node and given information nodes.

information nodes, we would only retain the path that contributes the most to the disclosure of privacy. Figure 4 illustrates the measurement process of the proposed semantic inference based privacy measurement method which is divided into offline and online parts, and reduces the calculation time of hundreds samples to several seconds.

The semantic inference based privacy measurement method can be further optimized. For example, the two nodes federal government.n.01 and state government.n.01, both of which are the children nodes of government.n.01 in the hyponym relationship, are located in the relatively same position in the WordNet hierarchy. Each of the semantic transfer path from one of the two nodes to a target node such as wage.n.01 is the same. Since WordNet is an unweighted graph, such connections would result in the same contribution from the two nodes to the leakage of wage.n.01. This may not be in line with the privacy inference process since there may be big difference between working for state government and for federal government in terms of the salary. It is therefore necessary to develop a method for assigning weights to different paths to express the inference preference.

B. DETERMINATION OF PRIVACY INFERENCE PREFERENCE WEIGHT

Computing the preference in the process of privacy inference is essentially to solve the conditional probability of determining new information under the condition that some other information has become known. The higher the conditional probability, the higher the probability of selecting the transfer path. We assume that the attacker would always prefer such a transfer path. In this paper, we design a weight assignment method based on information content (IC).

According to Formula 1, the higher the probability of event x, the lower its information content. When the probability of the occurrence of event x is 1, the event is inevitable and the uncertainty becomes 0, so IC(x) is reduced to 0.

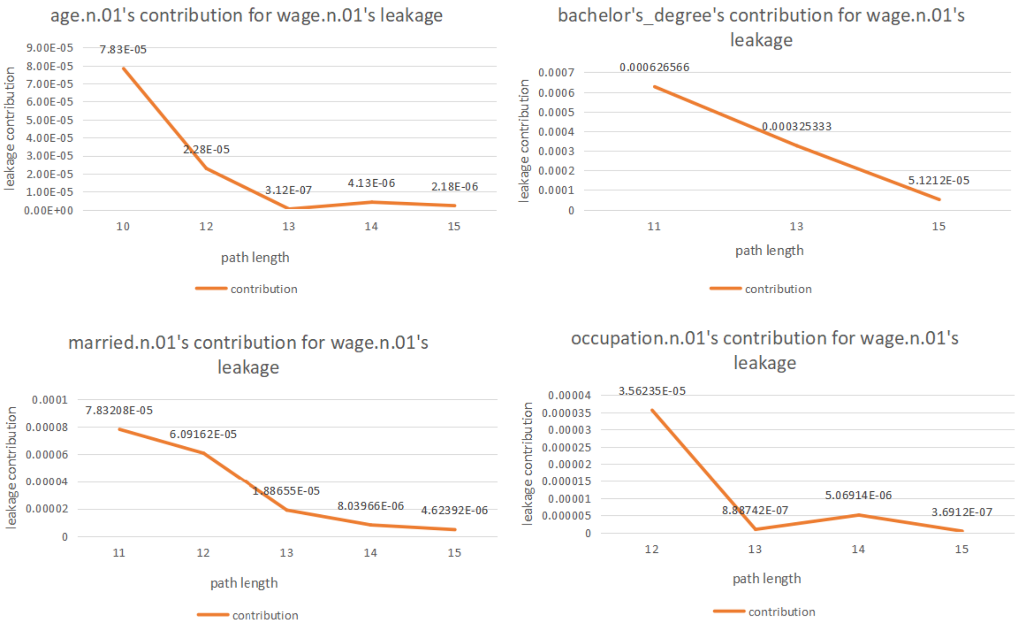

FIGURE 3. Contribution to privacy leakage vs. the length of path.

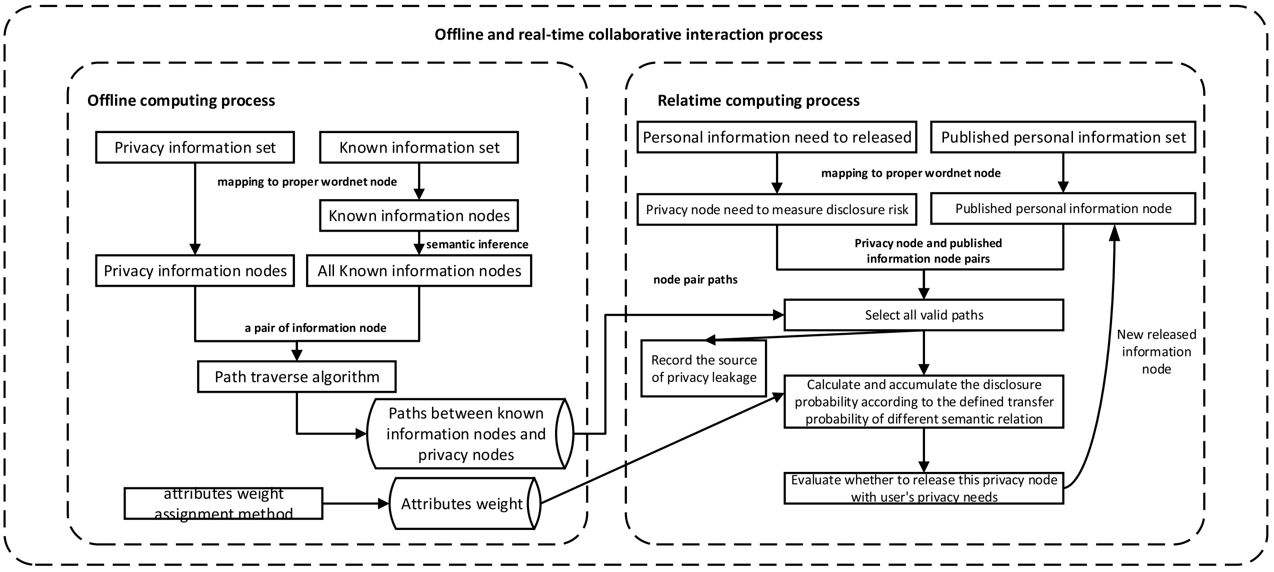
Offline and real-time collaborative interaction process
Offline computing process
Privacy information set
Known information set
mapping to proper wordnet node
Known information nodes
semantic inference
Privacy information nodes
All Known information nodes
a pair of information node
Path traverse algorithm
Paths between known information nodes and privacy nodes
attributes weight assignment method
Attributes weight
Relatime computing process
Personal information need to released
Published personal information set
mapping to proper wordnet node
Privacy node need to measure disclosure risk
Published personal information node
Privacy node and published information node pairs
node pair paths
Select all valid paths
Record the source of privacy leakage
Calculate and accumulate the disclosure probability according to the defined transfer probability of different semantic relation
New released information node
Evaluate whether to release this privacy node with user's privacy needs

FIGURE 4. Semantic inference based privacy measurement method.

On the contrary, when the probability of the occurrence of event x is low, its uncertainty and IC(x) is high. Since the probability of occurrence p(x) of event x is the only parameter for determining IC(x), calculating p(x) is the key to calculating IC(x). In calculating word similarity, p(x) represents the probability of the occurrence of concept x in a large number of corpora, which is generally obtained by the statistics the corpora. To eliminate the computational differences caused by different corpus selection, IC(x) can also be calculated by only using the node features in the structure of WordNet ontology without calculating p(x). In other words, we can calculate the statistical probability of nodes in the corpora in reverse. Since the path with the highest probability is assumed to be chosen by the attacker to conduct privacy inference, we can get the conditional probability of mutual reasoning according to the semantic transfer probability of different relationships in the WordNet to complete the setting of the privacy reasoning preference.

Consequently, under the assumption that the classification of concepts and organizational relationship of WordNet is reasonable, IC can be used to calculate the weight of the privacy information inference preference by the attacker.

The IC calculation method that we propose to use in this paper which is expressed in Formula 4 originally proposed by Meng *et al.* [30] only relies on the features of nodes in the WordNet without being affected by any external corpus.

$$\mathrm{IC}(c) = \frac{\log(deep(c))}{\log(deep_{\max})} * (1 - \frac{\log(\sum\limits_{a \in hypo(c)} \frac{1}{deep(a)} + 1)}{\log(node_{\max})}) \quad (4)$$

In the formula, c represents the node to be calculated, deep(c) represents the depth of the tree hierarchy where node c resides, $\mathrm{deep}_{\mathrm{max}}$ represents the maximum depth of the tree hierarchy, hypo(c) represents the set of hyponyms of node c, and $\mathrm{node}_{\mathrm{max}}$ represents the maximum number of nodes in the tree hierarchy.

We can now calculate the different conditional probability with the statistical probability of nodes and the different connection relationships between nodes. For example, under the condition that the information represented by the child node belongs to a subject, the conditional probability that the information represented by the parent node also belongs to the subject is still 1, which is the same as the case in which there is no weight.

$\mathrm{P(Node_{father}|\ Node_{child})}$
$= \mathrm{P(Node_{father},\ Node_{child})/P(Node_{child})}$
$= \mathrm{P(Node_{child})/P(Node_{child})} = 1$

However, when the information represented by the parent node is known to belong to the subject, the conditional probability will be different and not equal to 1/N. The inference process is as follows:

$\mathrm{P(Node_{child}|\ Node_{father})}$
$= \mathrm{P(Node_{father}, Node_{child})/P(Node_{father})}$
$= \mathrm{P(Node_{child})/P(Node_{father})}$
$= \mathrm{e}^{-\mathrm{IC(child)}} / \mathrm{e}^{-\mathrm{IC(father)}}$
$= \mathrm{e}^{\mathrm{IC(father)}-\mathrm{IC(child)}}$

Finally, by normalizing the conditional probability of the nodes at the same level, the distribution of the weights to the nodes can be completed. Take the node transfer of hyponym of relation as an example, the inference preference can be obtained by using Formula 5 in which Formula 4 is used as the IC calculate method.

$$\mathrm{weight}(c1) = \frac{e^{IC(father)-IC(c1)}}{\sum\limits_{c \in children} e^{IC(father)-IC(c)}} \quad (5)$$

According to the above analysis, we can calculate the assignment of attribute weight similarly for the other defined semantic transfer relationships to simulate the preference of the attacker inference in the process of conducting privacy attacks. It is clear that the inference probability of different attributes in the same level is obviously different since this is the empirical behavior of humans and attackers. There is no exact criteria to quantity the difference of inference. The proposed attribute weight assignment method is actually inspired by the use of IC and WordNet in some other semantic analysis work. The experiment that follows will serve to demonstrate the effectiveness of the proposed privacy measurement method.

Since the proposed method is only related to the structure of the WordNet, all the weights can be calculated and determined offline. After the calculation completes, the edges with 11 connection relationships considered in this paper can be associated with weight values. The weight information can also be stored in the database and then retrieved when they are needed in the privacy measurement algorithm.

V. EXPERIMENT AND ANALYSIS

A. VERIFICATION OF IC-BASED ATTRIBUTE WEIGHT

We select some common privacy concepts in different positions of the hierarchical tree and analyze the performance of weight calculation. The experiment considers the following two scenarios: (1) nodes in deeper positions of the hierarchy and (2) nodes in shallow positions of the hierarchy.

Take the hyponym relationship as an example, the deeper the position of a node in the hierarchy, the more specific the node. The depth of the sub-nodes tends to be the same as the number of hyponyms nodes, making IC the same and the weight distribution tend to be the same. Figure 5 shows the weight calculation results of nodes bus.n.01 and car.n.01 in which the darker the color of an edge, the greater the weight. Because both nodes are very close to the leaf level, the weight distribution is the average distribution.

It should be noted that not all the results calculated are consistent with the subjective understanding of the actual preference of privacy reasoning. For example, among all the hyponym children nodes of car.n.01, cab.n.03 gets the largest weight. This is because the number of hyponyms of cab.n.03 in the WordNet is more than that of sedan.n.01, causing its IC to be relatively small and thus the probability of occurrence to be higher. Therefore, the weight becomes higher. However, in the actual classification of car these days, the number of hyponyms of sedan.n.01 is often more than that of cab.n.03. This inconsistency can be attributed to the lack of more complete knowledge in the WordNet, causing some small deviations in the calculation results. Fortunately, most such problems happen to the deeper nodes where the weight distribution results are not much different from the average distribution. Therefore, it will not cause too much impact on the overall reasoning process.

As the position of a node gets shallower and thus more abstract, the depth and the density of its sub-nodes tend to be more differentiating, making the amount of information more different and the weight distribution more differentiating. As shown in Figure 6, the weight distribution of the nodes tends to be more differentiating and on both sides of the average value and the difference can be as large as in different magnitudes. The three nodes with the largest weights for dimension.n.01 are length.n.01, width.n.01 and height.n.01.

Similarly, the most likely transfer paths concept temperature.n.01 are hotness.n.01 and coldness.n.01. For occupation.n.01, it is position.n.06 while for concept vehicle.n.01, it is wheeled_vehicle.n.01. Concept entity.n.01, which is the root node of all the WordNet nodes and thus located at the

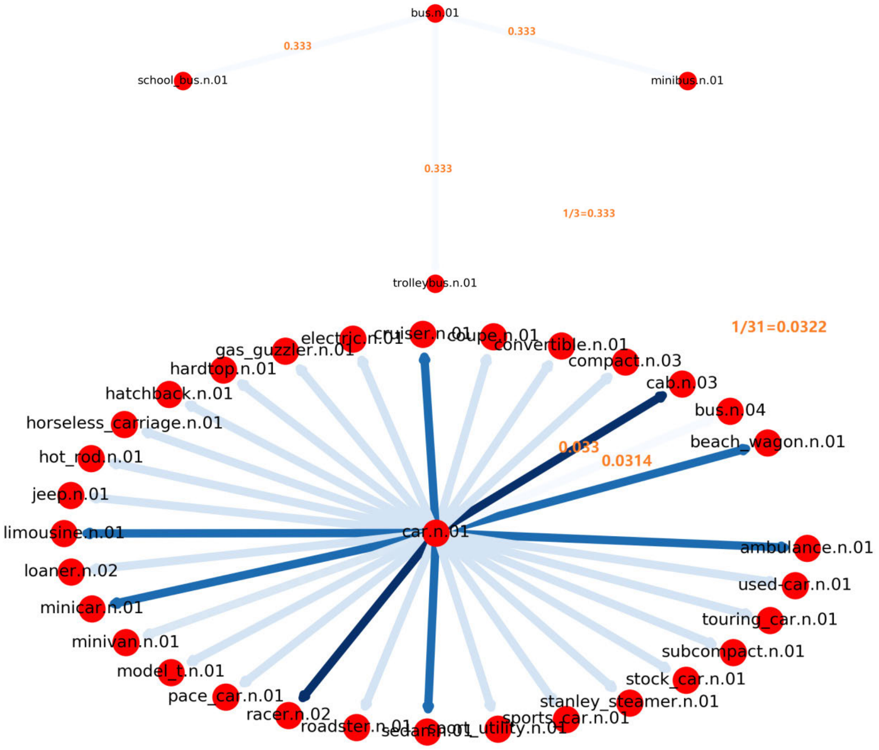

FIGURE 5. Weight allocation of nodes in deep positions of the hierarchy.

shallowest position, also shows weight differences. The high weight values correspond to high probabilities of occurrence, which is in line with the intuitive understanding of human beings, and can also describe the inference preference of privacy attacks.

B. VERIFICATION OF THE PROPOSED PRIVACY MEASUREMENT METHOD

It is not easy to verify the effectiveness of privacy measurement because of the characteristics that privacy is abstract and could be subjective. Although it is possible to analyze the rationality of specific measurement results through intuitive human analysis, this approach could be neither rigorous nor efficient. Consequently, a new verification strategy is used to verify the effectiveness of the proposed privacy measurement algorithm based on semantic inference.

We use the open Adult dataset of UCI Machine Learning Repository [31] to provide the experiment dataset, which was extracted from the 1994 census database by Barry Becker. The dataset contains age, work class, education and other personal information, and is often used to predict whether the annual salary of a given subject is above 50K.

1) SUBJECTIVE ANALYSIS OF PRIVACY DISCLOSURE

In order to have an intuitive understanding of the contribution of different information points to the leakage of a target node or privacy concept, we extracted all the information in the Adult dataset and map it to the nodes in the WordNet. We then took each and every node as the $\mathbf{K_{attacker}}$, which simulates the situation that the corresponding information has become known to the attacker, and measured the degree of leaking the target node wage.n.01 using the proposed privacy measurement algorithm. The results are shown in Figure 7. We can see from the figure that the contribution to the leakage of wage.n.01 mainly comes from education, family, marriage and occupation. The influence of age and gender is lower. Therefore, to prevent wage.n.01 from being leaked to the general public, measures should be taken to protect information related to education, family, marriage, etc. from being leaked.

2) OBJECTIVE EVALUATION OF THE PROPOSED PRIVACY MEASUREMENT METHOD

At present, in many data science competitions and company research or development, researchers use data mining techniques to predict certain information of users due to their high accuracy, which can undoubtedly be viewed as a process of privacy attack which may cause the leakage of the privacy of a large number of users. Therefore, on the basis of subjective analysis, we propose an objective strategy to verify the effectiveness of privacy measurement algorithm, which is more persuasive.

Firstly, we divide 3/4 of the dataset randomly into the training set and the validation set. Based on the personal information of different users, such as age, education, occupation, etc. in the training set, we can obtain a classification decision tree by applying the CART algorithm [32] using

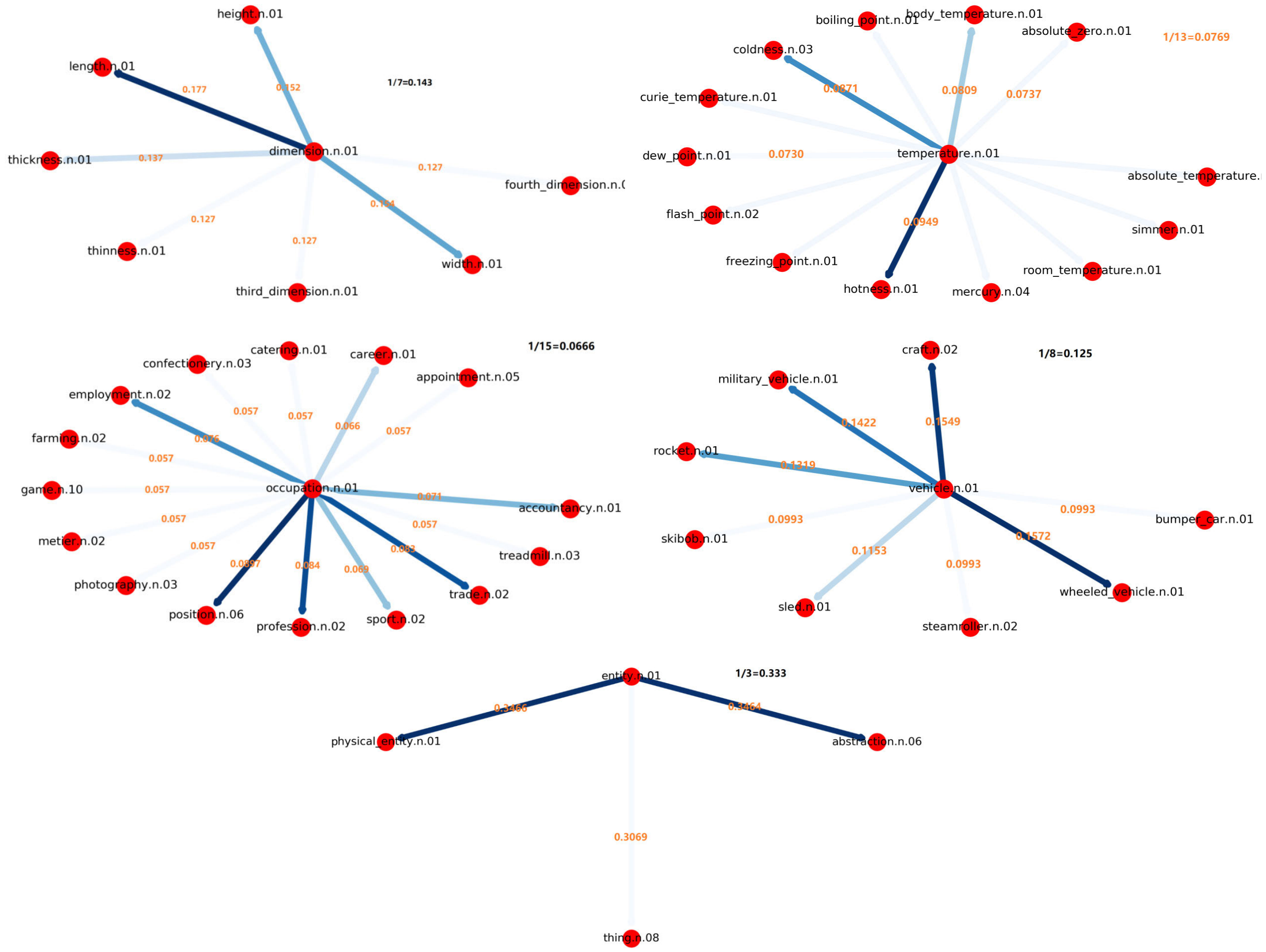

FIGURE 6. Weight allocation of nodes in shallow positions of the hierarchy.

the divided training set that has strong interpretability in the process of predicting the user's wage status (>50K or <50K). The trained model can achieve accuracy of 85.042% in the validation set. As the result, the wage information of the user is leaked to some degree.

Secondly, we selected two sets of samples derived from the trained decision tree model, i.e., the samples of correct prediction (True class samples) and samples of incorrect prediction (False class samples) of the user's wage, in our experiment. We then mapped the information in the samples to the nodes in the WordNet and measured the degree of leakage of wage.n.01 using the proposed privacy measurement algorithm with inference weight. Figure 8 shows the measurement results based on the two sets of samples from which we can see that the accuracy of privacy measurement is relatively high when using the True class samples and relatively low when using the False class samples with mean values being 0.04516380 and 0.02010454, respectively. The results demonstrate that the proposed privacy algorithm can find the privacy inference path and quantify the leakage of privacy well.

We also measured the degree of privacy leakage using the two sets of samples without the inference weight and the results are shown in Figure 9. Although differences can also be observed between the two sets of samples as in Figure 8, many samples are not distinguishable by the proposed measurement algorithm without inference weight. So, inference weight can improve the results of privacy measurement.

In both the data mining method and our proposed method, different information plays different importance. The value or degree of contribution of any information point can be easily calculated by analyzing the contribution path in the proposed privacy measurement method. In the data mining, however, since every decision path is visible, we can calculate the feature importance through entropy reduction brought by different nodes in the decision path according to Formula (6), as shown at the bottom of the next page.

The two types of contribution value are in different orders of magnitude, making it meaningless to compute correlations directly. Rather, ranking privacy leakage contributions of different information nodes can better reflect the correlations of nodes. So we analyze the degree of correlation between the

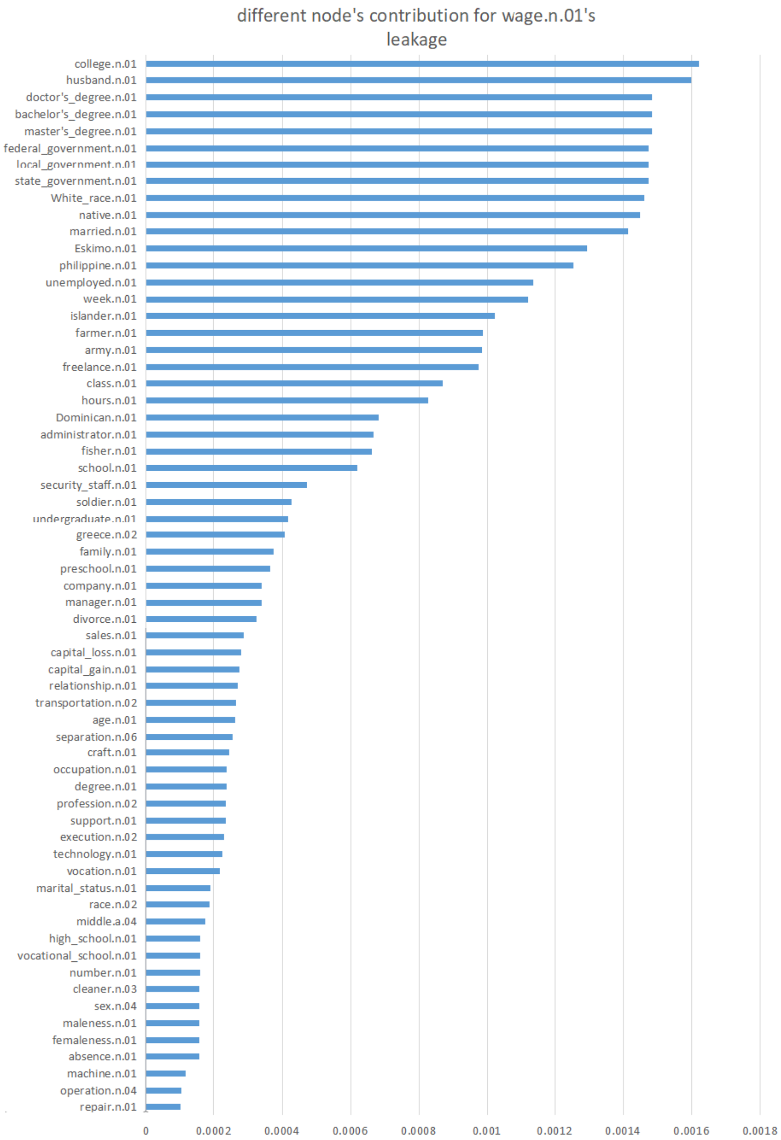

FIGURE 7. Ranking of the contribution of information to the leakage of wage.n.01.

two contribution results using Spearman's Correlation Coefficient which is commonly used to calculate the correlation coefficient of two ranking vectors. Taking the first sample as an example, the results about the ranking of the contribution of different information points computed using the decision tree and the proposed method are shown in Table 2.

We can see that the ranking of the contributions for marital status, relationship and occupation is similar. However, the contributions of capital gain are obviously different, which can be attributed to the lack of such knowledge in the WordNet. Native country gets the largest contribution value in the proposed privacy measurement method because there are just too many country related concepts in the WordNet, resulting in the contribution of all other nodes being suppressed. In addition, different countries have different number of related concepts due to the lack of a unified standard. For instance, U.S. has a lot more related concepts than other countries. Therefore, it should be removed when comparing

$$contribution = \frac{(node\_impurity * node\_samples - leftchild\_samples * leftchild\_impurity - rightchild\_samples * rightchild\_impurity)}{all\_node\_samples} \tag{6}$$

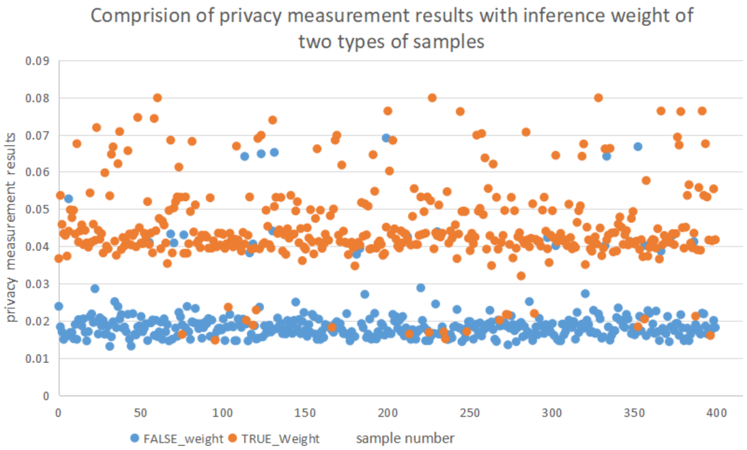

FIGURE 8. Results of privacy measurement with inference weight based on two sets of samples.

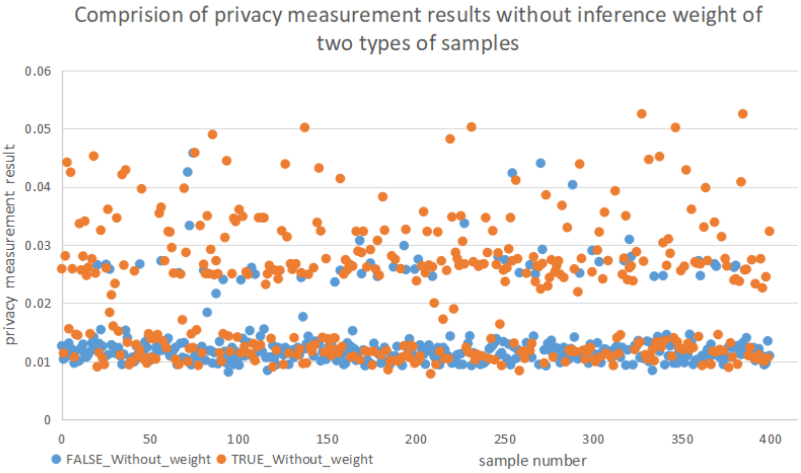

FIGURE 9. Results of privacy measurement without inference weight based on two sets of samples.

TABLE 2. Comparison of the ranking of the contribution to privacy leakage.

Decision tree		Proposed privacy measurement method	
Feature	Privacy leakage contribution	Feature	Privacy leakage contribution
marital status	0.089179279	native country	0.024855043
capital gain	0.0095381339	marital status	0.009211486
relationship	0.005647401	relation ship	0.001619248
occupation	0.001362118	occupation	0.000541957
native country	0.000208709	capital gain	0.000457241

the privacy leakage contribution of people from different countries. We can get value 1.0 as the Spearman's correlation coefficient by dropping native_country and capital_gain information. As the result, the average correlation value of the 400 selected True class samples becomes 0.7407. As the result, the proposed privacy measurement algorithm based on semantic reasoning is more effective according to the above experiment results.

C. COMPARISON TO OTHER PRIVACY MEASUREMENT METHOD

Before publishing data to the general public for scientific study, Internet companies could apply some privacy protection methods to protect user privacy information from being leaked. The general goal of such methods is to reduce the risk of leaking specific privacy information. K-anonymity is such a classical method for privacy protection with the aim of

TABLE 3. Description of the Adult dataset used in the experiment.

	Attribute	Type	Attribute type
1	age	Numeric	Quasi identifier
2	workclass	Categorical	Quasi identifier
3	education	Categorical	Quasi identifier
4	native country	Categorical	Quasi identifier
5	marital status	Categorical	Quasi identifier
6	race	Categorical	Quasi identifier
7	gender	Categorical	Quasi identifier
8	occupation	Categorical	Explicit Sensitive attribute
9	wage	Categorical	Implicit Sensitive attribute

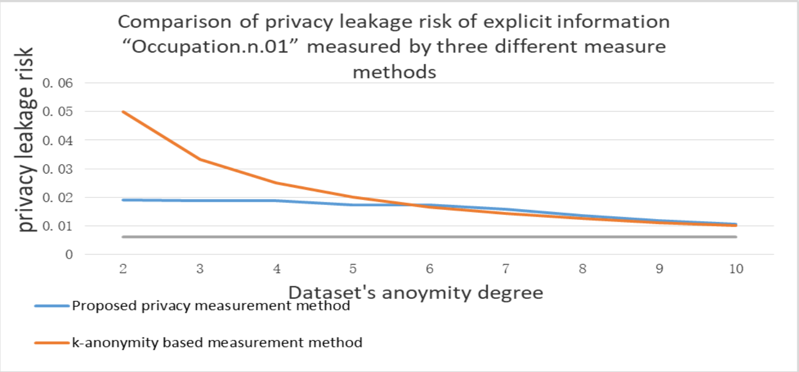

FIGURE 10. Privacy measurement of implicit information "Occupation.n.01".

reducing the risk of leaking sensitive attributes from linking attacks through anonymizing the quasi identifier attributes and the k value can be used as a measure of the privacy protection of certain specific sensitive attributes. The higher the value of k, the lower the risk of leaking the sensitive attributes. However, there exists the possibility in which some other leaked privacy attributes be used to infer the protected attributes.

The aim of this experiment is to compare the privacy measurement result of the privacy measurement method proposed in this paper to k-anonymity and to information entropy based privacy measurement such as that proposed by Clauß *et al.* [10] in terms of the degree of privacy protection. We included 9 attributes of the dataset as shown in Table 3 where the first 7 attributes were treated as the quasi identifiers, Occupation as the explicit sensitive attribute and Wage as the implicit sensitive attribute. We used the 400 TRUE sets of samples from the dataset that were to be released. In order to protect the sensitive attribute Occupation, we used the Mondrian algorithm [33] to anonymize the dataset by using different k values ranging from 2 to 10 and then measured the degree of privacy leakage of the explicit sensitive attribute Occupation and the implicit sensitive attribute Wage in the anonymized dataset, respectively, by applying the three methods involved in the experiment. Figure 10 shows the measurement results of the explicit sensitive attribute Occupation where the measurement results were converted to the same order of magnitude to make the comparison more intuitive. As can be seen, the results on the degree of leakage for each method have its own pattern. The key point here is to analyze the trend of change for each measurement method as the level of privacy protection varies. The results show that as the value of k increases, the risk of privacy leakage measured by the proposed privacy measurement method and by the k-anonymity based measure method goes down.

However, the situation changes when it comes to the implicit sensitive attribute Wage. Figure 11 shows the results for the three privacy measurement methods. The k-anonymity based method even could not identify the risk of privacy leakage of Wage while the proposed privacy measurement method could find the risk of privacy leakage of Wage as the risk value declines along with the increase in degree of anonymity. Lastly, the results for the information entropy based measurement method all stay the same. This is because each result depends only on the probability of possible values without concerning about what the information actually is. It is really not an appropriate way of measuring the risk of leaking both explicit and implicit sensitive attributes.

Experimental results show that the proposed semantic inference based privacy measurement method can be used to measure the leakage of implicit privacy information that most current privacy measurement methods cannot adequately do. Moreover, it can be used to find the source of privacy disclosure more precisely, providing the guidance for the development of effective privacy protection methods.

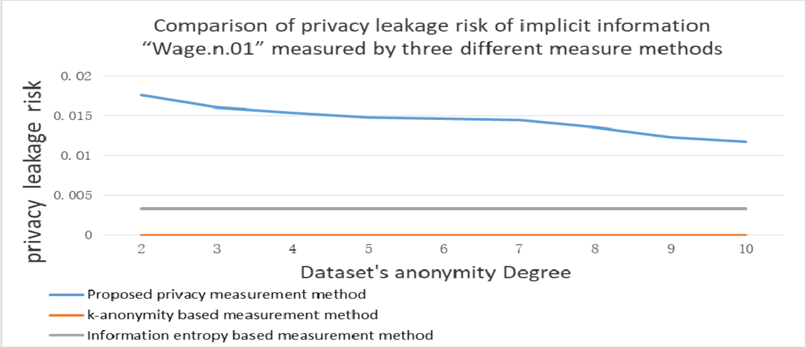

FIGURE 11. Privacy measurement of implicit information “Wage.n.01”.

TABLE 4. P3P categorization of privacy information.

Privacy category	Privacy words
Basic personal information	Age, race, party
Financial information	Income, debt, property
Occupation information	Occupation, position, salary
Address information	Address
Other information	Education, marriage, hobby, health

D. PRIVACY MEASUREMENT ON UNSTRUCTURED DATA

Along with the development of social media and other Internet applications, an increasing amount of unstructured data like blogs or tweets are posted which contains some personal information, creating the great risk of privacy leakage. In order to evaluate the effectiveness of measurement of the proposed privacy measurement method, we selected three influential people from different walks of life in our experiment as the privacy subjects: athlete Michael Jordan (M.J.), politician Barack Hussein Obama (B.H.O), the 44th president of the United States, and author Stephen Edwin King (S.E.K.). For each of the subjects, we used his name as the keyword to search for published articles using a commercial search engine and then randomly selected ten articles from thousands of search results after filtering out encyclopedia articles because they contain almost all the personal information. We then extracted personal information points from the ten articles and mapped them to the corresponding WordNet nodes using Word Sense Disambiguation method [34] to make the articles the background knowledge for the inference which can be viewed as the attacker's $\mathbf{K_{prio}}$. Then, the attacker can conduct the privacy attack through proper inference knowledge $\mathbf{K_{ob}}$. Finally, the attacker can get some privacy result $\mathbf{K_{post}}$ through inference, which can be used as $\mathbf{K_{prio}}$ for future attacks. The proposed privacy measurement method can simulate the privacy inference process.

In the experiment, we measured 14 pieces of privacy information condensed from the 17 privacy words defined in the P3P standard [35] in which these words are classified into five categories: basic personal information, financial information, occupation information, address information and other information, as listed in Table 4. We measured the risk of leaking these privacy words using the proposed the privacy measurement method by increasing the number of articles as the $\mathbf{K_{prio}}$ for the attacker. Figures 12-14 show the privacy measurement results of the three privacy subjects where the horizontal axes represent the number of articles applied and the vertical axes show the results of privacy leakage. When the value of a measurement result reaches 1, the corresponding privacy information is considered to be fully leaked. Otherwise, the value expresses the degree of privacy leakage.

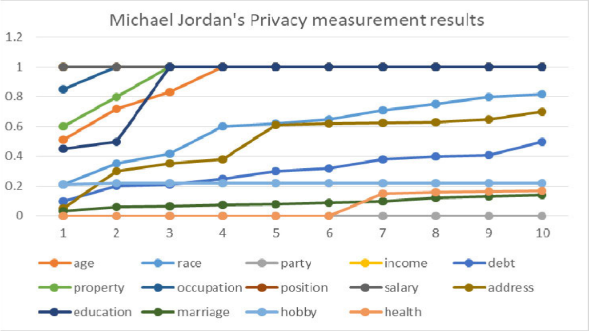

FIGURE 12. Privacy disclosure risk for Michael Jordan.

The measurement results in Figures 12-14 show that as the background knowledge accumulates as the result of applying more articles, the risk of leaking the privacy words increases. There are some similarities between the three subjects, e.g., marriage information is not easily leaked while age,

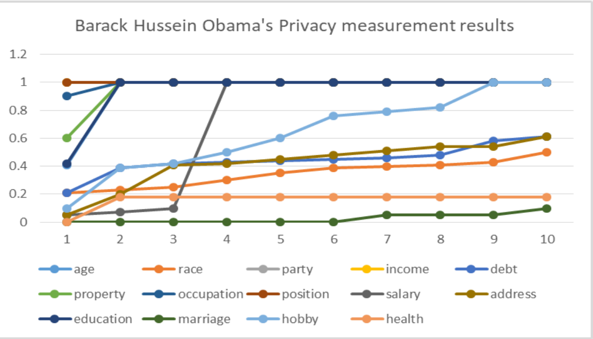

FIGURE 13. Privacy disclosure risk for Barack Hussein Obama.

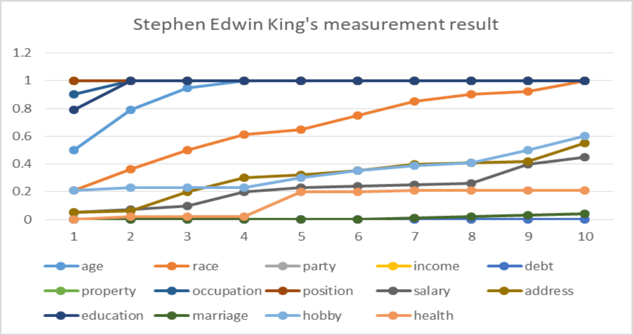

FIGURE 14. Privacy disclosure risk for Stephen Edwin King.

education and occupation information has a high risk of leakage. There are also differences found through applying the proposed privacy measurement method. For example, salary information for athlete Michael Jordan can be easily leaked while party information for politician Barack Hussein Obama can be easily leaked, but debt information for author Stephen Edwin King can hardly be leaked because the articles mostly talk about his book. So, the proposed privacy measurement method can also be applied to measure unstructured data effectively.

VI. CONCLUSION

There are currently not many effective methods for privacy measurement. Common methods that are based on probability statistics, information theory and set pair analysis can hardly meet the needs of privacy measurement at the era of the Internet and big data. In this paper, we proposed a semantic inference based privacy measurement method by incorporating an attribute weight calculation algorithm based on IC to express the inference preference during the process of privacy attack. Experiment was conducted to verify the effectiveness of the proposed privacy measurement method which includes a subjective strategy and an objective evaluation. For objective evaluation, the measurement results were also compared to those based on data mining and based on k-anonymity to demonstrate the advantages of the proposed privacy measurement method over the existing ones. In the future, to deal with the deficiency of the WordNet in terms of the lack of some knowledge, we plan to add some new nodes after exploring expert opinions based on the WordNet and delete some nodes that are not very relevant to privacy transmission so as to build a privacy-oriented ontology that is more suitable for privacy measurement. We will also apply the proposed privacy measurement method to other privacy protection technologies such as access control for better protection of implicit privacy information during access by legitimate users and during attack by malicious users.

REFERENCES

[1] L. Sweeney, "K-anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl. Based Syst.*, vol. 10, no. 05, pp. 557–570, Oct. 2002.

[2] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-Diversity: Privacy beyond k-anonymity," *ACM Trans. Knowl. Discov. Data.*, vol. 1, no. 1, pp. 1–52, Mar. 2007.

[3] N. Li, T. Li, and S. Venkatasubramanian, "T-closeness: Privacy beyond k-Anonymity and l-Diversity," in *Proc. IEEE 23rd Int. Conf. Data Eng.*, Apr. 2007, pp. 106–115.

[4] 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.

[5] J. Zhang, J. Xie, J. Yang, and B. Zhang, "A t-closeness privacy model based on sensitive attribute values semantics bucketization," *Jisuanji Yanjiu Fazhan*, vol. 51, no. 1, pp. 126–137, 2014.

[6] O. Gkountouna and M. Terrovitis, "Anonymizing collections of tree-structured data," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 8, pp. 2034–2048, Aug. 2015.

[7] Y. Yamaoka and K. Itoh, "K-presence-secrecy: Practical privacy model as extension of k-Anonymity," *IEICE Trans. Inf. Syst.*, vol. E100.D, no. 4, pp. 730–740, 2017.

[8] X.-Y. Li, C. Zhang, T. Jung, J. Qian, and L. Chen, "Graph-based privacy-preserving data publication," in *Proc. 35th Annu. IEEE Int. Conf. Comput. Commun.*, Apr. 2016, pp. 1–9.

[9] C. Díaz, S. Seys, J. Claessens, and B. Preneel, "Towards measuring anonymity," in *Proc. 2nd Int. Conf. Priv. Enha. Tech.* Berlin, Heidelberg: Springer, Apr. 2002, pp. 54–68.

[10] S. Clauß and S. Schiffner, "Structuring anonymity metrics," in *Proc. 2nd ACM Workshop Digit. Identity Manage.*, 2006, pp. 55–62.

[11] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady, "Preserving privacy in gps traces via uncertainty-aware path cloaking," in *Proc. 14th ACM Conf. Comput. Commun. Secur.*, 2007, pp. 161–171.

[12] Z. Ma, F. Kargl, and M. Weber, "A location privacy metric for V2X communication systems," in *Proc. IEEE Sarnoff Symp.*, Mar. 2009, pp. 1–6.

[13] Z. Ma, F. Kargl, and M. Weber, "Measuring location privacy in V2X communication systems with accumulated information," in *Proc. IEEE 6th Int. Conf. Mobile Adhoc Sensor Syst.*, Oct. 2009, pp. 322–331.

[14] R. Shokri, J. Freudiger, M. Jadliwala, and J.-P. Hubaux, "A distortion-based metric for location privacy," in *Proc. 8th ACM workshop Privacy Electron. Soc.*, 2009, pp. 21–30.

[15] X. Chen and J. Pang, "Measuring Query privacy in location-based services," in *Proc. 2nd ACM Conf. Data Appl. Secur. Privacy*, 2012, pp. 49–60.

[16] Y. Yang, J. Lutes, F. Li, B. Luo, and P. Liu, "Stalking online: On user privacy in social networks," in *Proc. 2nd ACM Conf. Data Appl. Secur. Privacy*, 2012, pp. 37–48.

[17] C. G. Peng, H. F. Ding, Y. J. Zhu, Y. L. Tian, and Z. Fu, "Information entropy models and privacy metrics methods for privacy protection," (in Chinese), *J. Softw.*, vol. 27, no. 8, pp. 1891–1903, 2016.

[18] K. Zhao, *Set Pair Analysis and Its Preliminary Application*. Zhejiang, China: Zhejiang Science Technology Press, 2000.

[19] Y. Yan, X. Hao, and W. Wang, "A set pair analysis method for privacy protection measurement," *J. Wuhan University, Eng. Ed.*, vol. 48, no. 6, pp. 883–890, 2015.

[20] C. Dwork, "Differential privacy," in *Proc. 33rd Int. Colloq. Automata Lang. Program.*, Venice, Italy, Jul. 2006, pp. 1–12.

[21] X. Wu, W. Dou, and Q. Ni, "Game theory based privacy preserving analysis in correlated data publication," in *Proc. Australas. Comput. Sci. Week Multiconference*, 2017, pp. 1–10.

[22] P. Cuff and L. Yu, "Differential privacy as a mutual information constraint," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2016, pp. 43–54.

[23] C. Fellbaum, *WordNet: An electronic lexical database, Language, Speech, and Communication*. Cambridge, MA, USA: MIT Press, 1998.

[24] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul./Oct. 1948.

[25] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proc. 14th Int. Joint Conf. Artificial Intell.*, Montréal QC, Canada, Aug. 1995, pp. 448–453.

[26] S. Z. Li, *Markov Random Field Modeling in Image Analysis*. Berlin, Germany: Springer-Verlag, 2009.

[27] N. Zhu, S. Wang, J. He, D. Teng, P. He, and Y. Zhang, "On the suitability of applying WordNet to privacy measurement," *Wirel. Pers. Commun.*, vol. 103, no. 1, pp. 359–378, Nov. 2018.

[28] B. C. Chen, R. Ramakrishnan, and K. Lefevre, "Privacy skyline: Privacy with multidimensional adversarial knowledge," in *Proc. 33rd Int. Conf. Very large Data Bases.*, Vienna, Austria, Sep. 2007, pp. 770–781.

[29] T. Li and N. Li, "Injector: Mining background knowledge for data anonymization," in *Proc. IEEE 24th Int. Conf. Data Eng.*, Cancun, MX, USA, Apr. 2008, pp. 446–455.

[30] L. Meng, J. Gu, and Z. Zhou, "A new model of information content based on concept's topology for measuring semantic similarity in WordNet," *Int. J. Grid Distrib. Comput.*, vol. 5, no. 3, pp. 81–94, Sep. 2012.

[31] D. Dua and C. Graff, "UCI machine learning repository," School Inf. Comput. Sci., Univ. California, Irvine, CA, USA, 2019. [Online]. Available: http://archive.ics.uci.edu/ml

[32] L. Rutkowski, M. Jaworski, L. Pietruczuk, and P. Duda, "The CART decision tree for mining data streams," *Inf. Sci.*, vol. 266, pp. 1–15, May 2014.

[33] K. LeFevre and D. J. DeWitt, "Ramakrishnan R. Mondrian multidimensional k-anonymity," in *Proc. 22nd Int. Conf. Data Eng.*, Atlanta, GA, USA, Apr. 2006, p. 25.

[34] S. Vij, A. Jain, D. Tayal, and O. Castillo, "Fuzzy logic for inculcating significance of semantic relations in word sense disambiguation using a WordNet graph," *Int. J. Fuzzy Syst.*, vol. 20, no. 2, pp. 444–459, 2018.

[35] *The platform for privacy preferences 1.0 (P3P1. 0)*, World Wide Web Consortium, Cambridge, MA, USA, 2002.

BAOCUN CHEN received the bachelor's degree in computer science from the North China University of Technology. He is currently pursuing the master's degree in software engineering with the Beijing University of Technology, China. His research interests include privacy protection and information security.

NAFEI ZHU received the B.S. and M.S. degrees from Central South University, China, in 2003 and 2006, respectively, and the Ph.D. degree in computer science and technology from the Beijing University of Technology, Beijing, China, in 2012. She was a Postdoctoral Research Fellow with the State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing, from 2015 to 2017. She is currently an Associate Professor with the Faculty of Information Technology, Beijing University of Technology. She has published over 20 research papers in scholarly journals and international conferences. Her research interests include information security and privacy, wireless communications, and network measurement.

JINGSHA HE (Member, IEEE) received the bachelor's degree in computer science from Xi'an Jiaotong University, China, and the master's and Ph.D. degrees in computer engineering from the University of Maryland, College Park, MD, USA. He worked for several multinational companies in USA, including IBM Corp., MCI Communications Corp., and Fujitsu Laboratories. He is currently a Professor with the Faculty of Information Technology, Beijing University of Technology (BJUT), Beijing. He has published more than ten articles. He holds 12 U.S. patents. Since August 2003, he has been published over 300 papers in scholarly journals and international conferences. He also holds over 84 patents and 57 software copyrights in China and authored nine books. He was a principal investigator of more than 40 research and development projects. His research interests include information security, wireless networks, and digital forensics.

PENG HE received the master's degree in computer software from Xi'an Jiaotong University, in 1989. He was an Education Information Expert with the Ministry of Education and a Standing Director with the Hubei Education Information Technology Research Association. He is currently a Professor with the College of Computer and Information, China Three Gorges University. His research interests include network time synchronization and deep learning.

SHUTING JIN received the bachelor's degree in software engineering from the Taiyuan University of Technology, China. She is currently pursuing the master's degree in software engineering with the Beijing University of Technology, China. Her research interests include privacy protection and access control.

SHIJIA PAN received the bachelor's degree from the Wuhan University of Technology. She is currently pursuing the master's degree with the School of Software Engineering, Beijing University of Technology (BJUT), Beijing, China. Her research interests include information dissemination and privacy security.