Structural Mutual Information and Its Application

Youliang Tian[1*], Zhiying Zhang[1], Jinbo Xiong[2], Jianfeng Ma[3]

[1]*State Key Laboratory of Public Big Date, College of Computer Science and Technology, Guizhou University, Guiyang, China*

[2]*Fujian Provincial Key Laboratory of Network Security and Cryptology, College of Mathematics and Informatics, Fujian Normal University, Fuzhou, China*

[3]*School of Cyber Engineering, Xidian University, Xi'an, China*

***Abstract*-Shannon mutual information is an effective method to analyze the information interaction in a point-to-point communication system. However, it cannot solve the problem of channel capacity in graph structure communication system. This problem make it impossible to use traditional mutual information (TMI) to detect the real information and to measure the information embedded in the graph structure. Therefore, measuring the interaction of graph structure and the degree of privacy leakage has become an emerging and challenging issue to be considered. To solve this issue, we propose a novel structural mutual information (SMI) theory based on structure entropy model and the Shannon mutual information theorem, following by the algorithms for solving SMI. The SMI is used to detect the real network structure and measure the degree of private data leakage in the graph structure. Our work expands the channel capacity of Shannon's second theorem in graph structure, discusses the correlation properties between SMI and TMI, and concludes that SMI satisfies some basic properties, including symmetry, non-negativity, and**

Foundation Items: National Natural Science Foundation of China under Grant Nos. 61662009 and 61772008; Science and Technology Major Support Program of Guizhou Province under Grant No.20183001; Ministry of Education-China Mobile Research Fund Project under Grant No.MCM20170401; Key Program of the National Natural Science Union Foundation of China under Grant No.U1836205; Science and Technology Program of Guizhou Province under Grant No.[2019]1098; Project of High-level Innovative Talents of Guizhou Province under Grant No. [2020]6008; Innovative talent team of Guizhou ordinary colleges and Universitie(Guizhou-Education-Talent-Team[2013]09)

so on. Finally, theoretical analysis and example demonstration show that the SMI theory is more effective than the traditional privacy measurement methods to measure the information amount embedded in the graph structure and the overall degree of privacy leakage. It provides feasible theoretical support for the privacy protection technology in the graph structure.

***Index Terms*-Structural mutual information (SMI), mutual information, structure entropy, privacy measurement, privacy leakage.**

I. INTRODUCTION

The types and amounts of data generated in various fields have exploded exponentially due to the rapid and intuitive development of social informatization and network. Digitization has become the foundational force for building a modern society and is driving us to an era of profound change. The continuous emergence and development of new technologies make people realize the value of data and the importance of open sharing. However, in the big data environment, we left the data footprints of our visited on various Internet websites and app applications, which makes our privacy data more accessible. Therefore, the protection of user private data has naturally become the focus of peoples attention.

In the 21st century, data of new types, such as biological data, web data, topographical maps and medical data etc. have appeared. These data form a large scale and strong agglomeration dynamic complex network. In this complex network environment, analysing the new data and discovering the orders and knowledge of them are new challenges. The existing work mainly focuses on the research of complex networks, using the traditional mutual information (TMI) to measure the correlation between nodes in the static graph structure. How to measure the information amount embedded in dynamic complex networks and describe the interaction among graph structures are a big challenge. For this new mission, Shannon's definition of information is apparently insufficient, for which a new metric of structural information is urgently called for.

A. Related Work

Complex networks were assumed to be randomly evolved in the early history of network study. Erdös and Rényi (ER) [7, 8] proposed a classic model, referred to as the ER model, which was used to explore many properties of random graphs. In 1998, Watta and Strogatz [9] proposed a simple model in which random edges are added to a grid graph. Newman and Girvan [10] defined the notion of modularity to measure the quality of community structure in a network. Regarding the complexity of the structural network. Kamisinski *et al.* [11] presented and compared 17 structural complexity indices, and determined the advantages and disadvantages of each metric according to different measurement

methods. Nevertheless, these metrics methods are suitable to examine the complexity of simple, small, usually not very dense graphs. Glantz *et al.* [12] proposed a new edge rating by defining weights for the edges of a network that express the edges' importance for connectivity via shortest paths, computed a minimum weight spanning tree with respect to these weights, and rating the network edges based on the conductance values of the tree's basic cuts. Then, the authors stressed that edge ratings are not only of interest for graph partitioning, but also for graph clustering and network analysis. However, for social networks with huge data. Wang *et al.* [13] aim to measure the transport difficulty of data dissemination in online social networks $(OSNs)$. The authors defined the fundamental limit of transport load as a new metric, called transport complexity. Furthermore, they derived the tight bounds on transport complexity of Social-InterestCast. Sansavini and Parigi [14] studied multimode Continuous Variables entangled states, named cluster states, where the entanglement structure is arranged in typical real-world complex networks shapes. The authors analyzed in the quantum regime different complex shapes corresponding to different models of real-world networks. The optimal graph state is obtained by analysis method optimization. But it has to be stressed that, in the cases where no solution is found, they cannot conclude with certainty that the solution does not exist, as the Derandomized Evolution Strategy (DES) algorithm can be stuck in a local minimum in the parameters space. With the arrival of the bigdata and $5G$, large public networks of various types have come to existence. Nevertheless, the societal and research benefits of these networks have also given rise to potentially significant privacy issues. Tanima Chatterjee *et al.* [15] formalized three such privacy measures for large networks and provided non-trivial theoretical computational complexity results for computing these measures. However, authors only provided a logarithmic approximation algorithm for $Adim_{k=1}$. When $k > 1$, whether the non-trivial approximation algorithm holds becoming a question. Therefore, the actual structure of complex networks cannot be effectively measured. In the communication environment of complex networks, the existing methods cannot effectively measure the real structure of dynamic complex networks, leading to private data leakage. Thus, how to effectively measure the real structure in complex networks has become an important issue that we consider at present.

Shannon's information theory [6] has done a lot of meaningful works on privacy protection technology. Hsiao *et al.* [16] studied conditional computational entropy: the amount of randomness a distribution appears to have to a computationally bounded observer who is given some correlated information. By considering conditional versions of Håstad, Impagliazzo, Levin, Luby $(HILL)$ entropy (so named after the author [35]) and Yao entropy, it proposed a new, natural notion of unpredictable entropy. As a method of privacy measurement, mutual information can be applied to privacy protection algorithms to increase privacy protection effectiveness. Whitnall and Oswald [17] assessed the effectiveness of mutual information based differential power analysis within a generic and comprehensive evaluation framework.

They presented several notions/characterisations of attack success with direct implications for the amount of data required and observed an interesting feature unique to the mutual information-based distinguisher. It could potentially enhance the effectiveness of such attacks over other methods in certain noisy scenarios. But it is difficult to find the estimator that most effectively transforms theoretical advantage into an actual advantage. Taghia and Martin [18] investigated the speech intelligibility prediction from the viewpoint of information theory and introduced novel objective intelligibility measures based on the estimated mutual information between the temporal envelopes of clean speech and processed speech in the subband domain. Chitambar *et al.* [19] investigated the problem of extracting secret key from an eavesdropped source p_{XYZ} at a rate given by the conditional mutual information under three different scenarios. For each of the above scenarios, strong necessary conditions are derived on the structure of distributions attaining a secret key rate of $I(X:Y|Z)$. Perotti *et al.* [20] introduced hierarchical mutual information, comparison of hierarchical division and hierarchical community structure. In order to quantitatively describe the community and module structure of complex networks. Li *et al.* [21] proposed a novel entropy and mutual information-based centrality approach (EMI), which attempts to capture a far wider range and a greater abundance of information for assessing how vital a node was. Viegas *et al.* [22] developed a complex framework inspired by the allometric scaling laws of living biological systems in order to evaluate the structural features of networks. It is realized by aligning entropy and mutual information with degree correlation. However, as pointed out by Brooks in [5], it had been a huge and longstanding challenge to define the information that is embedded in a physical system. Therefore, the TMI measurement can not support data analysis in communication networks, and it is difficult to measure the information amount embedded in graphs and the degree of correlation among graph structures.

Therefore, how to effectively quantify the structural information in dynamic complex networks is a new question. The answer to this question depends on a clear definition of the structure entropy. That is, structure entropy is used to measure the information embedded in a communication network or graph structure. With the emergence of new types of big data, these data have complex relationships. The knowledge and laws of data are embedded in the large-scale noisy structure. The knowledge extraction and data structure analysis of these big data require us to measure the information embedded in the complex network. This information determines and decodes the real structure of the complex network. Therefore, the quantification of structural information is becoming increasingly important. Li and Pan [1] defined the concept of structure entropy of graph G to measure the information embedded in the graph and determined and decoded the basic structure of graph G. And they proposed the first metric for structural information. Given a graph G, they defined the K-dimensional structural information of G (or structure entropy of G), denoted by $H^K(G)$, to be the minimum overall number of bits required to determine the K-dimensional

code of the node that is accessible from random walk in G. The K-dimensional structural information provides the principle for completely detecting the natural or true structure. Li *et al.* [2,3,4] firstly proposed a community finding algorithm by measure of structure entropy of networks. They found that our community finding algorithm exactly identifies almost all natural communities of networks generated by natural selection. Subsequently they proposed the notion of resistance of a graph as an accompanying notion of the structure entropy to measure the force of the graph to resist cascading failure of strategic virus attacks. They show that for any connected network G, the resistance of G is $R(G) = H^1(G) - H^2(G)$, where $H^1(G)$ and $H^2(G)$ are the one- and two- dimensional structure entropy of G, respectively. Further proposed a fast and normalization-free method to decode the domains of chromosomes (deDoc) that utilizes structural information theory. Wan *et al.* [24] considered the graph entropy measures based on the number of independent sets and matchings. Author established some upper and lower bounds as well as some information inequalities for these information-theoretic quantities. Those results reveal the two entropies possess some new features. Guo *et al.* [25] proposed a framework to convert the protein intrinsic disorder content to structural entropy using Shannon's information theory. Liu *et al.* [26] proposed a community-based structural entropy to express the information amount revealed by a community structure. By exploiting the community affiliations of user accounts, an attacker may infer sensitive user attributes. This raised the problem of community structure deception (CSD) and focuses on the privacy risks of disclosing the community structure in an online social network. In extensive application scenarios, conducting community detection over blockchain networks has potential effects on both discovering hidden information and enhancing communicating efficiency. However, the decentralised nature poses a restriction on community detection over blockchain networks. In coping with this restriction, Chen and Liu. [27] proposed a distributed community detection method based on the Propose-Select-Adjust (PSA) framework that runs in an asynchronous way. They extend the PSA framework used the concept of structural entropy and aimed to detect a community structure with low entropy.

B. Our Contribution

To sum up, most of the existing methods for studying the relevant characteristics of point-to-point communication focus on using information entropy frameworks, without fully considering the multidimensional data or complex structure in the graph. As a result, if we apply these approaches to complex networks, it is difficult to measure users' private data security and the overall degree of privacy leakage. Here, this paper proposes the structural mutual information (SMI) based on structural information theory. We can detect the true structure of complex networks based on SMI, and further effectively measure information amount embedded in the graph structure. Finally, it measures the degree of privacy leakage

of graph structure and describes the correlation among graph structures. Specifically, the contributions of this paper are three aspects as follows:

- Aiming at the problem that the TMI cannot effectively describe the interactions among graph structures, we propose SMI with structure entropy. The K-dimensional structural information provides the principle for completely detecting the natural or true structure. So, we measure the information amount embedded in the graph by using the K-dimensional structural entropy minimization. The SMI is further constructed based on the TMI solution formula. The SMI can be used as a privacy measurement method for graph structure and reflect the correlation among graph structures.
- We propose a series of algorithms for solving SMI of connected graphs and disconnected graphs. The connected graph is randomly divided into subgraphs G_1 and G_2. By combining the K-dimensional structure entropy $H^K(G)$ of the connected graph with Shannon mutual information, the SMI: $I(G_1;G_2)$ of the connected graph is calculated. The SMI of connected graphs mainly solves the interaction and correlation degree of two connected graph structures. The connected components of the disconnected graph are marked as subgraphs G_1 and G_2. By combining the K-dimensional structure entropy $H^K(G^{'})$ of the disconnected graph with Shannon mutual information, the SMI: $I(G_1^{'};G_2^{'})$ of the disconnected graph is calculated. The SMI of disconnected graphs mainly solves the interaction and correlation degree of two unconnected graph structures. The non-negativity, symmetry and other properties of SMI are further proved.
- We apply the proposed SMI to the privacy measurement of the graph structure. That is, the SMI is used to describe the leakage risk of privacy information in the graph structure. By calculating the SMI before and after the known graph G_2, the amount of uncertainty reduction about graph G_1 to reflect the privacy leakage risk. The greater the SMI, the greater the degree of privacy leakage of the graph structure. Compared with traditional measurement methods, we solve the shortcomings of mutual information and other measurement methods, and illustrate the effectiveness of SMI.

The rest of this paper is organized as follows. Section II introduces the concepts of coding tree and structure entropy proposed by Li and Pan [1]. Section III presents the concept of mutual information and communication channel. Based on the structural information of connected graphs and disconnected graphs, we propose the SMI of graph structures and prove the related properties of SMI. Section IV proposes a series of algorithms for solving connected graphs and disconnected graphs. Section V presents four privacy information measurement methods and compares the differences between the SMI and the four measurement methods. Finally, we conclude this paper and discuss some unsolved problems in Section VI.

II. STRUCTURAL INFORMATION

Shannon information theory is mainly based on the hypothesis of discrete, memoryless and lossy transmission, which realizes point-to-point transmission and solves the problem of person-to-person reliable communication. With the advent of an intelligent society, communication scenarios are becoming more diverse. Nowadays, there are many communication modes, such as single point to multi-point, multi-point to multi-point, etc.. Therefore, the quantification of structural information has always been one of the challenges in computer science, that is, there is currently no practical theoretical support to measure the information embedded in the structure. The same question Shannon himself mentioned in his 1953 paper. Shannons information theory can only use information entropy to analyze a number in each graph G, but we can't analyze the properties of the entire graph G in one point. Therefore, Shannon information is not very helpful for us to analyze the charts or structured data. When we need to analyze large-scale network data and unstructured data, the situation is even less optimistic. In order to better understand and solve the huge challenges left by Shannon, we need a new measure of structural information that supports the analysis of graphics, networks, structured data, and even unstructured data. Li and Pan [1] put forward the theory of structural information in order to measure uncertainty. The K-dimensional structural information provides the principle for completely detecting the natural or true structure. That is to decode the knowledge and laws embedded in the dynamic complex network from the large-scale noisy structure, and for analyzing communication systems, solving the Shannons problem and opening up new directions. Authors have observed that the K-dimensional structure entropy minimization is the principle for detecting the natural or true structures in real-world networks and is also the first metric of dynamical complexity of networks.

According to the structural information theory [1], the K-dimensional structural information of the graph G is exactly the measure of the uncertainty of the K-dimensional structure of the graph G. All the definitions of structural information were put forward by Li and Pan [1]. Before introducing the structure entropy, we understand the coding tree [1, 2] of the graph according to Huffman coding [29].

A. Coding Tree of A Graph

Definition 1 (Coding tree of a graph): Let $G = (V, E)$ be a graph. Coding tree T of graph G makes each tree node $\alpha \in T$, there is a subset T_α of the vertices V, and such that the following properties hold:

- For the root node λ, define a set $T_\lambda = V$.
- For each node $\alpha \in T$, $\alpha^\wedge\langle j\rangle$ represents the child node of α, j is the natural number from 1 to N increasing from left to right. Each internal node has at least two direct successors, when $i < j$, $\alpha^\wedge\langle i\rangle$ is on the left-hand side of $\alpha^\wedge\langle j\rangle$.

- For every $\alpha \in T$, there is a subset $T_\alpha \subset V$ that is associated with α. For α and β, we use $\alpha \subset \beta$ to denote that α is an initial segment of β. For every node $\alpha \neq \lambda$, we use α^- to denote the longest initial segment of α, or the longest β such that $\beta \subset \alpha$.
- For every i, $\{T_\alpha | h(\alpha) = i\}$ is a partition of V, where $h(\alpha)$ is the height of α (note that the height of the root node λ is 0, and for every node $\alpha \neq \lambda$, $h(\alpha) = h(\alpha^-) + 1$).
- For every α, T_α is the union of T_β for all β's such that β^-=α; thus, $T_\alpha = U_{\beta^-=\alpha} T_\beta$.
- For every leaf node α of T, T_α is a singleton; thus, T_α contains a single node of V.
- For each node $\alpha \in T$, if $T_\alpha = X$ is a set of vertices X, α is the code word for X, denoted as $c(X) = \alpha$, and termed X the marker of α, denoted as $M(\alpha) = X$.

B. Structural Information of Connected Graphs

Definition 2 (One-Dimensional Structural Information of Connected Graphs): Let $G = (V, E)$ be a connected and undirected graph with n vertices and m edges. For each node $i \in 1, 2, ..., n$, let d_i be the degree of i in G, and let $p_i = d_i/2m$. Then the stationary distribution of graph G is described by probability vector $P = (p_1, p_2, ..., p_n)$. We define the one-dimensional structural information of graph G as follows:

$$\begin{aligned} H^1(G) = H(P) &= H(\frac{d_1}{2m}, ..., \frac{d_n}{2m}) \\ &= -\sum_{i=1}^{n} \frac{d_i}{2m} \cdot \log_2 \frac{d_i}{2m}. \end{aligned} \tag{1}$$

Definition 3 (Two-Dimensional Structural Information of Connected Graphs): Given a connected and undirected graph $G = (V, E)$, suppose that $P = \{X_1, X_2, ..., X_L\}$ is a partition of V, in which each X_j is called a module or a community. We define the structural information of G by P as follows:

$$\begin{aligned} H^P(G) &= \sum_{j=1}^{L} \frac{V_j}{2m} \cdot H(\frac{d_1^{(j)}}{V_j}, ..., \frac{d_{n_j}^{(j)}}{V_j}) - \sum_{j=1}^{L} \frac{g_j}{2m} \log_2 \frac{V_j}{2m} \\ &= -\sum_{j=1}^{L} \frac{V_j}{2m} \sum_{i=1}^{n_j} \frac{d_i^{(j)}}{V_j} \log \frac{d_i^{(j)}}{V_j} - \sum_{j=1}^{L} \frac{g_j}{2m} \log_2 \frac{V_j}{2m}. \end{aligned} \tag{2}$$

Where L is the number of modules in partition P, n_j represents the number of nodes in module X_j, $d_i^{(j)}$ is the degree of the $i\text{-}th$ node of X_j, V_j represents the volume of module X_j, which is the sum of degrees of nodes in X_j, and g_j represents the number of edges with exactly one endpoint in module X_j. Therefore, two-dimensional structural information of graph G, also known as module entropy, can be defined as follows:

$$H^2(G) = \min_P \{H^p(G)\}. \tag{3}$$

Where P is all the partitions that can be formed in the graph G.

Definition 4 (K-Dimensional Structural Information of Connected Graphs): Let $G = (V, E)$, suppose T be the partition tree of graph G. We define the K-dimensional structural information of graph G by partition tree T as follows:

1) For every $\alpha \in T$, if $\alpha \neq \lambda$, λ is the root node of the tree, then define the structural information of vertex α as follows:

$$H^T(G;\alpha) = -\frac{g_\alpha}{2m}\log_2\frac{V_\alpha}{V_{\alpha^-}}. \tag{4}$$

Where g_α represents the number of edges from nodes in T_α to nodes outside T_α, V_α is the volume of set T_α, namely, the sum of the degrees of all the nodes in T_α.

2) We define the structural information of G by the partitioning tree T as follows:

$$H^T(G) = \sum_{\alpha\in T,\alpha\neq\lambda} H^T(G;\alpha). \tag{5}$$

3) The definition of K-dimensional structural information of G is as follows:

$$H^K(G) = \min_T\{H^T(G)\}. \tag{6}$$

Where the value range of T is all possible coding trees of graph G, and the maximum height is K.

C. Structural Information of Disconnected Graphs

Definition 5 (One-Dimensional Structural Information of Disconnected Graphs): Given a disconnected and undirected graph $G^{'} = (V, E)$,

1) If $E = \emptyset$, then define $H^1(G^{'}) = 0$.

2) Otherwise, suppose that $G_1^{'}, G_2^{'}, ..., G_L^{'}$ are the induced subgraphs of all the connected components of $G^{'}$. Then we define the one-dimensional structure entropy of $G^{'}$ is the weighted average of the one-dimensional structure entropies of $G_j^{'}$ for all j's. That is,

$$H^1(G^{'}) = \frac{1}{Vol(G^{'})}\cdot\sum_{j=1}^{L} Vol(G_j^{'})\cdot H^1(G_j^{'}). \tag{7}$$

Where $Vol(G^{'})$ is the volume of $G^{'}$, $Vol(G_j^{'})$ is the volume of $G_j^{'}$.

In Definition 5, we have that: 1) is reasonable, the reason is that there is no vertex random walk in $G^{'}$; and 2) is reasonable since it simply follows the additivity of the Shannon entropy function.

For a non-weighted graph, we regard the weight of an edge as 1.

Definition 6 (Two-Dimensional Structural Information of Disconnected Graphs): Given a disconnected and undirected graph $G^{'}$, suppose that $G_1^{'}, G_2^{'}, ..., G_L^{'}$ are the induced subgraphs of all the connected

components of $G^{'}$. Then we define the two-dimensional structural information of $G^{'}$ to be the weighted average of the two-dimensional structure entropies of all the subgraphs $G_j^{''}s$. That is,

$$H^2(G^{'}) = \frac{1}{Vol(G^{'})} \cdot \sum_{j=1}^{L} Vol(G_j^{'}) \cdot H^2(G_j^{'}). \tag{8}$$

Where $Vol(G^{'})$, for each j, $Vol(G_j^{'})$ is the volume of $G_j^{'}$.

To make sure that Definition 6 is well-defined, we notice that it is possible that there is a j such that $G_j^{'}$ contains a single isolated node. In this case, the two-dimensional structural information of $G_j^{'}$ is 0. So, for any graph $G^{'}$, if there is no edge among the nodes of $G^{'}$, then $\xi(G^{'}) = 0$.

Definition 7 (K-Dimensional Structural Information of Disconnected Graphs): For a disconnected and undirected graph $G^{'}$, the K-dimensional structural information of $G^{'}$ is defined similarly to Definition 6 as follows:

$$H^K(G^{'}) = \frac{1}{Vol(G^{'})} \cdot \sum_{j=1}^{L} Vol(G_j^{'}) \cdot H^K(G_j^{'}). \tag{9}$$

Where $Vol(G^{'})$, for each j, $Vol(G_j^{'})$ is the volume $G_j^{'}$, and $G_1^{'}, G_2^{'}, ..., G_L^{'}$ are all the connected components of $G^{'}$.

In particular, the K-dimensional structural information of a graph G implies that minimisation of non-determinism is the principle of the self-organisation of a K-dimensional structure of a graph for $K > 1$. For $K = 1$, the K-dimensional structural information of a graph G is the positioning entropy of G.

III. Structural Mutual Information

A. Mutual Information

Before defining SMI, let us review the definition of mutual information in Shannon's information theory[6]. Mutual information is the reduction of the uncertainty of the original random variable given the knowledge of another random variable. Specifically, let X and Y as two random variables, then the mutual information is defined by:

$$I(X;Y) = \sum_{x,y} p(x,y) log \frac{p(x,y)}{p(x)p(y)}. \tag{10}$$

Mutual information $I(X;Y)$ is also a measure of the degree of independence between two random variables. It is symmetric about X and Y and non-negative. If and only if X and Y are independent of each other, the mutual information is 0. Then, the relationship between mutual information and entropy can be defined as follows:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ I(X;Y) &= H(X) + H(Y) - H(X,Y). \end{aligned} \tag{11}$$

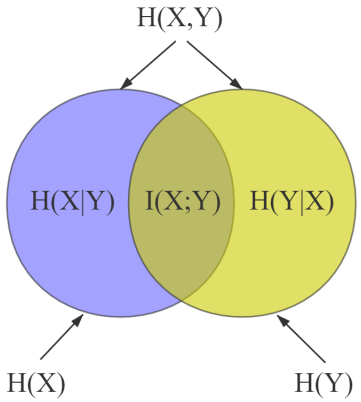


Fig. 1. The relationship between entropy and mutual information.

The relationship between $H(X)$, $H(Y)$, $H(X,Y)$, $H(X|Y)$, $H(Y|X)$ and $I(X,Y)$ can be represented by the Venn diagram, as shown in Fig. 1 [30].

It can be noted that mutual information $I(X;Y)$ corresponds to the intersection of information of X and information of Y. Therefore, mutual information has the following properties [30]:

a) Symmetry: The amount of information about X extracted from Y is the same as the amount of information about Y extracted from X, i.e., mutual information $I(X;Y) = I(Y;X)$.

b) Independence: Where X and Y are two information quantities respectively, and mutual information represents the measure of the correlation degree of two information quantities. Generally speaking, after knowing the amount of information Y, the amount of information X will be more certain, indicating that there is a certain correlation between the amount of information X and the amount of information Y. In contrast, if X and Y are irrelevant, then knowing Y cannot have any influence on the certainty of X. When events X and Y are independent of each other, i.e., $I(X;Y) = 0$.

c) Non-negativity: Extracting information from one event about another, in the worst case 0. It will not increase the uncertainty of another event by knowing an event. Through Jensen inequality and KL divergence [28], we can know that one random variable is always helpful to know another random variable, and prove that mutual information $I(X;Y) \geq 0$.

d) Extremum property: The amount of information extracted from one event about another event is at most the entropy of the other event. It will not exceed the amount of information contained in the other event itself. So mutual information can not provide more information than two random variables, i.e., mutual information $I(X;Y) \leq min\{H(X), H(Y)\}$ satisfies extremum property.

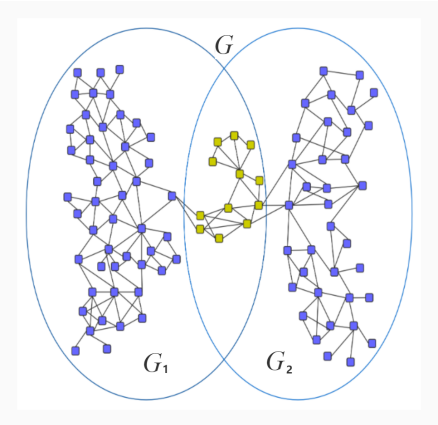


Fig. 2. Connected graph.

B. Structural Mutual Information

In the following, we propose the theorem of SMI in two cases of the connected graphs and disconnected graphs, and analyze and prove the relevant properties. Firstly, we randomly construct a connected graph of network model with n vertices and m edges, and divide the vertices in the connected graph into two subgraphs by random segmentation, as shown in Fig. 2.

It can be seen from Fig. 2 that we divided graph G into subgraphs G_1 and G_2. The mutual information between the two subgraphs is the overlap between subgraphs G_1 and G_2. Therefore, we define the SMI of connected graphs as follows.

Theorem 1: SMI of connected graphs. Given a connected and undirected graph $G = (V, E)$, with n vertices and m edges. We define the SMI of the connected graphs as follows:

$$I(G_1; G_2) = H^K(G_1) + H^K(G_2) - H^K(G). \tag{12}$$

$H^K(G_1)$ and $H^K(G_2)$ represent the K-dimensional structural information of subgraphs G_1 and G_2, $H^K(G)$ is the K-dimensional structural information of the whole connected graph G.

Symmetry of connected graphs: SMI of connected graphs $I(G_1; G_2)$ has symmetry.

Proof 1:

$$\begin{aligned} I(G_1; G_2) &= H^K(G_1) + H^K(G_2) - H^K(G) \\ &= H^K(G_2) + H^K(G_1) - H^K(G) \\ &= I(G_2; G_1). \end{aligned} \tag{13}$$

The amount of information about the graph G_1 extracted from the graph G_2 is the same as the amount of information about the graph G_2 extracted in the graph G_1. $I(G_1; G_2)$ and $I(G_2; G_1)$ are just different standpoints for the observer.

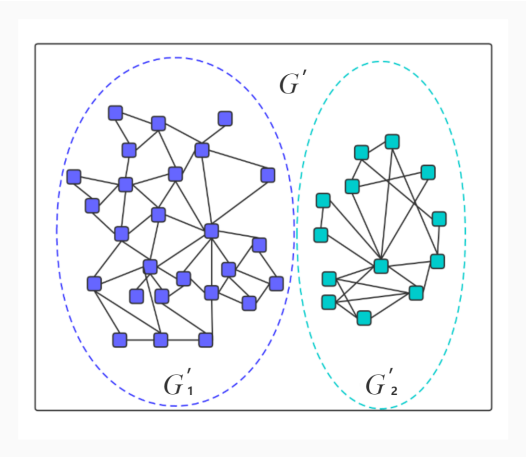


Fig. 3. Disconnected graph.

Non-negativity of connected graphs: SMI of connected graphs $I(G_1;G_2) \geq 0, i.e.$, satisfies non-negativity.

Proof 2: Due to the non-negativity of mutual information, we do know the condition reduces entropy, that is, $H(X) \geq H(X|Y)$. When we have knowledge and information about another random variable, the uncertainty of the random variable we are concerned about must have a non-negative gain. That is, if we know Y, then the remaining uncertainty of X will not increase, at least remain unchanged. Similarly, when we know the amount of information of connected graph G_2, the uncertainty of connected graph G_1 will have a non-negative gain. Therefore, the SMI of connected graphs also satisfies the condition to reduce entropy:

$$\begin{aligned} H^K(G_1) &> H^K(G_1|G_2) \\ I(G_1;G_2) &= H^K(G_1) + H^K(G_2) - H^K(G) \\ &= H^K(G_1) - H^K(G_1|G_2) \\ &\geq 0. \end{aligned} \tag{14}$$

The SMI of connected graphs satisfies non-negativity. The greater the conditional structural entropy of the privacy, the smaller the SMI of the connected graphs, and they have consistency.

We randomly construct a disconnected graph of network model with n vertices and m edges. Disconnected graph, that is, there are two or more connected components. Here, we define a disconnected graph $G^{'}$ to have two connected components. So, there are two subgraphs $G_1^{'}$ and $G_2^{'}$, as shown in Fig. 3.

We can see from Fig. 3 that the subgraphs $G_1^{'}$ and $G_2^{'}$ are the two connected components of the disconnected graph $G^{'}$. Therefore, we define the SMI of disconnected graphs as follows:

Theorem 2: SMI of disconnected graphs. Given an undirected and disconnected graph $G^{'} = (V, E)$, with n vertices and m edges. We define the SMI of the disconnected graphs as follows:

$$I(G_1^{'}; G_2^{'}) = H^K(G_1^{'}) + H^K(G_2^{'}) - H^K(G^{'}). \tag{15}$$

$H^K(G_1^{'})$ and $H^K(G_2^{'})$ represent the K-dimensional structural information of subgraphs $G_1^{'}$ and $G_2^{'}$, $H^K(G^{'})$ is the K-dimensional structural information of the whole disconnected graph $G^{'}$.

Symmetry of disconnected graphs: SMI of disconnected graphs $I(G_1^{'}; G_2^{'})$ has symmetry.

Proof 3:

$$\begin{aligned} I(G_1^{'}; G_2^{'}) &= H^K(G_1^{'}) + H^K(G_2^{'}) - H^K(G^{'}) \\ &= H^K(G_2^{'}) + H^K(G_1^{'}) - H^K(G^{'}) \\ &= I(G_2^{'}; G_1^{'}). \end{aligned} \tag{16}$$

It can be seen that the SMI of disconnected graphs $I(G_1^{'}; G_2^{'})$ has symmetry.

Non-negativity of disconnected graphs: It can be known from the non-negative property of connected graph, The SMI of disconnected graphs $I(G_1^{'}; G_2^{'}) \geq 0, i.e.,$ satisfies non-negativity.

Proof 4:

$$\begin{aligned} I(G_1^{'}; G_2^{'}) &= H^K(G_1^{'}) + H^K(G_2^{'}) - H^K(G^{'}) \\ &= H^K(G_1^{'}) + H^K(G_2^{'}) - \frac{1}{Vol(G^{'})} \cdot \sum_{j=1}^{L} Vol(G_j^{'}) \cdot H^K(G_j^{'}) \\ &= H^K(G_1^{'}) + H^K(G_2^{'}) \\ &\quad - \frac{Vol(G_1^{'}) \cdot H^K(G_1^{'}) + Vol(G_2^{'}) \cdot H^K(G_2^{'})}{Vol(G_1^{'}) + Vol(G_2^{'})} (because \quad L = 2) \\ &= \frac{[Vol(G_1^{'}) + Vol(G_2^{'}] \cdot [H^K(G_1^{'}) + H^K(G_2^{'})]}{Vol(G_1^{'}) + Vol(G_2^{'})} \\ &\quad - \frac{Vol(G_1^{'}) \cdot H^K(G_1^{'}) + Vol(G_2^{'}) \cdot H^K(G_2^{'})}{Vol(G_1^{'}) + Vol(G_2^{'})} \\ &= \frac{Vol(G_1^{'}) \cdot H^K(G_2^{'}) + Vol(G_2^{'}) \cdot H^K(G_1^{'})}{Vol(G_1^{'}) + Vol(G_2^{'})} \\ &\geq 0. \end{aligned} \tag{17}$$

$$Because \quad Vol(G^{'}), Vol(G_1^{'}), Vol(G_2^{'}), H^K(G_1^{'}), H^K(G_2^{'}) \geq 0.$$

$$Therefore \quad I(G_1^{'}; G_2^{'}) \geq 0.$$

Therefore, it is proved that the SMI of disconnected graphs is also satisfied non-negativity.

It can be seen from the above property, SMI: $I(G_1; G_2)$ has symmetry and non-negativity. However, whether there is extremum property in SMI. Whether the SMI: $I(G_1; G_2) = 0$ when the graph is independent. We can verify this in the next section with case analysis.

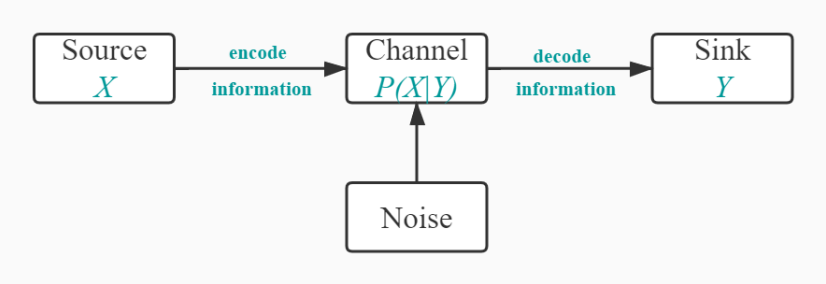


Fig. 4. A communication system model.

C. Communication Channel of Graph Structure

Definition 8 (Communication Channel): It is a system that the output signal is probabilistic dependent on the input signal. Its characteristics are determined by a transfer probability matrix $P(X|Y)$, which gives the conditional probability distribution of the output with a given input.

The general communication channel model is shown in Fig. 4. When the information transmission rate of the channel does not exceed the channel capacity, the appropriate channel coding method can be used to achieve any high transmission reliability. But if the information transmission rate exceeds the channel capacity, it is impossible to achieve reliable transmission. In general, the capacity of a communication channel is defined as $C = max\{I(X;Y)\}$ for the input X and the output Y.

The channel capacity C is the maximum amount of information transmission that the channel can accommodate under the guarantee of reliable communication in advance. For a fixed channel, the channel capacity C is a fixed value. For different channels, C is different, and C is a function of the channel transition probability $P(X|Y)$.

$I(X;Y) = H(X) - H(X|Y)$ represents the information amount that the channel transmits in the whole communication process between X and Y, while the channel capacity is the maximum information amount that we can transmit. For the graph structure, the SMI: $I(G_1;G_2)$ represents the information amount transmitted by the channel in the whole communication process between graphs G_1 and G_2, and the channel capacity is the maximum information amount we can transmit. Therefore, the channel capacity between source X and sink Y can also be used to define the channel capacity of G_1 and G_2 in the graph structure.

Theorem 3: Channel capacity of graph structure. The channel capacity C is equal to the maximum value of SMI.

$$C = max\{I(G_1;G_2)\}. \tag{18}$$

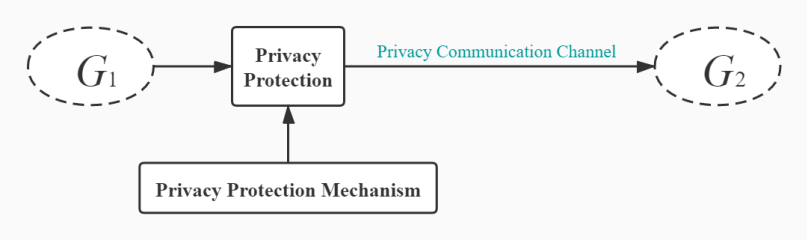


Fig. 5. Communication system model of privacy protection.

The TMI is to measure the strength of the correlation between two points, measure the uncertainty between points and quantify the degree of leakage of their private data. However, for the graph structure, Shannon entropy is no longer sufficient to measure the information amount between one or more graph structures. Firstly, we construct two model subgraphs G_1 and G_2 as shown in Fig. 5, which are transmitted through the privacy communication channel and added the privacy protection mechanism. Further introduce the average privacy SMI: $I(G_1;G_2)$ to describe the degree of privacy leakage on the channel. $I(G_1;G_2)$ represents the average amount of information exchanged between subgraphs G_1 and G_2, that is, the amount of privacy information transmitted on the channel. It can precisely describe the overall degree of privacy leakage, which can be used as a measure of privacy leakage.

IV. Theoretical Analysis and Experimental Evaluation

In this section, we evaluate the performance of our theory via case analysis, and give detailed experimental results and discussion.

A. Structural Mutual Information of Connected Graphs

Given an undirected connected graph $G = (V, E)$ with n vertices and m edges. Then we randomly divide the graph G into G_1 and G_2. Firstly, calculate the structure entropy of the graph G and the subgraphs G_1 and G_2 according to Algorithm 1, and then calculate the SMI: $I(G_1;G_2)$ of the connected graph G. In this algorithm, we calculate the SMI of the subgraphs G_1 and G_2 in the connected graph G based on the structure entropy. It can be seen that line 2 needs to scan graph G and two subgraphs G_1 and G_2, and line 3 needs to scan all nodes in the graph. So, the time complexity of Algorithm 1 is $O(n)$.

Case study 1: Given an undirected connected graph $G = (V, E)$, with 5 vertices and 7 edges. Then we randomly divide the graph G into G_1 and G_2 as shown in Fig. 6.

Algorithm 1 Structural Mutual Information (SMI) of Connected Graphs.

Input:

Undirect and connected graph $G = (V, E)$.

Output:

SMI of connected graphs: $I(G_1; G_2)$.

Randomly divide the graph into two subgraphs G_1 and G_2, n is the number of vertices in different graphs
for each $G_j(j = 0, 1, 2)$ **do**
  for each $\alpha_i(i = 1, 2, ..., n)$ **do**
    $H^T(G_j; \alpha_0) = 0$
    $H^T(G_j; \alpha_i) = -\frac{g_{\alpha_i}}{2m} \log_2 \frac{V_{\alpha_i}}{V_{\alpha_i^-}}$
    $H^T(G_j) += H^T(G_j; \alpha_i)$
    $H^K(G_j) = \min_T (H^T(G_j))$
  end for
end for
$I(G_1; G_2) = H^K(G_1) + H^K(G_2) - H^K(G_0)$
return $I(G_1; G_2)$
G_0 represents the whole graph G

Combining the structural information of the connected graph proposed by the Li and Pan [1], we can use the following formula to calculate the K-dimensional structural information of the whole graph G and the subgraphs G_1 and G_2 respectively:

$$\begin{aligned} H^T(G; \alpha) &= -\frac{g_\alpha}{2m} \log_2 \frac{V_\alpha}{V_{\alpha^-}}. \\ H^T(G) &= \sum_{\alpha \in T, \alpha \neq \lambda} H^T(G; \alpha). \\ H^K(G) &= \min_T \{H^T(G)\}. \end{aligned} \tag{19}$$

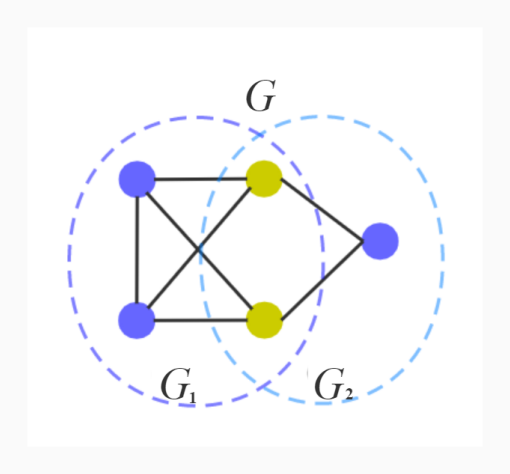


Fig. 6. An example of a connected graph based on algorithm 1.

Then the SMI of connected graph is calculated by Algorithm 1:

$$\begin{aligned} I(G_1;G_2) &= H^K(G_1) + H^K(G_2) - H^K(G) \\ &= \min_T\{H^T(G_1)\} + \min_T\{H^T(G_2)\} - \min_T\{H^T(G)\} \\ &= \min_T\{\sum_{\alpha\in T,\alpha\neq\lambda} H^T(G_1;\alpha)\} + \min_T\{\sum_{\alpha\in T,\alpha\neq\lambda} H^T(G_2;\alpha)\} \\ &\quad - \min_T\{\sum_{\alpha\in T,\alpha\neq\lambda} H^T(G;\alpha)\} \\ &= 1.57 + 1.29 - 1.74 \\ &= 1.12. \end{aligned} \tag{20}$$

Therefore, it can be seen from the calculation that the K-dimensional structural information of G, G_1 and G_2 are 1.74, 1.57 and 1.29, respectively. Finally, the SMI of the connected graph is $I(G_1;G_2) = 1.12$. The analysis of the example reveals that the SMI of connected graph is the reduced uncertainty of graph G_1 due to the known graph G_2.

Extremum property of connected graphs: SMI of connected graphs has extremum property, i.e.,

$$I(G_1;G_2) \leq \min\{H^K(G_1), H^K(G_2)\}.$$

Proof 5:

$$\begin{aligned} I(G_1;G_2) &\leq H^K(G_1) \\ I(G_1;G_2) &\leq H^K(G_2) \\ I(G_1;G_2) &= H^K(G_1) - H^K(G_1|G_2) \\ &= H^K(G_2) - H^K(G_2|G_1). \end{aligned} \tag{21}$$

Conditional structure entropy $H^K(G_1|G_2)$ and $H^K(G_2|G_1)$ is non-negative. When subgraphs G_1 and G_2 are in an overlapping relationship, $I(G_1;G_2) = H^K(G_1) = H^K(G_2)$, then $H^K(G_1|G_2) = 0$. Information about another graph can be fully obtained from one graph. At this time, the channel capacity reaches its maximum value, which means that the amount of information of subgraph G_1 can all pass through the channel. Therefore, the two inequalities are valid and satisfy the extremity property. Because $1.12 < 1.57$ and $1.12 < 1.29$, thus, the experimental results satisfied the extremum property.

B. Structural Mutual Information of Disconnected Graphs

Given an undirected disconnected graph $G^{'} = (V, E)$ with n vertices and m edges, and then mark the connected components of graph $G^{'}$ as $G_j^{'}(j = 1, 2, ...L)$. Here, we assume that L is 2. Firstly, calculating the structure entropy of the graph $G^{'}$ and the subgraphs $G_1^{'}$ and $G_2^{'}$ according to Algorithm 2, and then calculate the SMI: $I(G_1^{'}; G_2^{'})$ of the disconnected graph $G^{'}$. In this algorithm, we calculate the SMI of the subgraphs $G_1^{'}$ and $G_2^{'}$ in the disconnected graph $G^{'}$ based on structure entropy. It can be seen that line 3 needs to scan all connected components in the graph $G^{'}$, and line 4 needs to scan all nodes in the graph. So the time complexity of Algorithm 2 is $O(n^2)$.

Case study 2: Given an undirected and disconnected graph $G^{'} = (V, E)$, with 12 vertices and 19 edges as shown in Fig. 7. We can see that there are two connected components, and we mark the two connected components as subgraphs $G_1^{'}$ and $G_2^{'}$ respectively.

Combining the structural information of the disconnected graphs proposed by the Li and Pan [1], we can use the following formula to calculate the K-dimensional structural information of the whole disconnected graph $G^{'}$ and the subgraphs $G_1^{'}$ and $G_2^{'}$ respectively:

$$\begin{aligned} H^T(G^{'};\alpha) &= -\frac{g_\alpha}{2m}\log_2\frac{V_\alpha}{V_{\alpha^-}} \\ H^T(G^{'}) &= \sum_{\alpha\in T, \alpha\neq\lambda} H^T(G^{'};\alpha) \\ H^K(G^{'}) &= \min_T\{H^T(G^{'})\} \\ H^K(G^{'}) &= \frac{1}{Vol(G^{'})}\cdot\sum_{j=1}^{L} Vol(G_j^{'})\cdot H^K(G_j^{'}) \end{aligned} \tag{22}$$

Algorithm 2 Structural Mutual Information (SMI) of Disconnected Graphs.

Input:

Undirect and disconnected graph $G^{'} = (V, E)$.

Output:

SMI of disconnected graphs: $I(G_1^{'}; G_2^{'})$.

L is the number of connected components of a disconnected graph, and set $L = 2$.

$G_1^{'}$ and $G_2^{'}$ are two connected components of a disconnected graph, n is the number of vertices in different graphs.

for each $G_j^{'}(1, 2, ..., L)$ **do**

  for each $\alpha_i (i = 1, 2, ..., n)$ **do**

    $H^T(G_j^{'}; \alpha_0) = 0$

    $H^T(G_j^{'}; \alpha_i) = -\frac{g_{\alpha_i}}{2m} \log_2 \frac{V_{\alpha_i}}{V_{\alpha_i}}$

    $H^T(G_j^{'}) += H^T(G_j^{'}; \alpha_i)$

    $H^K(G_j^{'}) = \min_T (H^T(G_j^{'}))$

  end for

end for

for each $G_j^{'}(1, 2, ..., L)$ **do**

  $H^K(G^{'}) += \frac{1}{Vol(G^{'})} \cdot \sum_{j=1}^{L} Vol(G_j^{'}) \cdot H^K(G_j^{'})$

end for

$I(G_1^{'}; G_2^{'}) = H^K(G_1^{'}) + H^K(G_2^{'}) - H^K(G^{'})$

return $I(G_1^{'}; G_2^{'})$

Then the SMI of disconnected graph is calculated by Algorithm 2:

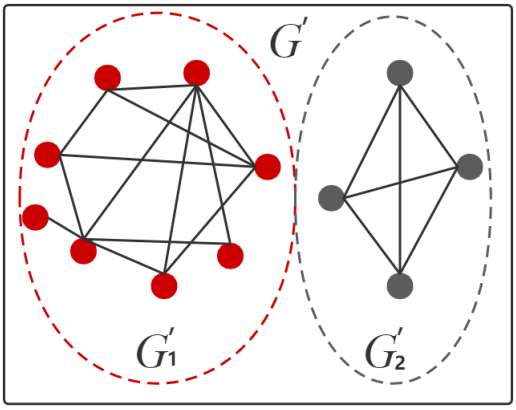


Fig. 7. An example of a disconnected graph based on algorithm 2.

$$
\begin{aligned}
I(G_1^{'};G_2^{'}) &= H^K(G_1^{'}) + H^K(G_2^{'}) - H^K(G^{'}) \\
&= \min_T\{H^T(G_1^{'})\} + \min_T\{H^T(G_2^{'})\} - \frac{1}{Vol(G^{'})} \cdot \sum_{j=1}^{2} Vol(G_j^{'}) \cdot H^K(G_j^{'}) \\
&= \min_T\{\sum_{\alpha\in T,\alpha\neq\lambda} H^T(G_1^{'};\alpha)\} + \min_T\{\sum_{\alpha\in T,\alpha\neq\lambda} H^T(G_2^{'};\alpha)\} \\
&\quad - \frac{Vol(G_1^{'}) \cdot H^K(G_1^{'}) + Vol(G_2^{'}) \cdot H^K(G_2^{'})}{Vol(G^{'})} \\
&= \min_T\{\sum_{\alpha\in T,\alpha\neq\lambda} H^T(G_1^{'};\alpha)\} + \min_T\{\sum_{\alpha\in T,\alpha\neq\lambda} H^T(G_2^{'};\alpha)\} \\
&\quad - \frac{Vol(G_1^{'}) \cdot \min_T\{\sum_{\alpha\in T,\alpha\neq\lambda} H^T(G_1^{'};\alpha)\}}{Vol(G^{'})} - \frac{Vol(G_2^{'}) \cdot \min_T\{\sum_{\alpha\in T,\alpha\neq\lambda} H^T(G_2^{'};\alpha)\}}{Vol(G^{'})} \\
&= 2.07 + 1.67 - 1.94 \\
&= 1.8.
\end{aligned}
\tag{23}
$$

So, it can be seen from the calculation that the K-dimensional structure information of $G^{'}$, $G_1^{'}$ and $G_2^{'}$ are 1.94, 2.07 and 1.67, respectively. Finally, the SMI of the disconnected graph is $I(G_1^{'};G_2^{'}) = 1.8$. The analysis of the example shows that when the two subgraphs do not have the same part, that is, they are independent of each other, the SMI: $I(G_1^{'};G_2^{'}) \neq 0$. Because $1.8 > 1.67$, there is no extremum property in the SMI of disconnected graphs. It can be seen that the time complexity of Algorithm 2 is $O(n^2)$.

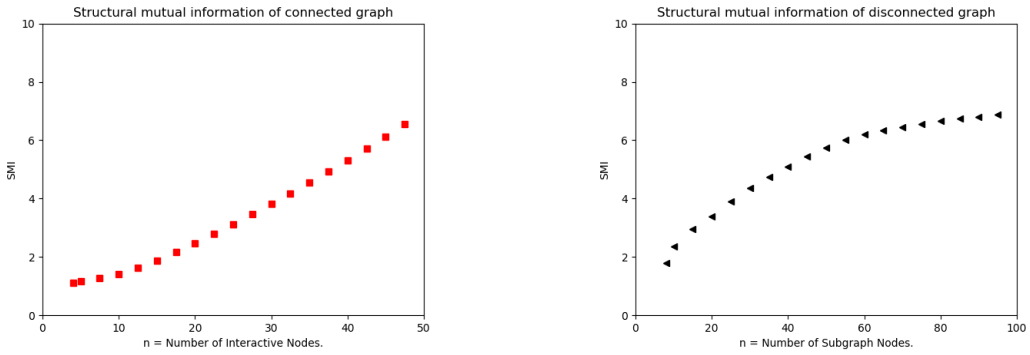
Subsequently, we conducted a lot of experiments and found that when the number of nodes interacting between connected graphs is increasing, the SMI of connected graphs is increasing. When the scale gap between disconnected graphs is expanding, the SMI of disconnected graphs is also increasing. As shown

TABLE I

THE COMPARISON OF TMI AND SMI.

	Symmetry	Non-negativity	Extremum property	Independence
TMI [6]	Yes	Yes	Yes	Yes
SMI	Yes	Yes	No	No

in Fig. 8. When the graph is a connected graph, the SMI of the connected graph increases with the increase of the interaction nodes between the subgraphs G_1 and G_2. When the graph is a disconnected graph, the number of nodes in the subgraph $G_1^{'}$ keeps increasing, while the number of nodes in the subgraph $G_2^{'}$ remains unchanged or decreases, the SMI of the disconnected graph gradually increases. Therefore, the SMI can be used to measure the mutuality among graph structures and reflect the channel capacity among graph structures.



(a) (b)

Fig. 8. The value of the SMI of the connected graph changes with the change of the interaction nodes between subgraphs G_1 and G_2. When there are more interaction nodes between two subgraphs, the greater the SMI: $I(G_1; G_2)$ (a). The value of the SMI of the disconnected graph changes with the difference in the size of the subgraphs G_1 and G_2. When the difference in the number of nodes in the two subgraphs is greater, the SMI: $I(G_1^{'}; G_2^{'})$ is greater (b).

Finally, we can clearly understand the difference in the nature of SMI and TMI through the definition and example analysis of SMI of connected and disconnected graphs, as shown in Table 1. Therefore, the SMI can also be used as a tool for privacy measurement. Measure the level of privacy among graph structures, and describe the degree of privacy leakage and correlation among graph structures.

Intuitively speaking, The TMI measures the information shared by X and Y. It measures the extent to which one of these two variables knows and the uncertainty decreases for the other. As can be seen

from Table 1, the TMI has symmetry, non-negativity, extremum property and independence. The TMI mainly solves the information transmission between point-to-point communication, and cannot effectively measure the amount of information embedded in the graph structure. The SMI is also symmetry and non-negativity, so we can use SMI to measurement information amount embedded in the graph structure and the degree of privacy leakage. The non-extremum property and non-independence of SMI can more effectively describe the degree of correlation between disconnected graphs and the degree of leakage of private data.

Therefore, we propose SMI based on structure entropy. Using SMI to measurement information amount embedded in graph structure and the uncertainty between them. Furthermore, the SMI can be applied to large complex networks or social networks to describe the overall degree of privacy leakage.

V. Privacy Measurement Methods

Privacy metrics is an essential method for evaluating privacy protection technology. It can be used to measure the protection intensity of privacy protection technology to users' privacy. Privacy metrics mainly include the amount of privacy information contained in the graph structure itself, the adversary's attack capabilities, privacy protection algorithms, and the degree of privacy leakage. The amount of information in the graph structure itself can be described by the size of the structure entropy $H^K(G)$, representing the inherent amount of privacy information in the graph structure. However, The description of the degree of correlation among graph structures can be expressed by SMI: $I(G_1;G_2)$. As a measure of privacy leakage, the SMI: $I(G_1;G_2)$ can be used to metric the level of privacy among graph structures and describe the overall degree of privacy leakage. Thus, the SMI: $I(G_1;G_2)$ indicates that subgraph G_1 is protected by the privacy protection mechanism, the amount of privacy information captured by subgraph G_2. The SMI should be as small as possible.

In research in the field of privacy protection, privacy measurement usually reveals the risk of privacy information leakage in privacy protection methods through measurement indicators or measurement methods. From the perspective of leaking privacy information, it reflects the privacy protection strength of privacy protection methods. Privacy measurement originated from privacy anonymity technology. In the research process of privacy anonymity protection technology, privacy measurement has been paid much attention by researchers. After that, scholars have successively proposed privacy measurement methods based on mutual information, set pair theory and differential privacy. Traditional privacy measurement methods are mainly aimed at small-scale, structured, and data stored in traditional relational databases, but can not effectively measure data in large-scale and unstructured groups. Therefore, SMI can effectively measure and describe the degree of correlation between data privacy in massive social networks or

complex communication networks. Then, we analyze and compare our method with the other four privacy measurement methods.

A. Privacy Measurement Method Based on Anonymity

The K-anonymity value is one of the classic indicators in privacy measurement. In K-anonymity, the K-anonymity value indicates the anonymity of the quasi-identifier attribute in the data set. The greater the value of K-anonymity in anonymous data, the more difficult it is for an attacker to speculate on privacy information, and the stronger the privacy protection. Because K-anonymity only anonymizes deal with quasi-identifier attributes in the data, there are no constraints on sensitive attributes. The attacker can use the background knowledge related to privacy information to infer the correspondence between users and sensitive attribute values based on the distribution of sensitive attribute values in the anonymous data set. Therefore, relying only on K-anonymity values as the measurement index for the anonymous data set is not comprehensive enough and the measurement results are not accurate enough. The author [31] proposed a measurement method based on calculating the value distribution of sensitive attributes based on K-anonymity and L-diversity. Under the condition of K-anonymity, EMD (earth mover's distance) method is used to calculate the difference between the global distribution of sensitive attribute values in data and the distribution of the same sensitive attribute values in any equivalence class. The larger the K-anonymity value, the smaller the difference, and the higher the anonymity, the smaller the risk of privacy information leakage. However, EMD method does not consider the stability of the sensitive attribute value distribution between the equivalence class and the data, resulting in a limited application range of the measurement method.

B. Privacy Measurement Method Based on Mutual Information

Mutual information $I(X;Y)$ is a method for describing the degree of privacy leakage in information entropy. Mutual information $I(X;Y)$ represents a measure of the interdependence between two random variables. The risk of privacy leakage is reflected by the amount of uncertainty reduction about the privacy information in the original data before and after the attacker knew information is calculated by mutual information. The greater the amount of uncertainty reduction, the greater the risk of leakage of private information. The author [32] uses mutual information to quantify the degree of independence between the data owned by the attacker and the original data. An attacker can obtain a subset of data associated with privacy information from the data that he owns, and mutual information can reflect the degree of independence between data subsets and privacy information. The larger the mutual information value, the more the uncertainty of privacy information in the data decreases, the stronger the correlation between the

two, and the higher the risk of privacy information leakage. However, because the measurement results are easily affected by outliers, erroneous data, and incomplete data, the accuracy of the measurement is low.

C. Privacy Measurement Method Based on Set Pair Analysis

Set pair analysis is an analysis method that combines qualitative and quantitative and can solve certainty and uncertainty. The author [33] proposed a set pair analysis privacy measurement method based on set pair analysis theory, which measures and analyzes the relationship between data sets. Firstly, the degree of identity, degree of difference and degree of opposition of the adjacent data subset pairs of the data set are analyzed and quantified, and the connection degree expression is obtained. In order to prevent the influence of the attribute with a larger or smaller starting value on the weight in the data set, data processing is performed on the inconsistency and noise data in the data set through numbers or special symbols. The processing method of the attribute value of the alignment identifier is to take the minimum value $minB_i$, the maximum value $maxB_i$ and the actual value B_i in the set to construct a ternary interval, i.e., : $[B_i] = [minB_i, B_i, maxB_i,]$. Then use set pair analysis to establish a set pair connection degree expression for each attribute value in the data set, and convert the ternary interval to $\mu(B_i) = a_{B_i} + b_{B_i} + c_{B_i}$. Then calculate the potential value or scoring function between the degree of identity, degree of difference and degree of opposition in the expression to reflect the degree of privacy protection leakage. According to the calculated potential value, the privacy information is divided into the same potential, equal potential and counter potential. The risk of privacy information leakage in the process of measuring data publishing or sharing can be analyzed and obtained. The same potential means that the risk of leakage of private information is low, and the counter potential means that the risk of privacy information leakage is high, and the equal potential is between the same potential and the counter potential. Because it is difficult to establish the set pair relationship between data, and is not combined with the attacker's background knowledge, the accuracy and efficiency of measurement are relatively low.

D. Privacy Measurement Method Based on Differential Privacy

While achieving privacy protection, differential privacy protection technology provides a method for quantifying and assessing the risk of privacy information leakage. This technology usually achieves the division of privacy protection strength by setting different ε values (differential privacy protection budget). The smaller of ε, the more noise is added, the lower the risk of privacy information leakage, and the stronger the privacy protection. When there is a correlation between tuples in the data, an attacker can infer the privacy information of the tuples associated with it by observing one or more tuples in the data.

Therefore, under differential privacy protection, the measurement of data privacy protection intensity is not only limited by its own ε, but also affected by the ε' of the associated data set. Information about the existence of some users is directly or indirectly related to tuples of different data. The author [34] proposed a differential privacy measurement method associated with multiple data sets, used to measure the risk of privacy information leakage when tuples in a data set are affected by tuples in other data sets. Therefore, this method has high accuracy and efficiency. However, the scope of application is limited, and when data tuples are related to each other, it is difficult to reflect the protection intensity of the associated privacy.

E. Comparison and Analysis of Privacy Measurement Methods

Through the above analysis, we compared the scale, application range, accuracy and efficiency of the four measurement methods with our methods. As shown in Table 2, the performance is between 0 and 1.

Privacy measurement methods based on anonymity measure the strength of privacy protection depending on the degree of data anonymity. The measurement method can be combined with the attacker and the background knowledge of the attacker, which makes the measurement results more comprehensive and accurate. However, the process of measuring large-scale multi-dimensional data is relatively complicated and processing efficiency is low, resulting in a limited application range.

The privacy measurement method based on mutual information can establish the corresponding measurement model for the attacker and the background knowledge owned by the attacker to make the measurement more comprehensive. The principle of this method is simple and easy to implement, and large-scale multi-dimensional data can be measured using big data computing technology. However, mutual information as a privacy metric is susceptible to abnormal and erroneous data. The accuracy of the measurement results will also be affected by the probability distribution.

The key to privacy measurement method based on set pair analysis is to analyze and measure the degree of identity, degree of difference and degree of opposition of adjacent data sets. This method does not combine the attacker and the attacker's background knowledge to analyze and measurement, thus affecting the accuracy of privacy measurement results, and at the same time, it has low efficiency for large-scale data.

The privacy measurement method based on differential privacy measurement the strength of privacy protection mainly depends on the value of ε in differential privacy. Analyzing the value of ε can reflect the strength of privacy protection. It is simpler to implement and more efficient. This method can not only measure small-scale data, but also use big data technology to realize the measurement of larger-scale

multi-dimensional data. However, when the data tuples are related to each other, this method cannot reflect the strength of protecting of privacy information only through the value of ε. Therefore, this method is mainly applicable to the measurement of privacy protection strength for differential privacy protection technology, and the application range is relatively small.

We use SMI: $I(G_1; G_2)$ to describe the privacy information leakage risk. Assuming that the privacy information of data in the graph structure is P. After the graph structure is processed by privacy protection technology and the attacker obtains data D associated with privacy information. By calculating SMI before and after the known information D, the amount of uncertainty about the privacy information P in the graph structure reflects the risk of privacy leakage. The greater the amount of uncertainty reduction, the risk of privacy information leakage bigger. Further use the SMI to quantify the degree of correlation between the data D owned by the attacker and the data in the graph structure. When the SMI value is larger, the uncertainty of privacy information P in the data is reduced more, and the correlation between the two is stronger, the risk of privacy information leakage is higher.

TABLE II

PERFORMANCE ANALYSIS

	method	***Range***	***Scale***	***Accuracy***	***Efficiency***
Li's work[31]	Anonymity	0.5	0.58	0.78	0.54
Humbert's work[32]	Mutual information	0.7	0.78	0.55	0.8
Yan's work[33]	Set pair analysis	0.78	0.62	0.5	0.53
Wu's work[34]	Differential privacy	0.55	0.8	0.7	0.68
Our work	SMI	0.83	0.9	0.93	0.85

This paper proposed a measure method of SMI, which can measure the high-dimensional or depth information embedded in a graph structure. At the same time, the real structure of the original system is decoded, and measure the interaction among the graph structures and the risk of privacy leakage. Since the structure entropy is the natural extension of Shannon entropy, it extends from unstructured probability distribution to arbitrary structure graphs. Therefore, the SMI can measure the information embedded in graph structure more effectively and comprehensively. The detailed comparison results of the SMI and the other four privacy measurement methods are shown in Fig. 9. As can be seen from Fig. 9, the privacy measurement method of SMI is the highest accuracy and the lowest range. But all aspects are better than

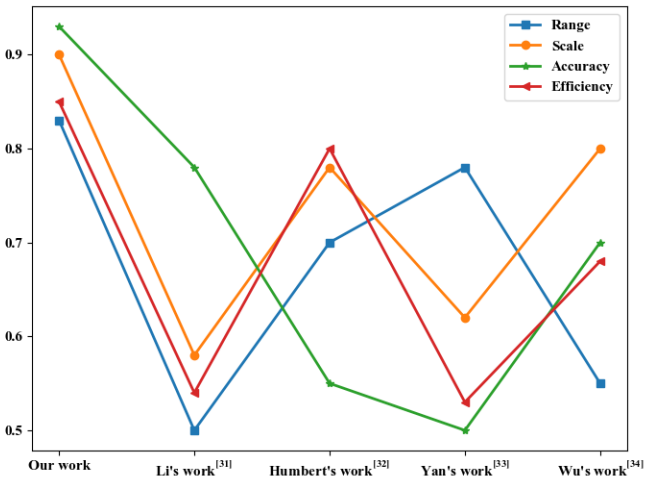


Fig. 9. Performance analysis and comparison of the SMI with other four privacy measurement methods in terms of range, scale, accuracy and efficiency.

the other four measurement methods. Because the SMI can not only accurately and effectively measure the information in the graph structure, but also can be applied to all structured or unstructured large-scale network graphs. So, compared with other methods, our method can not only use SMI to measure the risk of privacy information leakage in large-scale complex networks, but also have higher efficiency, wider application range and precise accuracy. In general, our method has good efficiency and usability in privacy measurement.

VI. Conclusion

In this paper, we proposed the SMI theory based on K-dimensional structural information and Shannon mutual information. The key starting point is to measure the interaction of the graph structure, describes the overall privacy leakage risk and the degree of correlation among graph structures. We introduced the definition of structure entropy of connected graphs and disconnected graphs in structural information, and then used the correlation characteristics of mutual information to propose SMI. In addition, we compared the properties of SMI and TMI through algorithm and case analysis, and found that SMI has different properties. However, the properties and methods proposed in this paper are not the end of our work.

In future work, we will further study some problems that have not been solved in this paper, e.g., (1) whether there are other properties of SMI proposed in this paper; (2) when the undirected graph is extended to the directed graph or weighted graph, how does the SMI describe or better describe the overall privacy leakage risk and the correlation degree among graph structures; (3) how to use SMI to better

apply in the existing privacy protection technology. Although the theory of SMI is proposed for the first time in this paper and the degree of privacy leakage and related properties among graph structures are not perfect, a feasible theoretical foundation has been established in order to privacy protection technology and privacy leakage analysis and evaluation.

ACKNOWLEDGMENT

We would like to thank the anonymous reviewers for their valuable comments and suggestions.

REFERENCES

[1] Li, A., Pan, Y.: Structural information and dynamical complexity of networks[J]. IEEE Transactions on Information Theory, 2016, 62(6):3290-3339.

[2] Li, A., Yin, X., Xu, B., et al.: Decoding topologically associating domains with ultra-low resolution Hi-C data by graph structural entropy[J]. Nature Communications, 2018, 9(1):3265-3226.

[3] Li, A., Pan, Y.: Structure entropy and resistor graphs[J]. CoRR abs, 2018, 1801.03404.

[4] Li, A., Li, J., Pan, Y.: Discovering natural communities in networks[J]. Physica A Statistical Mechanics & Its Applications, 2015, 436:878-896.

[5] Brooks, F.: Three great challenges for half-century-old computer science[J]. Journal of the ACM, 2003, 50(1):25-26.

[6] Shannon, C.: A mathematical theory of communication. The Bell System Technical Journal, 1948, 27(3):379-423.

[7] Erdös, P., Rényi, A.: On random graphs I. Publicationes Mathematicae, 1959, 6:290-297.

[8] Erdös, P., Rényi, A.: On the evolution of random graphs. Publicationes Mathematicae, 1960, 5:17-61.

[9] Watts, D., Strogatz, S.: Collective dynamics of 'small-world' networks. Nature, 1998, 393:440-442.

[10] Newman, M., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E, 2003, 69(2):026113.

[11] Kamisiński, A., Chołda, P., Jajszczyk. A.: Assessing the structural complexity of computer and communication networks. ACM Computing Surveys, 2015, 47(4):66:1-36.

[12] Glantz, R., Meyerhenke, H., Schulz, C.: Tree-based coarsening and partitioning of complex networks. ACM Journal of Experimental Algorithmics, 2016, 21(1):1.6:1-20.

[13] Wang, C., Zhu, H., Wang, C., et al.: Transport complexity of data dissemination in large-scale online social networks. In: ACM TURC'19, 2019, 39:1-5.

[14] Sansavini, F., Parigi, V.: Continuous variables graph states shaped as complex networks: optimization and manipulation. Entropy, 2020, 22(1),26.

[15] Chatterjee, T., DasGupta, B., Mobasheri, N., et al.: On the computational complexities of three problems related to a privacy measure for large networks under active attack. Theoretical Computer Science, 2019, 775:53-67.

[16] Hsiao, C., Lu, C., Reyzin, L.: Conditional computational entropy, or toward separating pseudoentropy from compressibility. In: EUROCRYPT 2007. LNCS, 2007, 4515:169-186. Springer (2007)

[17] Whitnall, C., Oswald, E.: A comprehensive evaluation of mutual information analysis using a fair evaluation framework. In: CRYPTO 2011. LNCS, 2011, 6841:316-334. Springer (2011)

[18] Taghia, J., Martin, R.: Objective intelligibility measures based on mutual information for speech subjected to speech enhancement processing. In: IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(1):6-16.

[19] Chitambar, E., Fortescue, B., Hsieh MH.: Distributions attaining secret key at a rate of the conditional mutual information. In: CRYPTO 2015. LNCS, 2015, 9216:443-462. Springer (2015)

[20] Perotti, J., Tessone, C., Caldarelli, G.: Hierarchical mutual information for the comparison of hierarchical community structures in complex networks[J]. Physical Review E, 2015, 92(6):062825.

[21] Li, Y., Cai, W., Li, Y., et al.: Key node ranking in complex networks: a novel entropy and mutual information-based approach. Entropy, 2020, 22(1), 52.

[22] Viegas, E., Goto, H., Kobayashi, Y., et al.: Allometric scaling of mutual information in complex networks: a conceptual framework and empirical approach. Entropy, 2020, 22(2), 206.

[23] Romashchenko, A., Zimand, M.: An operational characterization of mutual information in algorithmic information theory[J]. Journal of the ACM, 2019, 66(5):38:1-42.

[24] Wan, P., Chen, X., Tu, J., et al.: On graph entropy measures based on the number of independent sets and matchings. Information Sciences, 2020, 516:491-504.

[25] Guo, H., Ma, Y., Tuskan, G., et al.: A suggestion of converting protein intrinsic disorder to structural entropy using Shannon's information theory. Entropy, 2019, 21(6), 591.

[26] Liu, Y., Liu, J., Zhang, Z., et al.: REM: from structural entropy to community structure deception. In: NeurIPS 2019:12918-12928

[27] Chen, Y., Liu, J.: Distributed community detection over blockchain networks based on structural entropy. In: BSCI 2019:3-12.

[28] Markechová, D.: Kullback-leibler divergence and mutual information of experiments in the fuzzy case.[J]. Axioms, 2017, 6(4):5-18.

[29] Huffffman, D.: A method for the construction of minimum-redundancy codes[J]. Resonance, 2006, 11(2):91-99.

[30] Corso, G., Ferreira, G., Lewinsohn, T.: Mutual information as a general measure of structure in interaction networks. Entropy, 2020, 22(5), 528.

[31] Li, N., Li, T., Nkatasubramanian, S.: (n,t)-Closeness: A new privacy measure for date publishing. IEEE Transactions on Knowledge and Date Engineering, 2010, 22(7):943-956.

[32] Humbert, M., Ermanayda, Y., Hubaux, J., et al.: Quantifying interdependent risks in genomic privacy. ACM Transactions on Privacy & Security, 2017, 20(1):1-30.

[33] Yan, Y., Hao, X., Wang, W.: A set pair analysis method for privacy metric. Engineering Journal of Wuhan University, 2015, 48(6):883-890 (in Chinese with English abstract).

[34] Wu, X., Dou, W., Ni, Q.: Game theory based privacy preserving analysis in correlated data publication. In: Australasian Computer Science Week Multiconference. ACM, 2017:73-82.

[35] Håstad, J., Impagliazzo, R., Levin, L., et al.: Construction of pseudorandom generator from any one-way function. SIAM Journal on Computing, 1999, 28(4):1364-1396.