

基于结构熵约束的图聚类方法

张志颖^{1,2}, 田有亮^{1,2,3}

- (1. 贵州大学计算机科学与技术学院, 贵州 贵阳 550025;
2. 贵州省公共大数据重点实验室, 贵州 贵阳 550025;
3. 贵州大学密码学与数据安全研究所, 贵州 贵阳 550025)

摘要: 针对以大数据为中心的信息开放共享平台, 如何从嵌入大规模噪声结构的网络中解码出网络的真实结构, 进一步在挖掘关联信息的过程中得到较为准确的挖掘结果的问题, 提出基于结构熵的聚类方法实现对图中节点关联程度的划分。提出了计算二维结构信息的求解算法和基于熵减原则的模块划分算法, 对图结构中节点划分得到对应的模块; 利用 K 维结构信息算法对已划分的模块做进一步的划分, 实现对图结构中节点的聚类; 通过实例分析表明, 所提出的图聚类方法不仅能够反映图结构的真实结构, 而且可以有效地挖掘出图结构中节点之间的关联程度。同时对比了其他 3 种聚类方法, 实验表明该方法在执行时间上具有更高的效率和保证聚类结果的可靠性。

关键词: 数据挖掘; 结构信息; 结构熵; 图聚类

中图分类号: TP309

文献标识码: A

DOI: 10.11959/j.issn.2096-109x.2021098

Graph clustering method based on structure entropy constraints

ZHANG Zhiying^{1,2}, TIAN Youliang^{1,2,3}

1. College of Computer Science and Technology, Guizhou University, Guiyang 550025, China
2. Guizhou Provincial Key Laboratory of Public Big Data, Guiyang 550025, China
3. Institute of Cryptography & Data Security, Guizhou University, Guiyang 550025, China

Abstract: Aiming at the problem of how to decode the true structure of the network from the network embedded in the large-scale noise structure at the open information sharing platform centered on big data, and furthermore

收稿日期: 2021-04-15; 修回日期: 2021-06-31

通信作者: 田有亮, youliangtian@163.com

基金项目: 国家自然科学基金(61662009, 61772008); 贵州省科技重大专项计划(20183001); 国家自然科学基金联合基金重点支持项目(U1836205); 贵州省科技计划项目(黔科合基础[2019]1098); 贵州省高层次创新型人才项目(黔科合平台人才[2020]6008)

Foundation Items: The National Natural Science Foundation of China (61662009, 61772008), Science and Technology Major Support Program of Guizhou Province (20183001), Key Program of the National Natural Science Union Foundation of China (U1836205), Science and Technology Program of Guizhou Province ([2019]1098), Project of High-level Innovative Talents of Guizhou Province ([2020]6008)

论文引用格式: 张志颖, 田有亮. 基于结构熵约束的图聚类方法[J]. 网络与信息安全学报, 2021, 7(6): 155-166.

ZHANG Z Y, TIAN Y L. Graph clustering method based on structure entropy constraints[J]. Chinese Journal of Network and Information Security, 2021, 7(6): 155-166.

accurate mining results can be obtained in the mining related information process, the method of clustering based on structure entropy was proposed to realize divide the correlation degree of nodes in the graph. A solution algorithm for calculating two-dimensional structural information and a module division algorithm based on the principle of entropy reduction were proposed to divide the nodes in the graph structure to obtain corresponding modules. The K -dimensional structural information algorithm was used to further divide the divided modules to realize the clustering of nodes in the graph structure. An example analysis shows that the proposed graph clustering method can not only reflect the true structure of the graph structure, but also effectively mine the degree of association between nodes in the graph structure. At the same time, the other three clustering schemes are compared, and the experiment shows that this scheme has higher efficiency in execution time and guarantees the reliability of the clustering results.

Keywords: data mining, structural information, structure entropy, graph clustering

1 引言

在大数据和 5G 时代的环境下,云计算、大数据和物联网等为代表的信息技术得到了前所未有的飞速发展^[1-2]。新技术的不断涌现和发展,让人们意识到数据的价值和开放共享^[3]的重要性。例如,在物联网的大数据社交电子商务中,智能设备通过感知用户的偏好和浏览足迹,为用户提供精准的商品和便捷的服务^[4]。因此,在大型网络图中,迫切需要利用一种方法来发现网络图中的真实结构和相关程度。

图聚类是一种针对图中节点间的某种相近或相似性的划分方式^[5],发现图中各节点数据间的相关关系,并通过聚类效果为智能设备提供有效的服务。近年来,图聚类在社交网络^[6]、医学图像^[7]和生物数据网络^[8]等应用领域得到了广泛的发展。因此,根据网络图生成一个无向无权图,可以通过结构熵^[9]来检测图结构中的真实结构,进而更有效地挖掘出图中节点的关联程度。

目前,研究者对共享和挖掘图结构中数据的兴趣不断增长,对社交网络中隐私数据的保护力度也在逐渐加大。Keyvanpour 等^[10]在隐私保护数据挖掘中提出了一种新的自定义隐私模型。进一步扩展了有向无环图的隐私模型和讨论了发布社交网络时的隐私风险^[11],当数据挖掘算法应用于真实数据集时,可被恶意数据挖掘者用来威胁社交网络用户的隐私。Al-Saggaf 等^[12]在社交网络上使用决策森林数据挖掘算法,该算法不仅构建了决策树,而且还建立了一个森林,允许从数据集中探索更多的逻辑规则。数据集的庞大,

导致数据集中存在敏感信息。并且这项技术在信息提取、数据缩减和隐私方面也面临着巨大的技术挑战。Görnerup 等^[13]提出了一种从大量蜂窝网络数据中挖掘路线信息的复合方法,该方法可通过分布式集群有效减少序列数据来解决可伸缩性问题,以及通过将原始数据分组到聚合中之后,在需要时对有关聚合的统计进行扰动以实现进一步的隐私保护。

为了平衡网络的效用和隐私,周艺华等^[14]针对如何在保证隐私安全的前提下,挖掘出有效知识的问题,提出了一种基于聚类的社交网络隐私保护方法,该方法有效地减少了信息损失,提高了数据有效性。Yan 等^[15]提出了一种细粒度的差分隐私机制,用于在线社交网络中的数据挖掘。还考虑了针对差分隐私的共谋攻击,提出了保护隐私数据挖掘的对策。Tian 等^[16]介绍了一种用于将不确定性注入社区以生成不确定性图的隐私保护方法。关键思想是混淆原始图形的一部分,以实现隐私保护和有效地提高数据的实用性,并使用公开的真实数据集评估生成的不确定图的实用性。

针对网络数据挖掘研究成果中存在高冗余和不安全的问题。Jiang 等^[17]提出了一种基于 Apriori 算法的网络隐私数据挖掘方法,基于结合数据冗余分析、过滤和伪装,采用优化的 TS Apriori 算法对伪装后的数据集进行挖掘。实验结果表明,该方法存在冗余度低、安全性强和挖掘精度高等优点。Reza 等^[18]提出了一种 3LP+ (three layers of protection) 隐私保护技术,以保护用户的敏感信息泄露,将 3LP+应用于合成生

成的在线社交网络数据集，并证明 3LP + 优于现有的隐私保护技术。为了分析和查看人们对在线商人、社交网络和其他数字媒体的隐私和安全。Alshaikh 等^[19]使用最广泛的社交网络之一 Twitter，以分析 Twitter 用户之间的交互，为了提取有意义的数据以帮助确定用户对数字世界的隐私和安全的感知。然而以上研究方法并不能检测出图结构中数据结构的真实信息，从而在挖掘图结构中关联信息时可能会存在一定的误差。

因此，本文需要解码出网络结构的真实结构，进一步地挖掘出图结构的关联信息。Li 等^[20]定义了图的结构熵的概念，以度量图中嵌入的信息，并确定和解码图的基本结构。Li 等^[21-23]首先提出了一种通过网络结构熵的度量方法来寻找社区的算法。他们发现，社区检测算法可以准确识别几乎所有自然选择产生的网络社区。随后提出了图的抵抗力的概念，使用一个伴随结构熵的概念来度量图抵抗策略性病毒攻击的级联失败的能力。进一步提出了一种利用结构信息理论的快速且无归一化的算法来解码染色体域。Liu 等^[24]提出了一种基于社区的结构熵来表达由社区结构揭示的信息量。通过利用用户账户的社区隶属关系，攻击者可以推断出敏感的用户属性。这就提出了社区结构欺骗的问题，并着重于在在线社交网络中披露社区结构的隐私风险。

基于以上研究，本文针对大规模动态数据之间的关联程度，提出了一种基于结构熵约束的图聚类方法。相比已有工作，该方法通过引入结构熵来度量动态复杂网络，解码出网络的真实结构信息，解决了现有文献对动态复杂网络中信息关联刻画不足的问题。具体而言，本文的贡献如下。

1) 提出了利用二维结构信息的求解算法和熵减原则的聚类算法。通过算法将图结构中节点进行划分，关联程度强的划分为一个模块，从而挖掘出图结构中关联程度强弱的模块隐私信息。

2) 提出了 K 维结构信息的求解算法。通过算法将已划分的模块做进一步的划分，得到图中所有节点的关联程度。同时结构熵可以解码出在大规模噪声结构中图结构的信息，这个信息能区分图结构中的规律和噪声，从而挖掘出图结构中节

点的实质结构。

3) 通过实例分析和实验评估表明，所提出的图聚类方法不仅能够有效地反映图结构的真实结构，而且还能有效地挖掘出图结构中节点之间的关联程度。与其他 3 种聚类方法相比，该方法在执行时间上具有更高的效率，并且保证了聚类结果的可靠性。

2 相关概念定义

定义 1 (图的编码树^[20-25]) 设图 $G=(V, E)$ ，图的编码树 T ，使得图中的每个节点都在编码树中，存在顶点集 V 的子集 T_α ，并且以下属性成立。

对于根节点 λ ，定义一个集合 $T_\lambda = V$ 。

对于每个节点 $\alpha \in T$ ， $\alpha^{\wedge} \langle j \rangle$ 表示 α 的子节点， j 是 1 到自然数 N 从左到右递增。每个内部节点至少有两个直接后继节点，因此当 $i < j$ 时， $\alpha^{\wedge} \langle i \rangle$ 在 $\alpha^{\wedge} \langle j \rangle$ 左边。

对于 $\alpha \in T$ ，都有一个与 α 相关的子集 $T_\alpha \subset V$ ，对于 α, β 。 α 表示 β 的父节点，如果 $\alpha \neq \lambda$ ， α^- 表示 α 的父节点。

对于每一个 i ， $\{T_\alpha | h(\alpha) = i\}$ 是 V 的一个分区，其中 $h(\alpha)$ 是 α 的高度，根节点的高度为 0，当 $\alpha \neq \lambda$ 时， $h(\alpha) = h(\alpha^-) + 1$ 。

对于每个节点 α ， T_α 是 T_β 对所有 β 的联合， $\beta^- = \alpha$ ，因此 $T_\alpha = \bigcup_{\beta^- = \alpha} T_\beta$ 。

对于 T 的每个叶子节点 α ， T_α 是一个单例。因此 T_α 包含 V 中的单个节点。

对于每个节点 $\alpha \in T$ ，如果 $T_\alpha = X$ 为一组顶点 X ，则 α 是 X 的码字，记 $c(X) = \alpha$ ， X 是 α 的标记，记 $M(\alpha) = X$ 。

定义 2 (结构熵^[20]) 设图 $G=(V, E)$ ，根据图 G 的不同维度得到一维结构熵、二维结构熵和 K 维结构熵。将不同维度的结构熵的最小值定义为图在不同维度下的结构信息。

定义 3 (一维结构信息^[20]) 假设 $G=(V, E)$ 是一个 n 个顶点 m 条边的无向连通图，对于每个顶点 $i \in \{1, 2, \dots, n\}$ 。用 d_i 表示 G 中顶点 i 的度， m 为图边的数量，设 $Q_i = d_i / 2m$ 。然后用概率向量 $\mathbf{Q} = (Q_1, Q_2, \dots, Q_n)$ 来描述图 G 中顶点的平稳分布。由此可定义图 G 的一维结构信息：

$$H^1(G) = H(Q) = H\left(\frac{d_1}{2m}, \dots, \frac{d_n}{2m}\right) = -\sum_{i=1}^n \frac{d_i}{2m} \cdot \log \frac{d_i}{2m} \quad (1)$$

定义 4 (二维结构信息^[20]) 给定一个无向连通图 $G=(V, E)$, 假设 $P=(X_1, X_2, \dots, X_L)$ 是 V 的一个分区, 称 X_j 为一个模块或者社区。通过 P 定义图 G 的结构信息如下:

$$H^p(G) = \sum_{j=1}^L \frac{V_j}{2m} \cdot H\left(\frac{d_1^j}{V_j}, \dots, \frac{d_{n_j}^j}{V_j}\right) - \sum_{j=1}^L \frac{g_j}{2m} \lg \frac{V_j}{2m} = -\sum_{j=1}^L \frac{V_j}{2m} \sum_{i=1}^{n_j} \frac{d_i^j}{V_j} \lg \frac{d_i^j}{V_j} - \sum_{j=1}^L \frac{g_j}{2m} \lg \frac{V_j}{2m} \quad (2)$$

其中, L 是分区 P 中的模块数, n_j 是模块 X_j 中的节点数, d_i^j 是 X_j 中第 i 个节点的度, V_j 是模块 X_j 的体积, 也就是模块 X_j 中所有节点的度之和。 g_j 是模块中 X_j 只有一个端点的边的数量。

因此可以定义图 G 的二维结构信息, 也称之为模块熵:

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

其中, P 是图 G 中所有可能划分出的分区模块。

定义 5 (图聚类^[26]) 根据节点间的某种相近或相似性, 利用聚类方法将图中所有节点划分成一些聚类, 这个过程被称为图聚类。

为了更好地理解图聚类, 本文将图聚类表示为社交网络图 G 上的一种映射关系 $\varphi: (v_1, v_2, \dots, v_n) \rightarrow (X_1, X_2, \dots, X_m)$, φ 满足以下两个条件。

- (1) $\bigcup_{i=1}^m X_i = V(G), X_i \neq \varnothing; m$ 为聚类的数目。
- (2) $1 \leq i, j \leq m, i \neq j, X_i \cap X_j = \varnothing$ 。

3 问题描述

3.1 方案模型

结构熵被用来度量嵌入图结构中的不确定性, 结构熵的最小化是解码图结构中真实结构的一种直观的方法。通过图结构的结构熵最小化, 将图结构的随机变化和噪声引起的扰动减小到最小, 并且在比较图结构的节点模块合并前后结构熵的差异程度过程中, 可反映出图结构中节点之

间的关联程度。如图 1 所示, 在图结构交互过程中如何有效地挖掘出它们之间的关联程度是本文所要考虑的新问题。针对这一问题, 可以用图的二维结构熵度量图中的结构信息。对于图中的每一个节点模块, 在模块的生成过程中, 任意两个节点模块的合并必须满足熵减原则, 也就是说两个节点合并后的模块熵较合并之前减少的越多, 表示模块合并后的不确定性越低, 其图结构的二维结构熵也越小。这就反映了网络结构中本质的真实结构信息, 即网络结构的结构熵越小, 模块中的节点关联程度越强。

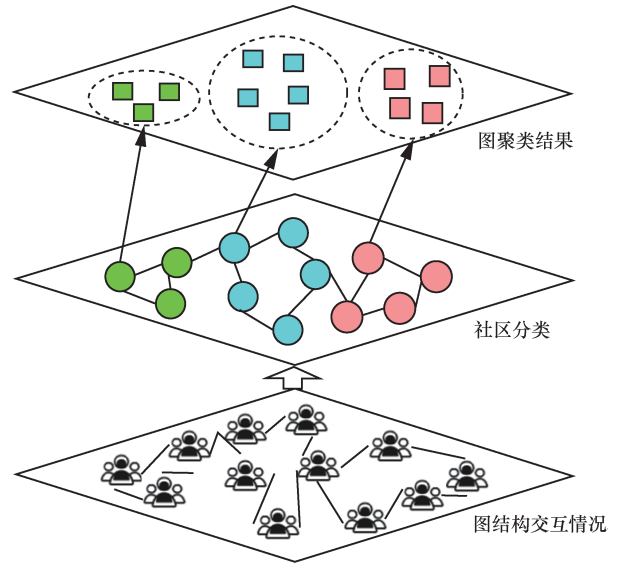


图 1 图结构节点聚类模型
Figure 1 Graph structure node clustering model

3.2 算法设计

为了能从大规模噪声结构中挖掘出图结构的隐私信息, 必须利用结构熵的最小值来解码图结构的真实结构, 从而更好地反映出图结构中任意两个节点的关联程度。在此, 本文提出利用二维结构熵最小值的计算算法和熵减原则的模块分区算法来得出图结构中节点的关联性。

为了解决网络结构中数据的高精度问题需要一个新的信息度量, 能度量嵌入在复杂系统中的信息, 这个信息能区分网络结构中数据的规律和噪声, 从而解码出嵌入在大规模噪声结构中的规律。因此需要设计算法 1。

算法 1 二维结构信息最小值求解算法

输入 一个 n 个顶点 m 条边的无向连通图

$G = (V, E)$

输出 二维结构信息 $H^2(G)$

- 1) $P = \{X_1, X_2, \dots, X_L\}$
- 2) $H^P(G) = 0$
- 3) for each $j \in (1, 2, \dots, L)$ do
- 4) $n_j = |X_j|$
- 5) $H(X_j) = -\sum_{i=1}^{n_j} \frac{d_i^{(j)}}{V_j} \lg \frac{d_i^{(j)}}{V_j}$
- 6) $H^P(G) += \frac{V_j}{2m} H(X_j) - \frac{g_j}{2m} \lg \frac{V_j}{2m}$
- 7) $H^P(G) += \frac{V_j}{2m} H(X_j) - \frac{g_j}{2m} \lg \frac{V_j}{2m}$
- 8) $H^2(G) = \min\{H^P(G)\}$
- 9) end for
- 10) return $H^2(G)$

算法 1 可计算出一个 n 个顶点和 m 条边的无向连通图 G 的二维结构信息，图 G 的二维结构信息就是图所能划分的所有模块熵的最小值。设图的二维结构熵在最小值时所对应节点的划分为分区 P ， $P = \{X_1, X_2, \dots, X_L\}$ 。其中，分区 P 是对所有节点的一个划分， $X_i (i=1, 2, \dots, L)$ 是划分 P 下的一个模块或者说是一个社区，模块是由部分节点组成的集合。然而针对不同的划分结果，计算出图的二维结构信息也不同。只有当图的二维结构信息取值为模块熵的最小值的时候，对应的图中节点分区 P 才能体现出该图的节点关联程度的最优结果。也就是当图的二维结构信息最小时，模块中节点的不确定性最小，节点的关联程度最强。

然而在计算二维结构信息时，为了得到模块熵的最小值，只能通过不同的分区和计算进行比较。然而无法快速地通过计算对比来减少计算成本，从而无法更有效地确定图中的顶点的划分结果是否完善。因此，进一步地利用二维结构信息和熵减原则来有效提高本文获取模块分区的结果。为了能够将图中节点有效分区，得到节点模块之间的关联程度设计了算法 2。

算法 2 网络节点模块分区算法

输入 一个 n 个顶点 m 条边的无向连通图 $G = (V, E)$

输出 网络节点模块

- 1) suppose $P = \{X_1, X_2, \dots, X_L\}$ be a partition of V
- 2) for each $i < j (i, j = 1, 2, \dots, L)$ do
- 3) $\Delta_{i,j}^P(G) = H_{X_i}^P(G) + H_{X_j}^P(G) - H_{X_i \cup X_j}^P(G)$
- 4) when $\Delta_{i,j}^P(G)$ is max, $i_0 = i, j_0 = j$
- 5) if there is no $i < j$ such that $\Delta_{i,j}^P(G) > 0$ then
- 6) return P
- 7) else:
- 8) $\Delta_{i_0, j_0}^P(G) > \Delta_{i,j}^P(G) > 0$
($i \neq j, i, j = 1, \dots, N$)
- 9) set $X = X_{i_0} \cup X_{j_0}$
- 10) set $P = \{X_1, \dots, X_{i_0-1}, X_{i_0+1}, \dots, X_{j_0-1}, X_{j_0+1}, \dots, X_{L-1}, X\}$
- 11) return step 7
- 12) end for

该算法的节点模块划分方法的主要过程如下：假设 $v_1, v_2, \dots, v_n$ 是节点集合 V 中所有按顺序排列出的节点。将 X_i 设置为所有 i 的单元 $\{v_i\}$ ，构成了所有节点集 V 的初始划分模块。如果不存在满足 $i < j$ 这样的 $\Delta_{i,j}^P(G) > 0$ ，则返回分区 P 。否则，有 i_0, j_0 使得 $\Delta_{i_0, j_0}^P(G)$ 是所有 i, j 在 $\Delta_{i,j}^P(G)$ 上的最大值，得到 $X = X_{i_0} \cup X_{j_0}$ 。其中， $\Delta_{i,j}^P(G)$ 表达式如下：

$$\begin{aligned} \Delta_{i,j}^P(G) = & -\frac{V_i}{2m} \sum_{k=1}^{n_i} \frac{d_k^i}{V_i} \lg \frac{d_k^i}{V_i} - \\ & \frac{V_j}{2m} \sum_{k=1}^{n_j} \frac{d_k^j}{V_j} \lg \frac{d_k^j}{V_j} + \frac{V_X}{2m} \sum_{k=1}^{n_i+n_j} \frac{d_k^{i,j}}{V_X} \lg \frac{d_k^{i,j}}{V_X} - \\ & \frac{g_i}{2m} \lg \frac{V_i}{2m} - \frac{g_j}{2m} \lg \frac{V_j}{2m} + \frac{g_X}{2m} \lg \frac{V_X}{2m} = \\ & \frac{1}{2m} [(V_i - g_i) \lg V_i + (V_j - g_j) \lg V_j - \\ & (V_X - g_X) \lg V_X + (g_i + g_j - g_X) \lg 2m] \end{aligned} \quad (4)$$

其中， $V_i = \sum_{k=1}^{n_i} d_k, V_j = \sum_{k=1}^{n_j} d_k, V_X = \sum_{k=1}^{n_i+n_j} d_k^{(i,j)}, X = X_i \cup X_j$ 。

根据二维结构信息最小值求解算法和网络节点模块分区算法可以有效地处理图的结构信息数

据,从大规模噪声结构中解码出图节点之间的关联程度,既不需要改变原始数据也不需要选择参数。这就排除了在归一化或在参数选择中可能产生噪声的影响。

本节使用了一种基于结构信息理论的图节点分区算法。通过节点的划分模块和整体的划分区,用结构熵的最小值来获取图中节点随机游动发生的不确定性,从而更好地反映出图节点之间的关联程度。因此,网络节点分区算法的本质是固定结构不确定性最大的网络节点,合并网络中连接紧密程度高的节点模块。节点的分区算法不仅挖掘出了使整个网络的不确定性最小化的基本结构,而且从模块分区结果中反映出网络中连接紧密程度最高的节点模块。

4 K 维结构信息的聚类算法

网络表现为特定群体中多个个体及它们之间相互联系组成的集合,最适合用图来表示。其中,节点表示个体,边表示个体间的联系。常见的网络图结构模型为简单的图模型 $G=(V,E)$,在图 G 中, $V=\{v_1, v_2, \dots, v_n\}$ 为图中节点集, $E=\{(v_i, v_j) | 1 \leq i, j \leq n, i \neq j\}$ 为边集。图中任何边 (v_i, v_j) 表示节点之间的联系,为结构边。根据实际网络中个体之间联系的紧密程度不同,结构边的权值不同。本文假设给定的是一个无向无权的连通图,结构边具有相等的权值,均为 1。

为了发现图中的簇,把图分割成若干块。每一块是一个簇,使得簇内的节点能够以利益最大化的方式有效地连接,而不同簇内的节点之间以较弱的方式连接。因此,本文通过利用结构熵能检测图结构的真实结构来更有效地对图中节点进行划分。其主要特点都是聚类,将图中所有节点划分为模块子集的过程,使得图中相似度最高的节点形成一个簇,簇内的节点与其他簇内的节点相似度最低。

图聚类问题的实质就是图划分问题,聚类的过程其实就是对图划分过程的优化。优化的目的是使划分的模块间的相似度变小,模块内的节点相似度变大。然而对于如何度量嵌入大规模噪声结构中的整体图结构的信息,仅是二维结构信息是远远不够的。因此,利用 K 维结

构信息来作为图聚类的一种方法,也就是在二维结构信息和熵减原则的基础上,进一步地计算出图结构的最小结构熵,从而更有效地反映图结构的真实结构。为了更加针对性地对图中节点间的关联程度做出有效的挖掘,设计了算法 3。

算法 3 K 维结构信息求解算法

输入 一个 n 个顶点 m 条边的无向连通图 $G=(V,E)$ 。

输出 K 维结构信息 $H^K(G)$

- 1) merge:
- 2) Let $\alpha, \beta \in T, M(\alpha) = \{x_1, x_2, \dots, x_M\}, M(\beta) = \{y_1, y_2, \dots, y_N\}$
- 3) Set $M(\alpha) = \{x_1, x_2, \dots, x_M, y_1, y_1, \dots, y_N\}$
- 4) combine:
- 5) Create a new tree node ε with $M(\varepsilon) = M(\alpha) \cup M(\beta), \varepsilon^- = \delta, \beta^- = \delta, \alpha^- = \delta$
- 6) Let the branch $(\delta - \alpha)$ be $(\varepsilon - \alpha)$ and let the branch $(\delta - \beta)$ be $(\varepsilon - \beta)$
- 7) for each T do
- 8) for each $\alpha_i (i=1, 2, \dots, L)$ do
- 9) $H^T(G; \alpha_0) = 0$
- 10) $H^T(G; \alpha_i) = -\frac{g_{\alpha_i}}{2m} \lg \frac{V_{\alpha_i}}{V_{\alpha_i^-}}$
- 11) $H^T(G) += H^T(G; \alpha_i)$
- 12) end for
- 13) $H^K(G) = \min_T (H^T(G))$
- 14) end for
- 15) return $H^K(G)$

定义 6 (K 维结构信息^[20]) 对于一个无向连通图 $G=(V,E)$, 假设 T 是图 G 的分区树。通过 T 定义图 G 的结构信息如下:

对于任意顶点 $\alpha \in T$, 如果 $\alpha \neq \lambda$, λ 为树 T 的根节点, 那么定义顶点 α 的结构信息为

$$H^T(G; \alpha) = -\frac{g_{\alpha}}{2m} \lg \frac{V_{\alpha}}{V_{\alpha^-}} \quad (5)$$

其中, g_{α} 是从 T_{α} 中的节点到 T_{α} 之外节点的边数, V_{α} 是集合 T_{α} 的体积。

通过分区树 T 定义图 G 的结构信息如下:

$$H^T(G) = \sum_{\alpha \in T, \alpha \neq \lambda} H^T(G; \alpha) \quad (6)$$

图 G 的 K 维结构信息定义如下：

$$H^K(G) = \min_T \{H^T(G)\} \quad (7)$$

其中, T 覆盖图 G 的所有分区树, K 表示 T 的高度。算法 3 计算出图结构的整体 K 维结构信息, K 维结构信息能挖掘嵌入在图结构中的信息, 这个信息能区分图结构的规律和噪声, 从而解码出嵌入在大规模噪声结构中的规律, 进一步决定并解码出该图结构的真实结构。通过在二维结构信息的基础上加入 merge 和 combine 两个算法, 将图结构由高度为 2 的分区树转化为高度为 K 的分区树, 从而进一步地划分图结构中节点的关联程度。因此, 当 K 维结构熵越小时, 越能反映出图结构中单个节点间的关联程度。

聚类分析可以对网络中节点实行分组管理, 使得组内的节点具有非常相似的特性或者需求, 这有利于应用在智能设备中更好地服务用户和加强对用户信息的保护。利用结构信息的最小化将复杂的网络抽象成一个图, 不仅对分析整体网络的结构信息有着重要的帮助, 而且还能对网络中用户的隐私信息进行挖掘和保护有着重要的意义。具体来说, 根据模块熵和熵减原则, 对图中节点进行合并得到不同的模块, 进一步根据图的 K 维结构熵的最小值 K 维结构信息, 最终得到对所有图节点的关联聚类。此时, 结构熵能度量出图在大规模噪声结构中的信息, 所得到的信息也反映了图的真实结构。其中节点的模块分区就表示图结构在挖掘图中关联信息条件下的聚类结果, 便可将图中隐私信息关联程度紧密的节点划分为一个模块, 而模块间节点隐私信息的关联程度远远小于模块内节点的关联程度。

5 实验评估与分析

本节通过实验来评估本文所提出的方法, 并给出详细的实验结果和讨论与其他方法的优缺点。

5.1 实验评估

本文随机生成一个 15 个顶点 60 条边的无向无权的网络模型图。如图 2 所示。

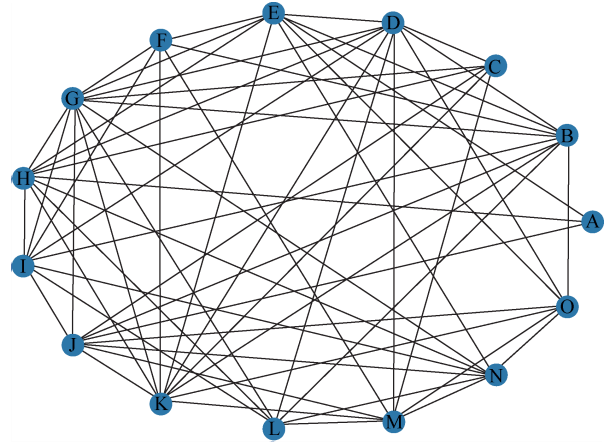


图 2 随机生成图
Figure 2 Randomly generated graph

首先通过算法 1 和算法 2 对图进行模块划分, 计算出图的二维结构信息 $H^2(G) = 3.23$ 。由此可见, 通过二维结构信息和熵减原则算法, 将图划分为大小不同的模块, 并计算不同模块的结构熵 $H^P(G)$, 最终得出图的二维结构信息 $H^2(G) = \min \{H^P(G)\}$, 此时模块 P 的划分结果便是图中节点属性相关程度大的为一个模块。因此, 如图 3 所示, 利用算法 1 和算法 2 将图划分为模块 $P = \{X_1, X_2, X_3, X_4\}$, 且 $X_1 = \{A, E, H\}$, $X_2 = \{B, F, G, I, L\}$, $X_3 = \{C, D, K, M, O\}$, $X_4 = \{J, N\}$ 。

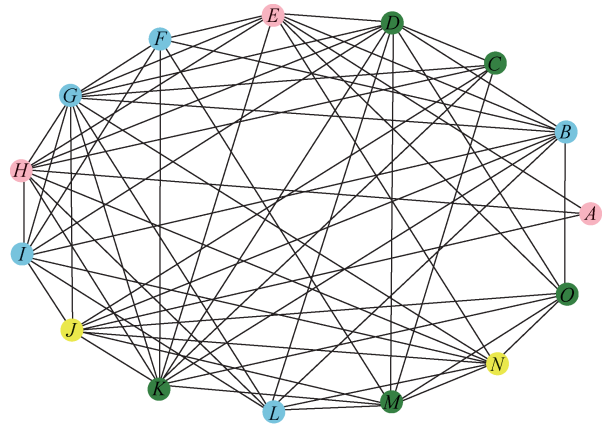


图 3 模块熵的关联节点
Figure 3 Associated node of module entropy

然后将图划分的模块 P 转化为高度 2 的分区树, 引入根节点 λ , 模块 P 中每一个子模块 $X_i (i=1, 2, 3, 4)$ 为树节点, 每个子模块中的节点为该树节点下的叶子节点。二维分区树如图 4 所示。

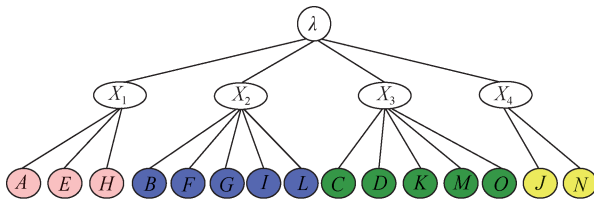


图 4 二维分区树
Figure 4 Two-dimensional partition tree

最后将以模块 P 的最优划分结果通过算法 3 做进一步的节点聚类, 并计算出图的 K 维结构信息 $H^K(G)=3.07$ 。因此, 通过 K 维结构信息聚类算法, 将所有的模块 P 生成的二维分区树转化成高度为 K 的分区树, 并计算出不同的 K 维分区树的结构熵 $H^T(G)$, 最终得出图的 K 维结构信息 $H^K(G)=\min\{H^T(G)\}$, 此时所得到的分区树便是将图中每个节点属性关联程度最大的聚类在一起。如图 5 所示, 从图 3 中可以看出, 利用算法 3 将模块中的节点再一次进行划分, 得到划分结果更为详细。因此, 可以得到在模块 X_1 中节点 A 、 E 关联程度更大, 模块 X_2 和 X_3 中节点进一步得出两两关联程度: $X_2:((G,(B,L)),(F,I))$, $X_3:((K,(C,M)),(D,O))$, 模块 X_4 中的关联程度没有变化。

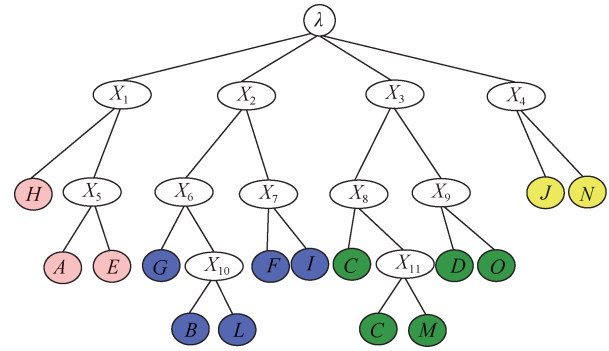


图 5 K 维分区树
Figure 5 K -dimensional partition tree

当图中顶点和边的数量在不断变化时, 图结构的本质信息也会改变, 从而导致图的划分结果也会随之变化, 如表 1 所示。随机生成 4 个不同大小的网络图, 通过二维结构信息的划分方法将图中相似度高的节点分成一个模块, 再通过 K 维结构熵最小值的原理将模块中的节点进一步划分, 最终得到图中所有节点间的关联程度。所以, 本文通过结构熵挖掘出图中节点关联程度大的节点都属于一个模块, 而节点之间关联程度小的则被划分在不同的模块中。因此可以利用结构熵的最小值来体现网络节点的关联程度, 从而挖

表 1 不同网络图的聚类效果
Table 1 Clustering result of different network graphs

顶点 n	边 m	二维结构信息划分结果	K 维结构信息聚类结果
10	21	$X_1:(3,4,8)$	$X_1:(8,(3,4))$
		$X_2:(0,5,6,9)$	$X_2:((0,5),(6,9))$
		$X_3:(1,2,7)$	$X_3:(7,(1,2))$
20	80	$X_1:(0,1,8,11)$	$X_1:((0,11),(1,8))$
		$X_2:(4,5,6,15)$	$X_2:((4,5),(6,15))$
		$X_3:(2,10,14,16,19)$	$X_3:((10,19),(2,(14,16)))$
		$X_4:(3,7,13,17,18)$	$X_4:((3,(13,17)),(7,18))$
		$X_5:(9,12)$	$X_5:(9,12)$
30	274	$X_1:(2,4,6,8,9,10,20,23,24,25)$	$X_1:(((6,24),(9,23)),((2,8),(10,25)),(4,20))$
		$X_2:(0,3,5,11,12,14,19,21,26,27,28,29)$	$X_2:(((10,11),(5,21)),((29,(12,26)),(3,28)),(14,(19,27)))$
		$X_3:(1,7,13,15,16,17,18,22)$	$X_3:(((1,(16,18)),(15,22)),(13,(7,17)))$
40	442	$X_1:(0,1,2,3,5,11,14,16,18,20,21,22,29,34,36,37,39)$	$X_1:(((34,(3,20)),(36,(11,37))),((18,(1,2)),(39,(16,29))),((5,22),(14,(0,21))))$
		$X_2:(4,6,7,9,10,12,13,17,23,24,25,27,32,33,35,38)$	$X_2:(((6,(13,32)),(33,35)),((17,(10,25)),(4,23)),((24,27),(7,12),(9,38)))$
		$X_3:(8,15,19,26,28,30,31)$	$X_3:((15,(19,32)),((8,28),(26,30)))$

掘出网络结构中节点之间的交互信息以及其他相关信息。

5.2 有效性比较与分析

对数据挖掘的研究领域中, 聚类往往可以直接或者间接地描述数据的关联程度, 从而能够有效地分析数据的相关特性。因此聚类分析算法是探索数据分析的关键要素, 其中 K 均值算法因其易于实现、可并行性强和计算成本相对较低而成为最受欢迎的方法。然而正如文献[27]提出的 K 均值对初始条件的高度依赖, 以及在大规模数据集中可能无法很好地扩展。文献[27]提出了一个递归并行逼近 K 均值算法, 该算法在不影响逼近质量的情况下, 能很好地根据问题实例的数量进行缩放。但是该算法在递归次数和计算开销方面都消耗较大, 无法有效地分析社交网络或图等结构化网络图中所嵌入的信息。

划分和层次方法旨在发现球状簇, 对于给定的数据很难发现任意形状的簇。然而为了发现任意形状的簇, 可以把簇看作数据空间中被稀疏区域分开的稠密, 从而可以发现任意非球状的簇。基于密度的噪声应用空间聚类 (DBSCAN, density-based spatial clustering of applications with noise) 是基于密度的聚类算法, 也是主要的聚类范式之一。尽管 DBSCAN 算法已被广泛应用, 但算法本身依然存在一定的缺点。DBSCAN 算法的输入参数包括扫描半径 ε 和阈值 MinPts, 在聚类的过程中需要计算每个数据点的密度。数据点的密度与从该点到其他所有点的距离有关, 使得聚类过程中存在很多冗余的距离计算, 复杂度高, 效率低。因此, 该算法同样不适用于大规模数据和高维数据。Li 等^[28]对 DBSCAN 的聚类原理进行了深入的研究, 发现 DBSCAN 的核心问题是寻找每个点的最近邻, 然后通过三角不等式提出了一种基于最近邻相似的快速密度算法, 并且利用覆盖树对每个并行点的领域进行检索, 可以有效减少聚类过程中距离计算的次数, 提高聚类效率。但是, 该算法的不足在于只基于最近距离来分析每个节点的相似度, 忽略了网络的自组织原理与网络节点所嵌入的结构信息对图聚类的影响。

图聚类主要是针对图的一种聚类方法, 用来

搜索图找出节点间某种相似的特性形成一个簇。

网络结构聚类算法 (SCAN, structural clustering algorithm for networks) 在处理大型网络图时, 其时间复杂性关于边数是线性的, 具有良好的可伸缩性。SCAN 是一种快速高效的集群技术, 用于查找隐藏的社区并隔离网络中的集线器或异常值。然而, 对于大型网络仍然需要相当多的时间。Stovall 等^[29]提出了一种基于计算统一设备架构 (CUDA, computer unified device architecture) 的 SCAN 并行实现 (GPUSCAN, graphic processing unit structural clustering algorithm for networks)。SCAN 的计算步骤经过仔细的重新设计, 通过将 SCAN 转化为一系列高度规则和独立的并发操作, 使其在图形处理单元 (GPU, graphic processing unit) 上高效运行。通过 GPUSCAN, 可以将大型网络或一批不相交的网络卸载到 GPU, 以进行非常快速且高效的结构聚类。GPUSCAN 是在考虑 CUDA 的情况下开发的, 启用 CUDA 的 GPU 充当了并行分布式系统的缩影。由于 GPUSCAN 专注于合并和本地资源访问方法, 因此具有可伸缩性, 并且可以轻松地适应并发或分布式环境。相比之下, SCAN 基于广度优先搜索, 尽管计算效率高, 但往往需要高频率的随机资源访问, 并迅速成为超大型网络的 I/O 绑定。GPUSCAN 当前受 GPU 上可用的设备内存量的限制。虽然可以通过将复杂的计算分布在多个连接的 GPU 上, 从而可以轻松地扩大规模, 但其开销仍是一个难以解决的问题。

因此, 表 2 对以上 3 种聚类方法与本文所提出的基于结构熵的图聚类方法进行了比较。从表 2 可以看出, 文献[27]所提方法的时间复杂度为 $O(nkt)$, 因此, 该方法在计算大规模数据集时, 将消耗大量的计算时间和计算开销。文献[28]所提方法的时间复杂度为 $O(n[\frac{1}{\text{SeedsNum}} \cdot \log_n + \text{MinPts}^2])$, 该算法在聚类过程中具有许多冗余计算, 导致效率低和复杂度较高的问题。文献[29]所提方法的时间复杂度为 $O(n^2)$, 该方法的计算效率与本文方法存在很小的差距, 但该方法受 GPU 上可用的设备内存限制, 将会增加聚类过程中的开销问题。由于图的结构信息能够度量出

表 2 聚类功能对比
Table 2 clustering function comparison

方法	规模	开销	效率	时间复杂度	适合数据
K-means ^[27]	小	大	低	$O(nkt)$	静态
DBSCAN ^[28]	适中	适中	适中	$O(n[\frac{1}{\text{SeedsNum}}\log_n + \text{MinPts}^2])$	静态
GPUSCAN ^[29]	大	大	高	$O(n^2)$	静态
本文方法	大	小	高	$O(n \cdot m)$	动态

图结构中所嵌入的信息, 这个信息量可以更好地反映出图结构的真实结构。因此, 可以在大规模噪声结构中解码出网络结构的真实结构信息的情况下, 利用结构熵更有利于对图结构中节点的关联信息的挖掘, 从而将达到节点属性有效地聚类。因此, 无论是在处理大规模数据集时, 还是在动态复杂的网络结构中, 本文方法在计算规模、计算开销以及计算效率等方面都优于其他方法。

5.3 仿真分析

本文实验环境是一台 CPU 为 Intel (R) Core(TM) i5-7500, 内存为 8 GB, 操作系统为 64 位 Windows 10 的计算机。选用 Python 作为编程语言, 模拟运行熵减原则图划分算法和图节点的聚类算法。如图 6 所示, 本文使用 0~10 000 的不同节点数量运行 6 个实验, 比较了本文所提出的二维结构信息和熵减原则下的图划分算法与 K 维结构信息对图节点的相似聚类算法的执行时间。

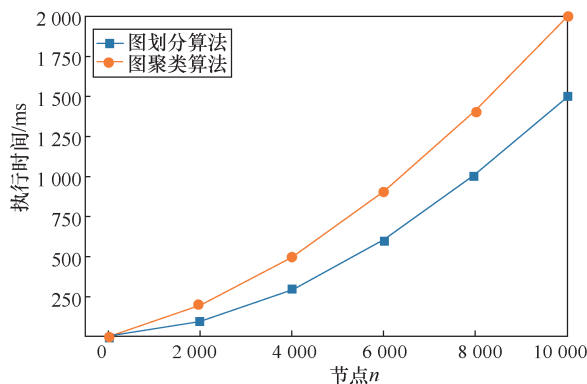


图 6 不同网络节点图划分与图聚类的执行时间
Figure 6 The execution time of graph division and graph clustering of different network nodes

结果表明, 随着节点数量的增大, 算法的执行时间在不断增加。在执行时间上, 图结构的节

点聚类算法的执行时间总是多于图结构的节点划分算法。然而, 图聚类与图划分相比, 聚类整个图结构中节点关联信息的成本更高, 但图聚类算法对图结构中节点聚类的效果更好。因此, 图节点的聚类算法是在图划分算法的基础上进行的。也就是在得到图中节点的模块划分之后, 进一步为了得出节点之间的关联程度, 利用 K 维结构熵的最小值对已划分的模块做进一步的聚类, 从而有效地得出整体图结构中节点之间的关联程度。

4 种节点聚类方案的执行时间如图 7 所示。

结果显示, 本文方法与其他 3 种方法在执行小规模数据时, 执行时间和成本开销相差较小。然而, 在执行大规模数据时, 本文所提出的聚类方法有较短的执行时间。从而反映了本文方法能够有效地执行大规模的动态数据, 缩短聚类时间和减小聚类成本。对于文献[27]方法而言, 随着数据集的增加, K -means 算法的执行时间大大增加, 导致计算开销也随之增加。文献[28]方法, 在数据集中选择的参数均为 $\text{eps} = 0.1$, $\text{minPts} = 10$ 。随着数据集的变化, DBSCAN 算法的执行时间也有较大幅度的上升。相对于文献[29]方法而言, 由于本文不需要计算分布在多个 GPU 上来扩大规模, 因此缩短了计算时间, 同时也存在相对较小的开销成本问题, 极大程度上提高了本文方法对图结构中节点关联信息的聚类效率。

为了实现一个高效的动态聚类方法, 本文利用结构熵的最小值来反映图结构节点的最优聚类效果。图 8 为模拟网络节点的数目在 100~500 变化的情况下进行的节点聚类得到二维结构熵和 K 维结构熵的最小值。从图 8 中可以看出, 随着节点数目的增大, 结构熵的值也增大。然而在相同节点下, K 维结构熵总是小于二维结构熵。因

此更好地说明 K 维结构信息能够有效地反映图结构的真实信息, 从而有利于针对图结构中节点得出最优的聚类结果。同时结构熵可以用来度量嵌入在一个图中的高维信息或深度信息, 并且在度量结构信息的同时解码出原始图结构的实质结构。因此, 可以将结构熵的最小值 K 维结构信息看作是解码图结构中节点真实信息的衡量标准, 有效地反映了图结构中节点之间的聚类效果。

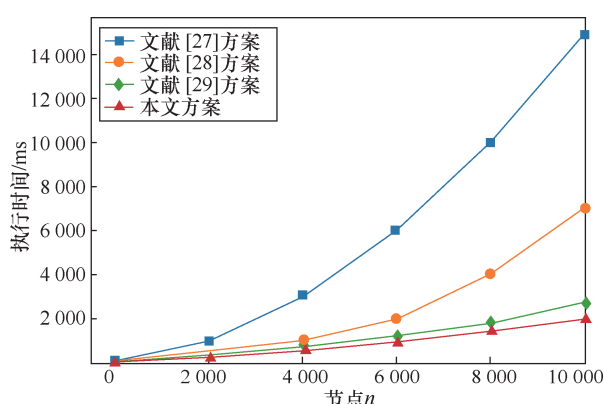


图7 不同网络节点4种聚类方案的执行时间

Figure 7 The execution time of the four clustering schemes for different network nodes

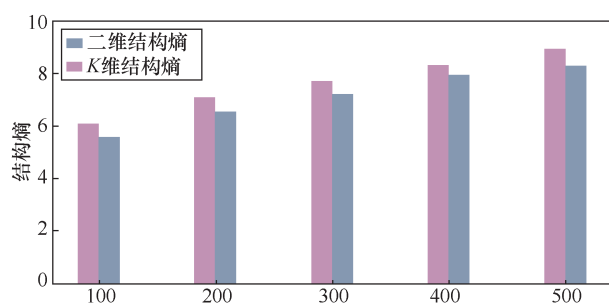


图8 不同网络图的熵值对比

Figure 8 Comparison of entropy values of different network diagrams

6 结束语

本文基于结构熵的聚类方法提出了对图结构中节点的划分。该方法主要是检测出图结构中的真实结构信息, 同时挖掘出图结构中节点之间的关联信息。具体来说, 本文方法不仅适用于大规模的网络图结构, 而且还在保证聚类结果可靠性的同时, 有效地提高了计算效率, 减少了开销。在未来的工作中, 将进一步研究当无向无权图扩展到有向加权图时, 如何挖掘图结构中的关联信息, 以及如何有效地保护挖掘出来的结构信息和

原始图结构中的隐私信息。

参考文献:

- [1] WU D P, ZHANG Z H, WU S E, et al. Biologically inspired resource allocation for network slices in 5G-enabled internet of things[J]. IEEE Internet of Things Journal, 2019, 6(6): 9266-9279.
- [2] WU S E, GUO H Q, XU J H, et al. In-band full duplex wireless communications and networking for IoT devices: Progress, challenges and opportunities[J]. Future Generation Computer Systems, 2019, 92: 705-714.
- [3] BI R W, CHEN Q X, CHEN L, et al. A privacy-preserving personalized service framework through Bayesian game in social IoT[J]. Wireless Communications and Mobile Computing, 2020, 2020: 1-13.
- [4] XIE C, XIAO X Y, HASSAN D K. Data mining and application of social e-commerce users based on big data of internet of things[J]. Journal of Intelligent & Fuzzy Systems, 2020, 39(4): 5171-5181.
- [5] MÜLLE Y, CLIFTON C, BÖHM K. Privacy-integrated graph clustering through differential privacy[C]//Proceedings of the Workshops of the {EDBT/ICDT} 2015 Joint Conference (EDBT/ICDT). 2015: 247-254.
- [6] PAUL A, SUPPAKITPAISARN V, BAFNA M, et al. Improving accuracy of differentially private kronecker social networks via graph clustering[C]//Proceedings of 2020 International Symposium on Networks, Computers and Communications (ISNCC). 2020: 1-6.
- [7] ZHAN Y, PAN H W, XIE X Q, et al. Graph entropy-based clustering algorithm in medical brain image database[J]. Journal of Intelligent & Fuzzy Systems, 2016, 31(2): 1029-1039.
- [8] LIU Y, NG M K, WU S. Multi-domain networks association for biological data using block signed graph clustering[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2020, 17(2): 435-448.
- [9] GUO H B, MA Y, TUSKAN G A, et al. A suggestion of converting protein intrinsic disorder to structural entropy using Shannon's information theory[J]. Entropy, 2019, 21(6): 591.
- [10] KEYVANPOUR M R, MORADI S S. A perturbation method based on singular value decomposition and feature selection for privacy preserving data mining[J]. International Journal of Data Warehousing and Mining, 2014, 10(1): 55-76.
- [11] TEO S G, CAO J, LEE V C. DAG: A general model for privacy-preserving data mining[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 32(1): 40-53.
- [12] AL-SAGGAF Y, ISLAM M Z. Data mining and privacy of social network sites' users: implications of the data mining problem[J]. Science and Engineering Ethics, 2015, 21(4): 941-966.
- [13] GÖRNERUP O, DOKOOHAKI N, HESS A. Privacy-preserving mining of frequent routes in cellular network data[C]//Proceedings of 2015 IEEE Trustcom/BigDataSE/ISPA. 2015: 581-587.
- [14] 周艺华, 张冰, 杨宇光, 等. 基于聚类的社交网络隐私保护方法[J]. 计算机科学, 2019, 46(10): 154-160.

ZHOU Y H, ZHANG B, YANG Y G, et al. Cluster-based social network privacy protection method[J]. Computer Science, 2019,

- 46(10): 154-160.
- [15] YAN S, PAN S R, ZHAO Y H, et al. Towards privacy-preserving data mining in online social networks: distance-grained and item-grained differential privacy[M]//Information Security and Privacy. Cham: Springer, 2016: 141-157.
- [16] TIAN Y P, YAN J, HU J, et al. A privacy preserving model in uncertain graph mining[C]//Proceedings of 2018 International Conference on Networking and Network Applications (NaNA). 2018: 102-106.
- [17] JIANG X L, ZHANG F, ZHANG H. Research on protective mining method for privacy data in network based on apriori algorithm[C]//Proceedings of 2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE). 2019: 1241-1246.
- [18] REZA K, ISLAM M, ESTIVILL-CASTRO V. Privacy preservation of social network users against attribute inference attacks via malicious data mining[C]//Proceedings of the 5th International Conference on Information Systems Security and Privacy. 2019: 412-420.
- [19] ALSHAikh M, ZOHDY M, OLAWOYIN R, et al. Social network analysis and mining: privacy and security on Twitter[C]//Proceedings of 2020 10th Annual Computing and Communication Workshop and Conference (CCWC). 2020: 712-718.
- [20] LI A S, PAN Y C. Structural information and dynamical complexity of networks[J]. IEEE Transactions on Information Theory, 2016, 62(6): 3290-3339.
- [21] LI A S, LI J K, PAN Y C. Discovering natural communities in networks[J]. Physica A: Statistical Mechanics and Its Applications, 2015, 436: 878-896.
- [22] LI A S and PAN Y C. Structure entropy and resistor graphs[J]. CoRR abs, 2018, 1801. 03404.
- [23] LI A S, YIN X C, XU B X, et al. Decoding topologically associating domains with ultralow resolution Hi-C data by graph structural entropy[J]. Nature Communications, 2018, 9(1):3265-3226.
- [24] LIU Y W, LIU J M, ZHANG Z J, et al. REM: from structural entropy to community structure deception[C]// Proceedings of 2019 Annual Conference on Neural Information Processing Systems, NeurIPS. 2019: 12918-12928.
- [25] HUFFMAN D A. A method for the construction of minimum-redundancy codes[J]. Resonance, 2006, 11(2): 91-99.
- [26] 姜火文, 曾国荪, 胡克坤. 一种遗传算法实现的图聚类匿名隐私保护方法[J]. 计算机研究与发展, 2016, 53(10): 2354-2364.
- JIANG H W, ZENG G S, HU K K. An anonymous privacy protection method for graph clustering based on genetic algorithm[J]. Computer Research and Development, 2016, 53(10): 2354-2364.
- [27] CAPÓ M, PÉREZ A, LOZANO J A. An efficient K -means clustering algorithm for tall data[J]. Data Mining and Knowledge Discovery, 2020, 34(3): 776-811.
- [28] LI S S. An improved DBSCAN algorithm based on the neighbor similarity and fast nearest neighbor query[J]. IEEE Access, 2020, 8: 47468-47476.
- [29] STOVALL T R, KOCKARA S, AVCI R. GPUSCAN: GPU-based parallel structural clustering algorithm for networks[J]. IEEE Transactions on Parallel and Distributed Systems, 2015, 26(12): 3381-3393.

[作者简介]



张志颖 (1996-), 男, 贵州铜仁人, 贵州大学硕士生, 主要研究方向为大数据安全与隐私保护、结构信息理论。



田有亮 (1982-), 男, 贵州盘县人, 博士, 贵州大学教授、博士生导师, 主要研究方向为算法博弈论、密码学与安全协议、大数据安全与隐私保护等。