Advances in Applied Mathematics 应用数学进展, 2023, 12(12), 4946-4954
Published Online December 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.1212487

Hans汉斯

基于节点拓扑属性的社团结构划分方法

骆　勇

上海理工大学管理学院，上海

收稿日期：2023年11月5日；录用日期：2023年11月28日；发布日期：2023年12月8日

摘　要

社团结构，作为复杂网络分析中的核心概念之一，指的是那些拥有相似特性和功能的节点所构成的子图结构。社团结构的划分不仅有助于深入挖掘网络内在的信息和特征，而且还有助于揭示复杂网络系统的演化机制。本研究提出了一种创新的社团结构划分算法，其核心围绕网络结构中节点的拓扑属性。在该算法中，我们首先通过节点的接近中心性和介质中心性构建二维决策矩阵。具体来说，接近中心性揭露了节点在网络中与其他节点的亲密程度，而介质中心性则考虑了节点在网络中的桥接作用。二维决策矩阵的构建旨在从网络的广度和深度双视角考察节点的重要性和功能。然后利用基于K-means的迭代策略找到最优的聚类中心。最后通过加权的Dijkstra算法将剩余节点依据就近的聚类中心进行分配完成社团划分。为验证所提出的算法的有效性，在多个实际存在的社会网络上进行了实验，并将其与已有的算法进行了综合比较。实验结果验证了所提算法的性能，特别是在识别具有强社交关系的Karate数据集的社团关系中，本文所提算法在NMI，ARI，Purity三个评价指标中分别达到了81.56%、83.34%、100.00%，提高了社团划分结果的准确性。

关键词

复杂网络，节点拓扑属性，社团结构，社团划分

Community Structure Division Method Based on Node Topology Attributes

Yong Luo

Business School, University of Shanghai for Science & Technology, Shanghai

Received: Nov. 5th, 2023; accepted: Nov. 28th, 2023; published: Dec. 8th, 2023

Abstract

Community structure, as a fundamental concept in the analysis of complex networks, refers to

文章引用: 骆勇. 基于节点拓扑属性的社团结构划分方法[J]. 应用数学进展, 2023, 12(12): 4946-4954.
DOI: 10.12677/aam.2023.1212487

subgraph structures composed of nodes with similar characteristics and functions. The division of community structure not only aids in the in-depth exploration of intrinsic information and features within networks but also helps uncover the evolutionary mechanisms of complex network systems. This study introduces an innovative algorithm for community structure detection, primarily focusing on the topological attributes of nodes within the network. In this algorithm, we initially construct a two-dimensional decision matrix using node closeness centrality and betweenness centrality. Specifically, closeness centrality reveals the intimacy of nodes with respect to other nodes in the network, while betweenness centrality considers the bridging role of nodes within the network. The construction of the two-dimensional decision matrix aims to examine the importance and functionality of nodes from both breadth and depth perspectives of the network. Subsequently, an iterative strategy based on K-means is employed to identify the optimal clustering centers. Finally, the remaining nodes are allocated to clusters based on their proximity to the designated clustering centers using a weighted Dijkstra algorithm, thus completing the community partition. To validate the effectiveness of the proposed algorithm, experiments were conducted on multiple real-world social networks, and comprehensive comparisons were made with existing algorithms. The experimental results have validated the performance of the proposed algorithm, particularly in identifying the community relationships within the Karate dataset, which is characterized by strong social ties. The proposed algorithm achieved 81.56%, 83.34% and 100.00% respectively in the three evaluation indexes of NMI, ARI and Purity. These metrics indicate an enhancement in the accuracy of the community detection results

Keywords

Complex Network, Node Topology Properties, Community Structure, Community Division

1. 引言

在复杂网络研究中，社团结构描述了网络内部的一种特殊组织形态，它由节点之间的紧密连接和相似性形成的子图或节点集合组成。这种结构是网络自发形成的，代表了网络中节点的内在关联和互动方式[1] [2] [3]。其特点是同一社团内的节点更相似，而不同社团之间则差异较大。很多系统都可以看作是网络，例如社交关系[4] [5]、知识图谱[6] [7] [8]等。这些网络中呈现出明显的模块化特征和层次化结构，其中社团结构是一个重要的组成部分。如何有效地发现和刻画社团结构，对于深入理解网络本质、揭示系统的规律和特点具有重要意义。

社团划分不仅可以帮助我们了解网络的特征，还能够预测未来网络的演化趋势，发现潜在隐秘节点和关系。社团划分算法的主要任务是将网络节点划分成若干个子社团，以确保社团内部的紧密性和社团之间的分隔性。这一过程旨在揭示网络的内在组织结构，有助于更深入地理解网络的复杂性。社团划分算法可以分为三个主要范畴，包括基于节点相似性的聚类方法、以网络模块度为基础的分析方法，以及运用随机游走策略的方法。这些方法各自侧重不同的网络结构特征和信息挖掘方式，为研究复杂网络中的社团结构提供了多样的工具和角度。同时，随着深度学习技术的飞速发展，神经网络的强大潜力被越来越多地应用于社团划分问题的解决，提供了新的创新方法和技术路径。这些算法借助神经网络的特征学习和表示学习能力，以更深入、更准确的方式来探索网络中的社团结构，为复杂网络研究提供了令人兴奋的前景。

基于聚类的方法依靠节点或社团的位置信息以及他们在网络中的关注度来进行社团结构的划分。Newman 利用模块度原理和模块度增量计算，通过逐步合并节点或社团进行社团划分，提出了 Fast Newman (FN)算法[1]。由 Raghavan 等人[9]提出的标记传播算法(Label Propagation Community Detection Algorithm, LPC)，是一种以节点相似性为基础的社团检测算法。该方法的核心思想在于将具有相似特性的节点赋予相同的标签，通过标签在网络中的传播来揭示社团结构。此方法为研究人员提供了一种有力工具，以更全面地理解网络中的社团组织和关联。Zhang 等[10]提出了一种基于真实连接概念的重叠社团检测方法，它通过预处理原始网络得到“真实连接”图，并利用凝聚式策略不断地合并相似度高的社团，最终形成重叠社团结构。基于模块度(modularity)的方法以一种优化固定的数量的模块度量为目标函数，并尝试将网络分成多个社区，其中每个社区都有高内部连通性和低交叉边数。Blondel 等[11]提出了 Louvain 算法。该算法将节点逐步聚类到社区中，并优化整个网络的模块度。Zhe 等[12]提出了基于贪心策略的最大化模块度算法，利用网络拓扑和属性信息图对网络进行划分，以获得网络的社团结构。该算法能够高效处理大规模复杂网络数据。Zhu 等人[13]引入了一种创新性的社团检测方法，采用了基于 k-plex 的策略。该方法的核心思路是通过生成 k-plex，将其作为社团的种子，然后运用模块度优化算法来智能分配剩余节点，形成社团结构。基于随机游走的社团发现方法是利用随机游走模拟在网络中漫步而创建的节点序列，对网络结构进行分析和建模来检测社团。Perozzi 等人[14]在其研究中提出了 DeepWalk 算法，该算法使用基于随机游走的节点表征策略来生成节点序列，随后应用 Skip-gram 模型将这些节点序列转换为低维向量表示。这一方法的独特之处在于它能够将网络中的节点转化为连续向量空间中的点，从而为节点之间的相似性和关联性提供了新的度量方式。Grover 等人[15]提出 Node2Vec 算法，通过定义 node-to-node 转移概率分布来实现对网络节点的高质量嵌入表示。与 DeepWalk 类似，Node2Vec 也利用随机游走的策略捕获网络局部结构，但同时采用深度优先搜索和广度优先搜索来平衡序列上下文和距离信息，建立更完整、更贴近真实网络的节点向量表达。

本文提出了一种基于聚类的方法，首先计算节点的接近中心性和介质中心性构建二维特征矩阵，其次利用基于 K-means 的迭代策略来找到聚类中心，并通过加权的 Dijkstra 算法将其余节点进行分配，直至所有节点完成社团划分。在几个真实网络上的实验结果证明了所提出算法的有效性。

2. 相关理论

2.1. 图的定义

给定网络模型 $G=(V,E,A)$，其中 V 代表节点的集合，表示为 $V=\{v_1,v_2,\cdots,v_n\}$；E 代表节点之间连接的边的集合，表示为 $E=\{e_1,e_2,\cdots,e_m\}$；A 代表网络的邻接矩阵，记作 $A=(a_{ij})_{n\times n}$，其中，当节点 v_i 和节点 v_j 之间存在连接时，$a_{ij}=1$；反之，则 $a_{ij}=0$。此外，w_{ij} 表示边 e_{ij} 的权重。

2.2. 网络拓扑中心性

2.2.1. 接近中心性

接近中心性，通常表示为 Closeness Centrality (CC) [16]，评估了一个节点与其他节点之间的平均最短路径距离。这一度量方法关注的是节点在网络中的紧密性，即节点与其他节点之间的距离越短，其接近中心性值越高。其公式为：

$$CC_i=\frac{1}{d_i} \tag{1}$$

$$d_i=\frac{1}{N-1}\sum_{j=1}^{N}d_{ij} \tag{2}$$

其中，d_i 表示节点 i 到其余各点的平均距离，N 表示所有节点的总数，d_{ij} 表示连接节点 i，j 的最短距离。在这里，接近中心性的概念用来描述一个节点与其他节点之间的联系程度，即这个节点与其他节点沟通的便捷程度。聚类中心作为网络中的中心，与其他节点的联系应该比普通节点要强。

2.2.2. 介质中心性

介质中心性，通常用 Betweenness Centrality (BC)表示[17]，用于评估网络中的节点在不同节点对之间的最短路径上出现的频率。这一概念有助于我们识别那些在网络中连接不同部分、促进信息流动的节点，从而更好地理解网络的结构和功能。介数中心性的计算公式为：

$$BC_i = \frac{2}{(N-1)(N-2)} \sum_{s \neq i \neq t} \frac{g_{st}^i}{g_{st}} \tag{3}$$

其中，N 表示节点的数量，g_{st} 表示由节点 s 到达节点 t 的所有最短路径的条数，g_{st}^i 则表示在满足 g_{st} 条件的路径中，恰好路径节点 i 的路径条数。聚类中心作为社团内的中心，起到其他节点之间的连接沟通作用。

2.3. Dijkstra 最短路径算法

Dijkstra 算法[18]，被广泛认为是一种基于局部最优选择策略的最短路径搜索算法。该算法的核心思想是逐步探索并更新到达各个节点的最短路径，以便有效地确定源节点与其他节点之间的最短距离。具体算法步骤如下：

第一步：初始化节点，将起点 s 的距离 dist[s]赋值为 0，把除 s 外的其它顶点的已知最短距离 dist 数组置为一个很大的数(表示暂时不可达)。再定义一个 S 集合，存放已经确定了最短路径的节点。

第二步：迭代，找到当前未被访问的、距离最近的节点 v(即在 dist 中最小的)，加入 S 集合；并通运算更新与 v 相邻节点到起点 s 的距离，直到所有顶点都被加到 S 集合中。算法一给出了 Dijkstra 最短路径的算法流程。

算法一带权值的 Dijkstra 最短路径算法

```
输入：带权图 Graph；源节点 Source；目标节点 Target
输出：由源节点 Source 出发达到目标节点 Target 的最小步长
d ← {}
for each node v in Graph：
    d[v] ← infinity
d[source] ← 0
p ← {}
visited ← {}
while some nodes are unvisited：
    u ←a node having the minimum value of d[u] among the ramian nodes that have not been visited.
    if u == target：
        break        // 如果当前节点是目标节点，提前终止算法
    visited.add(u)
```

Continued

```
        for each neighbor v of u:
            if v is in visited:
                continue
            alt ← d[u] + length(u, v)
            if alt < d[v]:
                d[v] ← alt
                p[v] ← u
        remove u from unvisited nodes
    path ← [target]
    while path[0] != source:
        path.insert(0, p[path[0]])
    return path
```

3. 基于节点拓扑属性的社团结构划分方

3.1. 基于迭代策略的聚类中心选择

根据网络图计算每个节点的接近中心性和介质中心性 CC 和 BC 值。对 $CC=\{CC_1,CC_2,\cdots,CC_n\}$ 和 $BC=\{BC_1,BC_2,\cdots,BC_n\}$ 进行 Max-Min 归一化处理，具体公式为：

$$CC_i^*=\frac{CC_i-\min(CC)}{\max(CC)-\min(CC)} \tag{4}$$

$$BC_i^*=\frac{BC_i-\min(BC)}{\max(BC)-\min(BC)} \tag{5}$$

值得注意的是，接近中心性 CC 和介质中心性 BC 都很小的节点将没有机会被选为聚类中心。具体来说，我们要保留前 80%节点的平均值，即对所有节点按照 CC 和 BC 的值从大到小排序，选取前 80%节点的平均值作为阈值。将所有 CC 和 BC 值都小于该阈值的节点丢弃。剩余节点组成的集合记为 V_c，并将 V_c 的 CC 与 BC 的乘积记为 ρ_i，其公式为：

$$\rho_i=CC_i^*\cdot BC_i^* \tag{6}$$

根据 ρ_i 值对节点进行 k-means 聚类，其中 $k=2$，即分为两组。根据启发式方法，在初始的聚类过程中，将 ρ_i 较高的节点设置为初始聚类中心(H-group)，ρ_i 较低的节点设置为被划分的一个集合(L-group)。然后，使用一个标准来确定新 H-group 中的节点是否被选为聚类中心，该标准涵盖了计算 L-group 中的最大值与 H-group 中的最小值，然后取其平均值。这一度量方式旨在从平均的角度审视两个不同群体之间的关联程度。其公式如下：

$$\frac{1}{|V_c|}\sum_{v_i\in V_c}\rho_i\geq\mu\left(\min_{v_i\in\text{H-group}}\rho_i+\max_{v_i\in\text{L-group}}\rho_i\right) \tag{7}$$

其中，$|V_c|$ 表示集合中节点的数量，$\min_{v_i\in\text{H-group}}\rho_i$ 表示 H-group 中最小的 ρ_i 值，$\max_{v_i\in\text{L-group}}\rho_i$ 表示 L-group 中最大的 ρ_i 值，μ 为平衡参数，根据实验最佳值为 0.5。如果满足不等式，则说明 L-group 中已经没有更多的节点应该被划分到 H-group 中作为新的聚类中心了，递归过程结束；否则，从 L-group 将 ρ_i 值最大

的节点分配给 H-group 形成新的 H-group′，继续进行 k-means 聚类的迭代过程。重复执行直至达到所需的收敛状态，即不再有节点被划分到 H-group 中。将最终的 H-group 节点集合视为网络的聚类中心。

3.2. 节点聚类分配

一旦确定了聚类中心，接下来的任务是使用 Dijkstra 最短路径算法将尚未分配的节点分配到它们最近的聚类中心所在的集群中。值得注意的是，这一聚类分配过程是在单一步骤内完成的。这意味着在确定聚类中心后，节点的分配过程更加高效，省去了多次迭代的需要，从而加速了整个聚类过程。不仅能够在实际应用中节省时间和计算资源，还有助于减少潜在的计算复杂性。

4. 实验与结果分析

选取 3 个已知社团标签的真实网络数据集 Karate [19]、Football [20]和 Polbooks [11]来进行实验，表 1 列出了真实网络的先关统计数据。

Table 1. Statistical analysis of real network datasets
表 1. 真实网络数据集的统计分

真实网络	类型	节点数	边数	社团数
Karate	加权	34	78	2
Football	无权	115	613	12
Polbooks	无权	105	441	3

本文应用标准互信息(Normalized Mutual Information, NMI) [21]、兰德指数(Adjusted Rand Index, ARI) [22]和纯度分数(Purity) [23]这三个评价指标，对本文所提出的算法本文算法与 LPC [9]、GN (Girvan-Newman Algorithm) [24]方法进行社团划分结果比照实验。

图 1 为三个真实网络的二维决策图，图中横纵坐标分别表示节点的拓扑结构属性接近中心性(CC)和介质中心性(BC)的标准化后的数值。此外，图中的颜色条表示节点聚类中心的可能性。节点颜色越接近红色，则越有可能成为聚类中心。最后使用基于 K-means 的迭代策略进行量化找到具体的聚类中心节点。

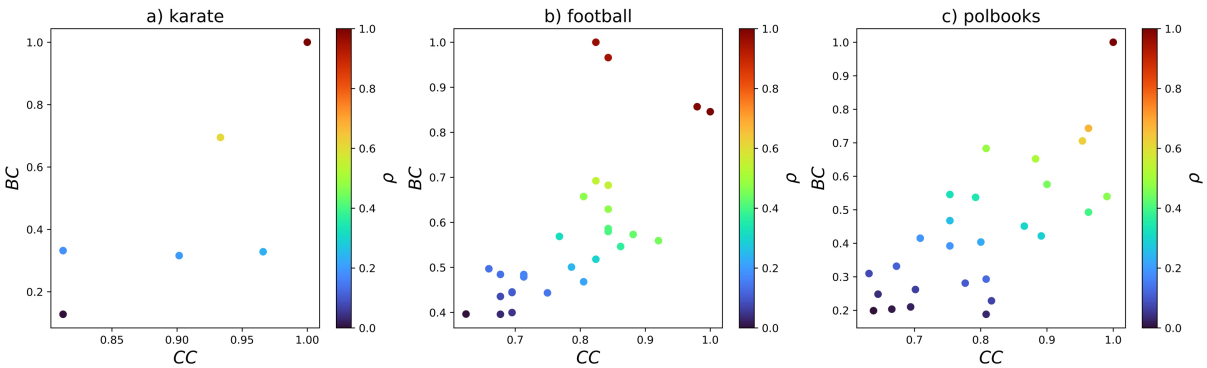

Figure 1. Two dimensional decision graphs of three real networks
图 1. 三个真实网络的二维决策图

表 2 展示了不同算法的标准互信息对比，本文的算法本文算法在三个真实网络中的结果相对较好，特别是在 Karate 真实网络中，本文算法达到性能最佳。表 3 和表 4 分别为 ARI 和 Purity 评价指标下不同

算法的对比，表现出同样的结果。从图中可以看出本文所提方法在 Karate 的性能表现最佳，在 NMI、ARI、Purity 三个评价指标中分别达到了 81.56%、83.34%、100.00%，在 Football 和 Polbooks 数据集中，并没有表现出最好的性能，但也并不是最差的。

Table 2. Comparison of NMI among various algorithms
表 2. 不同算法的标准互信息 NMI 对比

算法	LPC 算法	GN 算法	本文算法
Karate	0.363599	0.732378	0.835574
Football	0.869727	0.359228	0.484394
Polbooks	0.534105	0.548914	0.382228

Table 3. Comparison of ARI among various algorithms
表 3. 不同算法的兰德指数 ARI 对比

算法	LPC 算法	GN 算法	本文算法
Karate	0.383312	0.771725	0.813397
Football	0.750996	0.140016	0.196887
Polbooks	0.594237	0.630523	0.295418

Table 4. Comparison of Purity among various algorithms
表 4. 不同算法的纯度分数 Purity 对比

算法	LPC 算法	GN 算法	本文算法
Karate	0.852941	0.941176	1.000000
Football	0.869565	0.217391	0.486957
Polbooks	0.847619	0.838095	0.857143

分析其原因，考虑数据集之间的本质差异。Karate 数据集源自于一个空手道俱乐部，反映的是俱乐部成员之间的社交互动和联系。这种类型的网络往往表现出较为明显的社交结构特征，其中社团结构的形成与个人间的日常交流密切相关。因此，此类网络中的社团划分通常较为清晰，且社团内部的联系比较紧密。相比之下，Football 数据集代表的是 12 支足球队之间的比赛网络。这种网络的连接通常基于比赛关系而非日常社交互动，因此其社团划分可能不如基于社交互动的网络那么明显。此外，足球队之间的比赛并不一定能完全反映出队伍之间的社会关系或者社区归属感，而是更多受赛季赛程和比赛结果的影响。Polbooks 数据集则涉及政治图书之间的共引关系，映射的是图书在政治倾向上的二分性。综上所述，本文提出的算法在分析具有强社交结构特征的网络时表现得更为优秀。

此外，在基于迭代策略的聚类中心选择阶段，我们还对平衡参数进行了详尽的敏感性分析来确定最佳参数值，以确保聚类过程的准确性和效果。在图 2 中，横坐标对应着不同参数 μ 的取值，而纵坐标则展示了各种评价指标的具体数值，包括兰德指数 ARI、标准互信息 NMI 以及纯度分数 Purity。这一敏感性分析的目的在于确定最佳参数值，以确保聚类过程的准确性和效果。从图 2 中可以看出，当 $\mu = 0.5$ 时，在不同真实网络上的性能总体达到最佳。

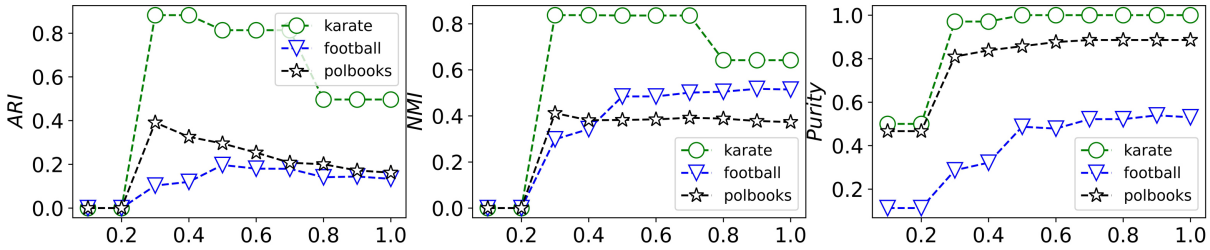

Figure 2. Sensitivity analysis for different μ parameter values
图 2. 不同 μ 参数值的敏感性分析

5. 结论

本文提出了一种基于节点拓扑属性的社团结构划分算法(记为本文算法)，首先计算节点的接近中心性和介质中心性来构建二维决策图，然后使用基于 k-means 聚类的迭代策略选择聚类中心，最后使用 Dijkstra 算法将剩余节点分配给就近的聚类中心所在的社团，直至所有节点完成分配实现网络的社团划分。在三个真实社会网络中，将所提算法与 LPC 算法和 GN 算法进行综合性能对比，进一步确认了本文算法在不同社会网络下的可行性和有效性。实验结果显著地展示了本文提出的算法在识别 Karate 数据集这类具有强烈社交关系的社群结构方面的卓越性能。针对 Karate 数据集，本文所提算法在 NMI、ARI、Purity 三个评价指标中分别达到了 81.56%、83.34%、100.00%，显著提高了社团划分结果的准确性。

参考文献

[1] Newman, M.E.J. (2006) Finding Community Structure in Networks Using the Eigenvectors of Matrices. *Physical Review E*, **74**, Article ID: 036104. https://doi.org/10.1103/PhysRevE.74.036104

[2] 李欢, 莫欣岳. 复杂网络重叠社团检测算法研究综述[J]. 传感器与微系统, 2017, 36(1): 1-4. https://doi.org/10.13873/J.1000-9787(2017)01-0001-04

[3] Ferligoj, A. and Batagelj, V. (1992) Direct Multicriteria Clustering Algorithms. *Journal of Classification*, **9**, 43-61. https://doi.org/10.1007/BF02618467

[4] 蒋忠元, 陈贤宇, 马建峰. 社交网络中的社团隐私研究综述[J]. 网络与信息安全学报, 2021, 7(2): 10-21.

[5] Souravlas, S., Sifaleras, A., Tsintogianni, M. and Katsavounis, S. (2021) A Classification of Community Detection Methods in Social Networks: A Survey. *International Journal of General Systems*, **50**, 63-91. https://doi.org/10.1080/03081079.2020.1863394

[6] 方倩, 窦永香, 王帮金. 基于 Web of Science 的社会化媒体环境下社区发现研究综述[J]. 中文信息学报, 2017, 31(3): 1-8.

[7] 黄琳凯. 基于知识图谱的 Web 信息关联网络分析与主题社区发现[J]. 无线互联科技, 2018, 15(10): 23-24.

[8] 向卓元, 利朝香. 知识图谱视域下国内外社区发现研究动态与热点分析[J]. 图书馆研究与工作, 2019(1): 52-57.

[9] Raghavan, U.N., Albert, R. and Kumara, S. (2007) Near Linear Time Algorithm to Detect Community Structures in Large-Scale Networks. *Physical Review E*, **76**, Article ID: 036106. https://doi.org/10.1103/PhysRevE.76.036106

[10] Zhang, Y., Zhang, Y., Chen, Q., Ai, Z. and Gong, Z. (2017) True-Link Clustering through Signaling Process and Sub-community Merge in Overlapping Community Detection. *Neural Computing & Applications*, **30**, 3613-3621. https://doi.org/10.1007/s00521-017-2946-3

[11] Blondel, V.D., Guillaume, J., Lambiotte, R. and Lefebvre, E. (2008) Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics*: *Theory and Experiment*, No. 10, P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

[12] Zhe, C., Sun, A. and Xiao, X. (2019) Community Detection on Large Complex Attribute Network. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage, 4-8 August 2019, 2041-2049. https://doi.org/10.1145/3292500.3330721

[13] Zhu, J., Chen, B. and Zeng, Y. (2020) Community Detection Based on Modularity and K-Plexes. *Information Sciences*, **513**, 127-142. https://doi.org/10.1016/j.ins.2019.10.076

[14] Perozzi, B., Al-Rfou, R. and Skiena, S. (2014) DeepWalk: Online Learning of Social Representations. *Proceedings of the* 20*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, 24-27 August 2014, 701-710. https://doi.org/10.1145/2623330.2623732

[15] Grover, A. and Leskovec, J. (2016) Node2vec: Scalable Feature Learning for Networks. *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 855-864. https://doi.org/10.1145/2939672.2939754

[16] Sabidussi, G. (1966) The Centrality Index of a Graph. *Psychometrika*, **31**, 581-603. https://doi.org/10.1007/BF02289527

[17] Freeman, L.C. (1977) A Set of Measures of Centrality Based on Betweenness. *Sociometry*, **40**, 35-41. https://doi.org/10.2307/3033543

[18] Dijkstra, E.W. (1959) A Note on Two Problems in Connexion with Graphs. *Numerische Mathematik*, **1**, 269-271. https://doi.org/10.1007/BF01386390

[19] Zachary, W.W. (1977) An Information Flow Model for Conflict and Fission in Small Groups. *Journal of Anthropological Research*, **33**, 452-473. https://doi.org/10.1086/jar.33.4.3629752

[20] Girvan, M. and Newman, M.E.J. (2002) Community Structure in Social and Biological Networks. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 7821-7826. https://doi.org/10.1073/pnas.122653799

[21] Church, K.W. and Hanks, P. (1989) Word Association Norms, Mutual Information, and Lexicography. *Computational linguistics*, **16**, 22-29. https://doi.org/10.3115/981623.981633

[22] Rand, W.M. (1971) Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, **66**, 846-850. https://doi.org/10.1080/01621459.1971.10482356

[23] Hennig, C. (2007) Cluster-Wise Assessment of Cluster Stability. *Computational Statistics & Data Analysis*, **52**, 258-271. https://doi.org/10.1016/j.csda.2006.11.025

[24] Newman, M.E. and Girvan, M. (2004) Finding and Evaluating Community Structure in Networks. *Physical Review E*, **69**, Article ID: 026113. https://doi.org/10.1103/PhysRevE.69.026113