语义增强的大规模多元图简化可视分析方法

刘玉华[1)] 张汝敏[1)] 张靖宇[1)] 高 峰[1)] 高 远[1)] 周志光[1),2)]

[1)](浙江财经大学信息管理与工程学院 杭州 310018)

[2)](浙江大学 CAD&CG 国家重点实验室 杭州 310058)

摘 要 网络图可视化可以有效展示网络节点之间的连接关系,广泛应用于诸多领域,如社交网络、知识图谱、生物基因网络等.随着网络数据规模的不断增加,如何简化表达大规模网络图结构已成为图可视化领域中的研究热点.经典的网络图简化可视化方法主要包括图采样、边绑定和图聚类等技术,在减少大量点线交叉造成的视觉紊乱的基础上,提高用户对大规模网络结构的探索和认知效率.然而,上述方法主要侧重于网络图中的拓扑结构,却较少考虑和利用多元图节点的多维属性特征,难以有效提取和表达语义信息,从而无法帮助用户理解大规模多元网络的拓扑结构与多维属性之间的内在关联,为大规模多元图的认知和理解带来困难.因此,本文提出一种语义增强的大规模多元图简化可视分析方法,首先在基于模块度的图聚类算法基础上提取出网络图的层次结构;其次通过多维属性信息熵的计算和比较分析,对网络层次结构进行自适应划分,筛选出具有最优属性聚集特征的社团;进而设计交互便捷的多个关联视图来展示社团之间的拓扑结构、层次关系和属性分布,从不同角度帮助用户分析多维属性在社团形成和网络演化中的作用.大量实验结果表明,本文方法能够有效简化大规模多元图的视觉表达,可以快速分析不同应用领域大规模多元图的关联结构与语义构成,具有较强的实用性.

关键词 网络图可视化;简化表达;多维属性;语义;拓扑结构

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2020.00136

Semantic-Enhanced Visual Abstraction of Large-Scale Multivariate Graphs

LIU Yu-Hua[1)] ZHANG Ru-Min[1)] ZHANG Jing-Yu[1)] GAO Feng[1)] GAO Yuan[1)] ZHOU Zhi-Guang[1),2)]

[1)](*School of Information, Zhejiang University of Finance and Economics, Hangzhou* 310018)

[2)](*State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou* 310058)

Abstract As an effective tool to show the relationships between network nodes, graph drawing has been widely used in various fields, such as the social network, knowledge map, co-citation network and biological gene network. With the growth of scale of network graphs, it has been a research focus in the field of graph visualization to simplify the presentations of large-scale graphs. A variety of simplification visualization methods have been proposed to reduce the visual clutters caused by a large number of crossovers between nodes and edges, such as the graph sampling, edge bundling and graph clustering. However, these methods mainly focus on the topology of network, without taking the multi-dimensional attribute information associated with network nodes into consideration. It is always difficult to extract meaningful semantic information to help users understand the inherent association between topologies and multi-dimensional attributes. At present, both addressing the topology-based tasks and attribute-based tasks at the same time is still an open research problem. To solve the difficulty of visualizing such multivariate networks

收稿日期:2018-11-12;在线出版日期:2019-09-09.本课题得到国家自然科学基金(61872314,61802339)、教育部人文社会科学研究项目(18YJC910017)、浙江省自然科学基金(LY18F020024)、浙江省高校重大人文社科攻关计划项目(2018QN021)、浙江大学 CAD&CG 国家重点实验室开放课题(A1806)资助.**刘玉华**,博士,讲师,主要研究方向为数据可视化与可视分析.E-mail:liuyuhua@zufe.edu.cn.**张汝敏**,本科,主要研究方向为数据可视化与可视分析.**张靖宇**,本科,主要研究方向为数据可视化与可视分析.**高 峰**,本科,主要研究方向为数据可视化与可视分析.**高 远**,本科,主要研究方向为数据可视化与可视分析.**周志光**(通信作者),博士,副教授,中国计算机学会(CCF)会员,主要研究方向为数据可视化与可视分析.E-mail:zhgzhou1983@163.com.

arising from two conflicting goals: visualizing topology and visualizing node attributes, we propose a semantic-enhanced method in this paper for the visual abstraction of large-scale multivariate graphs. First, the hierarchal structure is extracted from the original network dataset by means of a modularity-based graph clustering to find such communities including a set of highly interconnected nodes. Then, a graph cut scheme is designed to restructure the hierarchy of networks by computing and comparing the information entropies of multi-dimensional attributes of communities, so that the communities with the best aggregation characteristic on one attribute at different scale will be well preserved and displayed in the resultant simplified visualizations. Thus, these communities both showing obvious aggregations on the topology and semantics are selected successfully. Next, several coordinated views are designed in a visual analytic system, enabling users to further analyze the roles of different attributes in the network evolution and community formation from different aspects. For example, a multi-scale force-directed layout is applied to create the abstract visualizations of a large-scale multivariate graph to show the community-centric topologies. Two kinds of visualizations namely non-nested and nested layouts are both provided to solve different analysis tasks for users. While a tree view is provided to help users identify the hierarchies among these communities, which are designed as a pie chart to compare the aggregation properties of multi-dimensional attributes. In addition, an attribute sankey view is designed to explore the difference between the semantic-enhanced community set formed by different single attribute, allowing users to observe the homophily's effects varying with different attributes. Three alternative glyphs are designed for the blocks in the attribute sankey view to show the macro and micro attribute information. Finally, a set of experimental results shows that our method can effectively simplify the visualizations of large-scale multivariate graphs in different fields such as the microblog retweeting and paper citation, and help users explore the topological and semantic structures of graphs. The utility of our approach is also demonstrated by domain experts through an in-depth case study.

Keywords　graph drawing; visual abstraction; multi-dimensional attributes; semantics; topology

1　引　言

网络图经常用于抽象表达实体与实体之间的关系. 例如,社交网络(social network)中,节点代表社交媒体用户,边代表用户之间的好友关系;引文网络(citation network)中,节点代表论文,边代表论文之间的引用关系;在蛋白质网络(protein-protein interaction networks)中,节点代表蛋白质,边代表两个蛋白质在表达生物功能时的相互作用. 因此,研究者们广泛利用网络图分析各自领域(包括社会学、生物学、交通地理学等)的实体关系结构,探索和洞察网络的结构特性. 网络图可视化充分利用了人类的视觉感知能力和信息处理能力,以其直观、易于理解的优点逐渐成为网络图分析的重要手段,其中以点线链接式的网络图可视化方法为代表,应用最为广泛.

随着数据规模的不断增加,节点的数量越来越多,关系结构越来越复杂. 传统的点线链接图出现大量的点线交叉和重叠,给用户造成严重的视觉紊乱和混淆,增加了网络结构的理解难度. 因此,大量研究提出了多种大规模网络图简化方法,如图采样、边绑定和图聚类等技术,一定程度上降低了用户对大规模网络图的认知负担,提高了探索效率. 大规模网络图不仅具有复杂的拓扑结构,而且具有多元属性描述. 例如,社交网络中,每个用户节点都有丰富的个人资料描述,包括性别、年龄、职业、地域等;引文网络中,每个论文节点都具有相应的发表时间、研究主题、引用次数、所属会议或者期刊等信息. 多元属性信息能够从不同方面反映实体之间语义关联结构的内聚特性,以及更高层次的聚类之间的耦合特性,已在诸多领域产生了重要应用,例如:McPherson 等人[1]发现种族和教育在日常人际关系构成中扮演了重要角色,证实了"物以类聚、人以群分"这一现象

的普遍存在；Yavas 等人[2]通过研究发现在社交网络中拥有相似兴趣爱好的好友更易互动和传播信息，这对我们如何借助社交媒体进行营销宣传等提供了理论依据；生物学家还可通过已知交互作用的蛋白质及其基因表达数据来推测其他蛋白质在表达某一功能时的交互作用[3]，这对研究生物组织、疾病机制和发现新药靶点等具有重要意义. 然而，传统的大规模网络图简化可视化方法，难以充分利用网络节点的多维属性信息、提取有效的语义知识，从而无法帮助用户更好地理解拓扑结构与多维属性的语义关联，不能深入探索网络特性等性质.

因此，本文提出一种语义增强的大规模图简化可视分析方法，首先，在基于模块度的图聚类算法基础上，提取出网络图的层次结构；其次，通过多维属性信息熵的计算和比较分析进行网络层次结构的自适应划分，迭代式地筛选出在具有最优属性聚集特征的社团；再次设计可交互的多个关联视图来展示社团之间的拓扑结构、层次关系和属性分布，从不同角度帮助用户分析多维属性在社团形成和网络演化中的作用；最后集成便捷的用户交互模式，设计与研制语义增强的大规模图简化可视分析系统，在维持网络图整体拓扑结构的基础上，结合多维属性信息构造多尺度的社团关系布局，在简化网络图的同时，帮助用户快速探索网络图的语义构成.

本文的主要贡献如下：

(1) 考虑多维属性信息，结合信息熵和图聚类提出了网络层次结构的自适应划分算法，查找出在拓扑结构和多维属性上具有明显聚集特征的社团.

(2) 针对多尺度的社团集，优化改进传统的力引导图、层次结构图和桑基图，分别展示社团之间的拓扑结构、层次关系和属性分布，来解决不同的分析任务.

(3) 集成上述方法和设计，开发了一套多元网络图的可视分析系统，并通过实验发现了不同领域内隐藏的关联结构与语义构成特征.

本文第 2 节阐述大规模图简化方法的相关工作；第 3 节介绍本文所提语义增强的大规模多元图简化可视分析方法的工作流程；第 4 节论述大规模多元图的数据处理和模型计算；第 5 节重点介绍语义增强的可视分析系统视图和相关的可视化设计；第 6 节讨论不同数据集的案例分析和专家反馈；第 7 节总结本文算法和展望未来工作.

2 相关工作

国内外研究学者面向大规模网络数据开展了大量的简化表达与可视分析研究工作. 本章将从图采样、边绑定以及图聚类等角度出发对相关工作进行详细的阐述.

2.1 基于图采样的简化方法

所谓图采样简化方法，是指通过筛选部分节点或边来构建原始网络图的一个子图. 采样获得的子图不仅可以逼近原始大规模网络图的拓扑特征，如度分布、聚类系数、连通分量等，而且能够有效降低大规模网络图引起的视觉混淆和分析复杂性[4]. 面向不同拓扑结构特征分析与简化目标，大规模图采样方法主要可以分为三类，包括图节点采样、边采样和遍历的采样策略.

随机节点采样[5-6]是常用的图节点采样策略，是从图中随机地抽取一组节点，如果节点之间有关联则填加相应的边完成子图创建. 与随机节点采样思想类似，随机边采样[5]通过筛选一组边集生成采样子图. 具体的实施过程中有多种改进方案，如随机节点—边采样策略，即随机地选择一个图节点，进而随机地选择与之相邻的一条边；而随机边—节点采样，则是先选择边，再选择边相连的节点.

图节点采样及边采样策略所获得的子图通常是稀疏连接的，难以保证原始大规模网络图的连通性. 因此，大量研究提出基于遍历的采样策略，主要包括深度优先采样和广度优先采样两种策略[7]. 前者从随机选择的节点开始，按照深度优先顺序选择节点，而后者按照广度优先顺序选择节点，二者均倾向于保留度数较高和 PageRank 值较高的节点. 之后研究者又提出多种改进方法，如滚雪球(Snow-Ball)和森林火灾(Forest Fire)[8]、随机游走采样[9]等.

图采样的思想主要是借助于规模较小的子图近似描述原始的大规模网络图，从而达到简化表达与优化视觉感知的目标. 为了有效评估图采样算法的效果，Wu 等人[10]从可视化的角度，重点研究不同的图采样策略对多种视觉特征感知的影响，如高度节点、聚类质量和覆盖区域等.

2.2 基于边绑定的简化方法

通常情况下，网络图可视化遵循“直线边原则”的美学标准. 然而，随着网络规模的扩大，大量的直线交叉严重影响了网络结构的可读性. 为有效简化

大规模网络图中边的可视化，Holten 等学者在网络图的层次布局中首次提出边绑定技术[11]，查找某条边所连两个节点的共同祖先节点，获得共同祖先节点到这两个节点的路径，利用 Bezier 曲线沿着节点路径绘制连线，进而减少线条交叉引起的视觉紊乱. 这种思想的实质是将视觉上相似的一组边捆绑成可辨识的、宏观的边束，从而提取出网络图的骨干结构，方便用户进行认知. 基于这种思想，研究者提出了基于不同策略的边绑定方法，如几何结构[12]、路径规划[13-14]、图像骨架[15-16]、力引导[17-18]、层级聚类[19]、核密度[20]等. 不同于上述在连线中间进行绑定的方法，Peng 等学者提出 SideKnot[21]，通过对连接至同一节点的边进行聚类，并将边绑定应用在连线两端.

边绑定技术能够降低大规模线条交叉导致的视觉紊乱，然而具体的节点之间的关联关系却被隐藏，容易带来图布局理解的歧义，存在局限性[22].

2.3　基于图聚类的简化方法

图聚类简化的核心思想是将网络图中的多个节点汇聚成类，在原始网络结构的基础上，构建一个更为粗糙的超节点图，进而实现大规模网络图的简化表达与可视化. 根据节点聚类或者分类的标准可以将图聚类简化方法分为如下 3 类：

(1) 网络图节点的属性常用于驱动节点的聚类和绘制. PivotGraph[23]将具有相同属性的节点作为一类，然后按照类别将节点布局于不同空间进行可视化. Pretorius 等学者[24]依次将节点的不同属性作为网络层次结构每一级的划分标准，用于分析和探索状态转移图. 类似的其他几种方法[25-28]，也是完全按照节点属性对网络图节点进行聚类和可视化.

(2) 拓扑结构特征是一种有效的图节点聚类依据. 该类方法将具有紧密拓扑关系的一簇节点聚为一类，形成多层级的聚类结构. ASK-GraphView[29]允许用户通过扩展或折叠节点，设置和浏览不同的聚类级别. Archambault 等学者[30]允许用户直接在层次结构上交互式地编辑聚类，实现聚类的分裂、融合、删除等操作. Vehlow 等学者[31]针对加权无向图设计一种分析模糊聚类的可视化方法. Batagelj 等学者[32]通过定义类内和类间的拓扑性质，自动生成原始网络图的聚类结构. Vehlow 等学者[33]提出一种基于流图的可视化方法，用以探索动态网络图中聚类结构的演化. Rieck 等学者[34]提出一种基于不同边权重阈值设置的交互式可视化工具来检测聚类，并展示这些聚类之间的演化关系.

(3) 亦有大量研究同时考虑节点的属性和拓扑结构对网络图节点进行聚类和绘制. Itoh 等学者[35]按照属性对节点进行聚类，进而根据聚类间的拓扑结构生成一个超图，利用力导向布局和矩形填充方法实现网络图简化可视化. OnionGraph[36]同时考虑了网络语义和拓扑结构，将异构图的节点进行分层聚类和展示. JauntyNets[37]将基于属性相似性产生的节点吸引力引入传统的力引导方法中，生成全新的网络图布局结果. 然而，以上方法中的节点属性的考虑局限于单一维度，而且属性在布局中起主导作用，一定程度上破坏了原有的网络拓扑结构. Archambault 等学者[38]根据数据属性中的模式来创建和修改图层次结构，主要依赖用户交互探索图层次空间. Liu 等学者[39]设计研制 HybridVis，在不同尺度上对一个或多个属性进行聚类分析，借助平行坐标等视图辅助查阅聚类的属性信息. 但这些方法都需要大量的用户学习时间.

3　系统流程概述

由相关工作分析可以看出，传统的大规模图简化可视化方法主要侧重于节点和边的拓扑结构特征，容易忽略网络节点本身所具备的多维属性信息，妨碍用户对网络深层语义及拓扑关系的理解和认知. 因此，本文提出语义增强的大规模多元图简化可视分析方法. 其中语义强调的是数据所对应现实世界中的事物所代表的具体含义，以及这些含义之间的关系，而本文结合网络图的多维属性信息关注社团在不同领域上的解释和逻辑表示，重点挖掘语义特征明显的社团以及它们之间的关联. 首先，结合模块度和多维属性信息熵构造多尺度的社团集合，这些社团在结构和属性上都具有较强的聚集特性，并为每个社团标记最优属性值；然后，利用多级力引导布局展示社团之间的拓扑关系，通过颜色映射增强显示社团的语义表达；进而，设计优化改进的层次视图和属性桑基视图，辅助用户查看社团的层次关系和属性分布；最后开发集成上述方法的多元图可视分析系统，在降低网络图规模的基础上，提供多种交互，从不同角度方便用户理解网络图的语义构成以及探索多维属性在社团形成和网络演化中的作用. 整个系统流程图如图 1 所示.

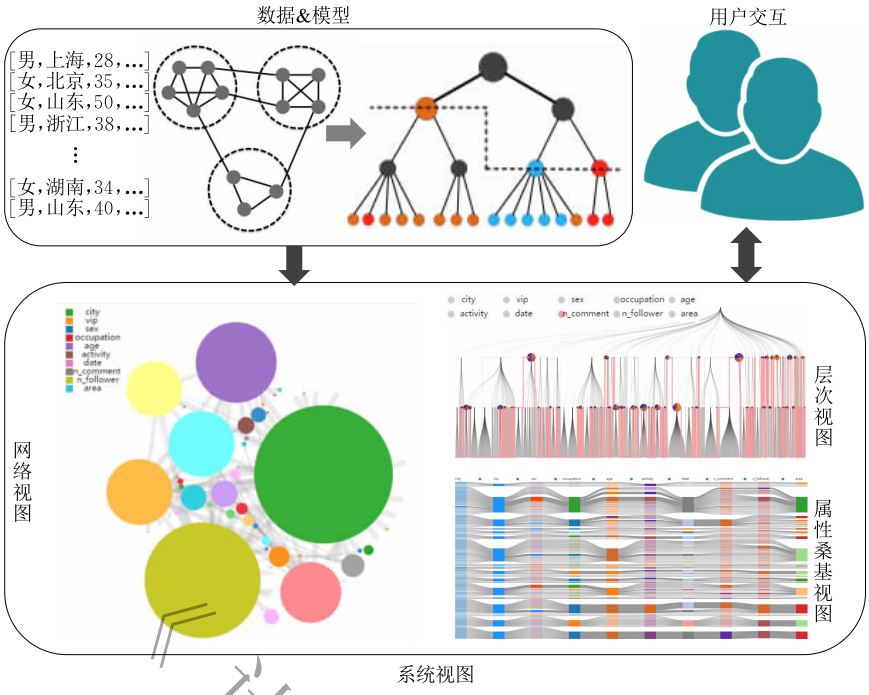


图 1 系统流程图

4 语义增强的大规模多元图聚类

本节详细介绍语义增强的大规模多元图聚类的数据模型和计算方法.

4.1 数据模型

利用 $G=(V,E)$ 表示一个网络图,其中 V 和 E 分别代表网络图中的节点集合和边集合,满足 $E\subseteq V\times V$. 在多元图中,每个节点都有 n 维属性 a_1, a_2,…,a_n. 可以利用 n 维向量来表示多维属性节点,比如 $\boldsymbol{v}=(v_1,v_2,\cdots,v_n)$,其中 $v_i\in D_i$,而 $D_i=\{d_{i,1},d_{i,2},\cdots,d_{i,q_i}\}$ 表示属性 a_i 的取值范围.

4.2 层次聚类

Blondel 算法[40]是一个基于模块度优化的图聚类检测算法,其形式化表达如式(1)所示:

$$Q=\frac{1}{2m}\sum_{ij}\left[A_{ij}-\frac{k_ik_j}{2m}\right]\delta(c_i,c_j) \tag{1}$$

其中,m 表示网络图中边的数量,A_{ij} 表示节点 i 和节点 j 之间连边的权重,k_i 和 k_j 分别表示节点 i 和节点 j 的出度,c_i 和 c_j 分别表示节点 i 和节点 j 所属社团的编号,而 $\delta(c_i,c_j)$ 判断节点 i 和节点 j 是否同属一个社团,如果同属一个社团则值为 1 否则为 0. 其核心思想是将图划分成一系列的社团,并且社团内部紧密相连,而社团之间的联系相对较弱,然后将该算法迭代地应用在新生成的加权网络图上不断地缩小网络规模直到模块度 Q 不再增加为止. 这样一个连续的划分过程(图 2)最终生成网络图的层次结构(图 3(a)). 相比于其他聚类算法[41-44],Blondel 算法兼顾质量与效率,用于大规模的网络图聚类时用时较短且聚类质量较高,因此本文使用该算法作为网

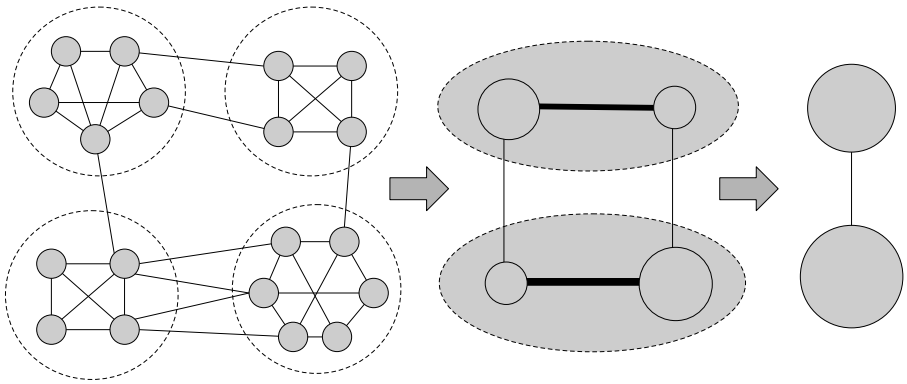


图 2 基于模块度的图聚类过程

络图层次聚类的基础算法.

基于模块度的图聚类算法可以查找出在拓扑结构上具有较强关联的社团，它们满足在结构上的聚集特性. 然而该算法只考虑了网络图的拓扑结构特征，忽略节点的多维属性描述，无法提取有效的语义特征，从而不能帮助用户区分社团是什么样的社团、社团在不同属性作用下的聚集规模以及不同种类社团之间的相互作用. 因此，综合考虑拓扑结构和多维属性信息，抽象大规模多元图的多尺度的网络结构，有效降低节点和边的数量，对于减轻用户的认知负担，帮助用户聚焦于重要的社团结构和属性信息具有重要的意义.

4.3　考虑属性聚集的层次划分

在经典的层次聚类的基础上，为有效增强表达具有较强属性聚集特点的社团，本文引入信息熵有效度量属性的分散程度，进而设计考虑属性聚集的网络图层次聚类方法.

一个网络图形成的层次聚类结构可以表示成一系列的图 $G_0, G_1, G_2, \cdots, G_{N-1}$. 其中 G_0 是最底层的原始图而 G_{N-1} 是抽象最高的图. 往往图聚类很难聚类成一个点，我们添加一个根节点作为层次结构图的最顶端，包含原始网络图的所有节点. 定义一个社团 $C_{x,y}$ 为处在第 x 层的第 y 个社团，我们用信息熵[45]来衡量社团内部所有底层节点的属性聚集度，例如 $C_{x,y}$ 在属性 a_i 的聚集度计算公式如下：

$$IE(C_{x,y}, a_i) = -\sum_{k=1}^{q_i} p(d_{i,k}) \log p(d_{i,k}) \quad (2)$$

其中，q_i 表示第 i 个属性的取值范围大小，即不同取值的数量，$p(d_{i,k})$ 表示社团 $C_{x,y}$ 中第 i 维属性值等于 $d_{i,k}$ 的底层节点数量占比. 根据信息熵的原理可知，如果社团内部在当前属性上出现一个占比很大的值，那么计算出的结果越接近 0，反之如果属性值分布的越分散越均匀，那么结果越大甚至趋向无穷. 如果某个属性为连续型数值时，我们可以根据经验或者平均分配的原则提前对数值进行分组，尽量保证不要出现某个值占比过大的情况. 因为一旦某个属性出现绝大部分集中在某个值上的情况，也就意味着当前属性对节点的判别价值降低，失去了探索属性对网络结构影响的意义.

因此，本文设置两个阈值 ε_1 和 ε_2 ($0<\varepsilon_1<\varepsilon_2$) 来限定信息熵的范围. 其中 ε_1 用于在原始网络上判断所有节点中某个属性的聚集程度是否过高，例如某个属性在整个网络节点上的信息熵小于 ε_1，则认为该属性的聚集程度异常较高，将在后面的社团查找过程中忽略该属性. ε_2 表示我们对社团在属性聚集程度上的容忍下限，如果某个属性在一个社团上的信息熵小于 ε_2，则认为该社团在当前属性上具有明显的聚集特征，那么这个社团将在所处的层级保留. 我们通过实验发现 ε_1 和 ε_2 取值为 0.1 和 0.2 时，能较好的应对大部分多元图，因此将这两个值设置为参数初始值，同时在最后的可视分析系统中提供修改参数的接口，方便用户根据特定的分析任务和需求调整参数来获得不同属性聚集程度下的社团结构. 如果出现多个属性值的信息熵都小于 ε_2，本文将选用信息熵最小的属性并且用其中占比最大的属性值标记该社团. 如果没有属性满足信息熵小于 ε_2，那我们用同样的方法依次检测它的下一级所有社团，直至遍历至底层叶子节点为止. 具体的考虑属性聚集的层次划分流程如下：

Step 1. 对属性集 $A=\{a_1, a_2, \cdots, a_n\}$ 中的每个属性 a_i，依次计算其在根节点 $C_{N,0}$ 上的信息熵，如果 $IE(C_{N,0}, a_i)<\varepsilon_1$，则从属性集中移除 a_i；

Step 2. 设置社团检测队列 Q 和社团存储集合 S，初始化 Q 和 S 都为空；

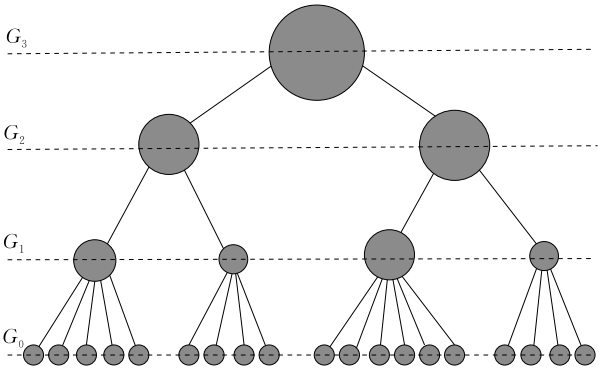
Step 3. 将 G_{N-1} 上的所有社团压入队列 Q；

Step 4. 如果队列 Q 为空，则算法结束，否则从队列 Q 中出队一个待检测社团，依次计算属性集 A 中保留的每个属性在该社团上的信息熵，并按照升序排列，筛选出信息熵值小于 ε_2 的前 L 个属性；

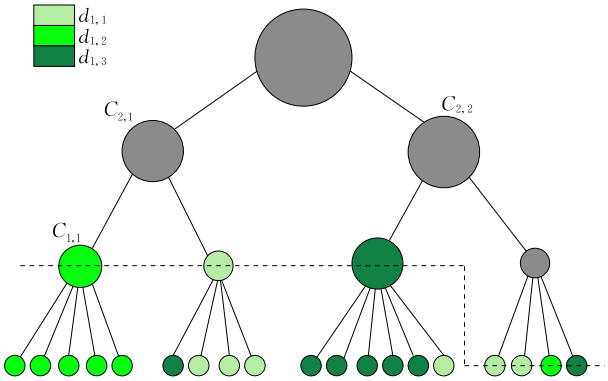
Step 5. 如果 $L\geq 1$，则选择信息熵最小的属性，并用其在该社团中占比最大的属性值标记该社团，将该社团存入集合 S 并返回执行 Step 4；

Step 6. 如果 $L=0$，则将当前社团的所有下一级子社团压入队列 Q，返回执行 Step 4.

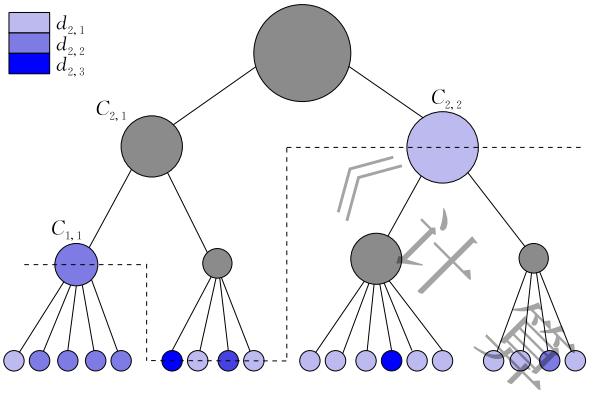
最后社团存储集合 S 记录的 K 个社团，$\{C_{x_1,y_1}, C_{x_2,y_2}, \cdots, C_{x_K,y_K}\}$，也就是我们要查找的符合在结构和属性上都具有较好聚集特性的社团. 属性集包含属性数量的多少决定了该算法既适用于多维属性，也适用于单维属性. 例如，图 3(b) 和图 3(c) 分别展示了在属性 a_1，a_2 上的层次划分，图 3(d) 展示了综合属性 a_1 和 a_2 的层次划分. 当检测社团 $C_{2,1}$ 时，因为 $IE(C_{2,1}, a_1)$ 和 $IE(C_{2,1}, a_2)$ 都大于阈值 ε_2，所以开始检测它的子社团 $C_{1,1}$，经对比发现 $IE(C_{1,1}, a_1)<IE(C_{1,1}, a_2)<\varepsilon_2$，而且值等于 $d_{1,2}$ 的节点数量最多，因此用 $d_{1,2}$ 来标记 $C_{2,1}$. 用类似的方法检测社团 $C_{2,2}$ 时，发现 $IE(C_{2,2}, a_2)<\varepsilon_2<IE(C_{2,2}, a_1)$，并且值等



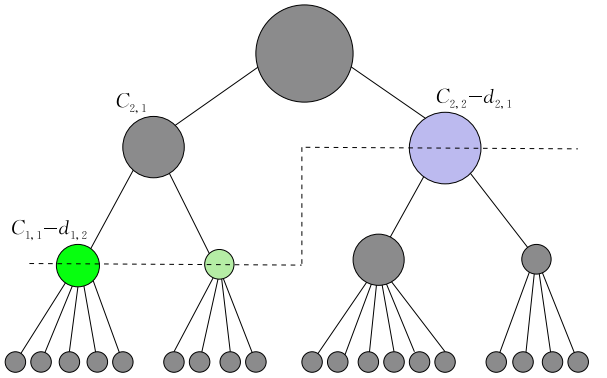
(a) 图2的聚类层次结构



(b) 基于属性a_1的社团查找



(c) 基于属性a_2的社团查找



(d) 基于属性a_1和a_2的社团查找

图 3 考虑节点属性的社团查找

于$d_{2,1}$的节点占比最大,因此用$d_{2,1}$来标记$C_{2,2}$.这样可以提取出具有显著语义信息的社团结构.

5 语义增强的大规模图简化可视分析

本节主要介绍语义增强的多元图可视分析系统的各个视图及其设计流程.图 4 所示为系统视图的概览,主要包含三个视图,我们将一一展开论述.

5.1 多尺度力引导图

根据第 4 节描述的多元图聚类算法,我们可以从原始图$G_0=(V_0,E_0)$抽象出一个新的图$G'=(V',E')$,$E'\subseteq V'\times V'$.其中V'代表筛选出的社团集合,任意的社团$v'\in V'$都有一个标记$d_{i,j}$,表示该社团在属性a_i上的聚集效果最好,并且第j个值占比最大.

我们首先利用考虑节点大小的力引导算法将

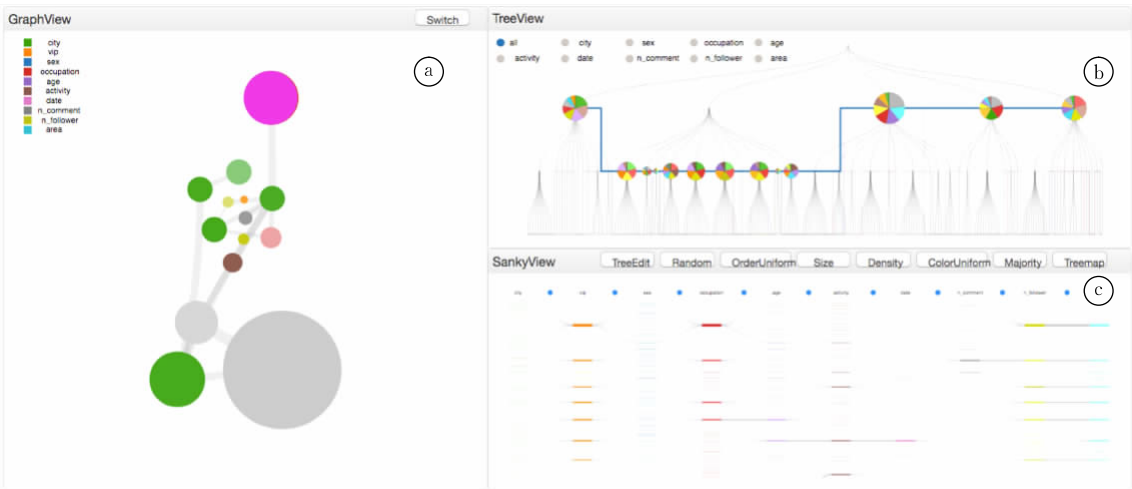


图 4 语义增强的多元图可视分析系统视图

G_{N-1}中的所有社团进行布局，每个社团用圆来表示，其半径与内部所含的节点数量成正比，这样布局的社团之间不会产生重叠. 之后判断G_{N-1}中的每个社团是否在V'中，如果在的话，跳过该社团检查另一个社团；否则在其相应的圆形区域内再次利用考虑节点大小的力引导算法对其下一级的社团进行布局，重复上述过程，直至V'中所有社团布局完成为止.

在可视化设计方面(图 4(a))，我们使用不同颜色映射不同的属性，同一色系下不同的颜色深度表示相应的属性值. 随即根据V'中社团的属性值标记来赋予对应圆形的颜色. 之后根据E'添加社团之间的连线，连线宽度表示两个社团底层节点之间的连边数量. 最后我们对连线借用边绑定算法进一步降低可视化效果的紊乱度. 我们在这种多尺度网络布局的基础上又设计了如图 10 所示的嵌套式展示方法，如果某个社团不在网络聚类的层次划分上，则将其绘制成圆环，圆环的半径表示其内部节点的数量，而扇区则表示不同的属性值占比，接着在其内部继续绘制下一级的社团. 当用户点击视图菜单栏右侧的“Switch”按钮时，网络视图会在上述两种可视化方法与原始的网络图之间进行切换.

5.2　语义增强的层次树图

展示多维属性和不同单维属性上的层次划分可以帮助用户直观地了解它们对应的社团所处层级和规模. 我们设计了图 4(b)所示的增强显示的层次树图. 视图的左上角显示数据的各个维度，当用户点击属性全集“All”或者某个属性a_i时，相应的层次划分线和在其上的社团将被高亮显示. 社团用饼状图绘制，其半径大小代表社团内部节点数量的多少. 当展示单维属性时，饼状图显示社团内部各属性值的占比信息；当展示多维属性时，饼状图则显示社团内部各属性的聚集特征，例如某个属性的信息熵越小，则聚集性越好，对应的扇形角度越大. 点击左上角的属性全集或者单个维度时，左侧的网络视图(图 4(a))也会进行更新，展示当前维度下的聚类图.

5.3　属性桑基图

多属性的聚类层次结构会隐藏掉各个维度上的聚类层次细节，为了解决这个问题，我们设计了属性桑基图，进一步对比社区在不同属性上的聚集关系. 首先，受平行坐标的启发，我们为每个属性设置了一个垂直的且不可见的轴，这些属性轴是平行且等间距的；其次，对于每个属性，聚类层次划分中的社团会被竖直放置在相应的属性轴上，如图 5 所示，C_1、C_2、C_3是检测出的在属性a_1上具有较好聚集特性的社团，对应于a_1轴上的三个矩形块B_1、B_2、B_3，每个块的高度代表社团内部的节点数量；接着，如果分布在相邻属性轴上的两个矩形块表示的两个社团之间存在包含关系，则在相邻的属性轴间添加水平条带来连接两个社团，条带的宽度映射子社团的大小. 这样分布在不同属性轴间的条带在视觉上形成了连续的跨过多属性轴的流图.

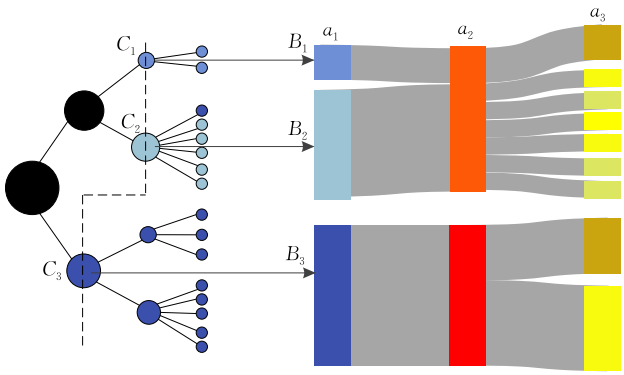


图 5　属性桑基视图的可视化设计

此外，矩形块可以设计成不同的图案来展示更丰富的信息，在我们的系统里提供了如下几种方式：矩形块完全由标记该社团的属性值所对应的颜色填充；矩形块用 ColorBar 的形式展示，可以呈现社团内部详细的属性值占比；矩形块用 TreeMap 的形式展示，每个格子表示一个节点，颜色根据节点属性值填充，这样既呈现了社团内部详细的属性值占比也反映了社团内部的层次结构. 三种方式展示的信息量依次递增.

属性之间的聚类关联主要取决于属性轴的排序. 如果随机排列属性轴，隐藏在数据中的一些模式将被打乱. 例如在两个属性上检测到的社团集如果是相同的，可以认为这两个属性在这些社团的形成中共同发挥作用. 但如果这两个属性轴不相邻甚至距离很远，那么用户就很难观察到这种模式. 虽然用户可以通过系统提供的交互来调整属性轴的顺序，但对于具有较多属性的网络图来说，其探索时间仍然较长. 因此我们构造属性对应的社团集到根节点形成的树状结构，通过比较这些树之间的编辑距离来衡量社团集的相似性. 如果某两维属性的树编辑距离越小，说明它们的社团集越相似，则对应的属性轴放置的也越近.

6　结果评估

这一节我们重点以微博转发数据和论文引用数

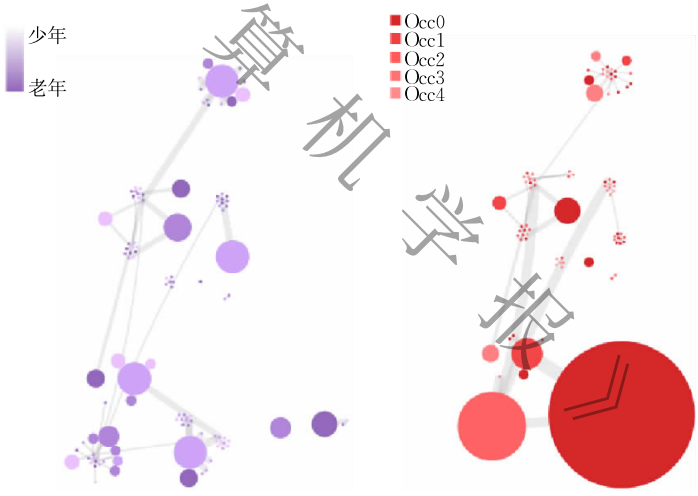
据为例,探讨如何用我们的系统简化分析不同应用背景下网络数据集中属性和拓扑结构的内在联系,以及邀请领域专家对我们的系统进行使用和评估.

6.1 微博转发

新浪微博是中国最流行的社交网站,我们通过新浪 API 跟踪了具有不同热度的三条微博的转发数据.第一条微博转发规模较小,参与用户在 500 个左右;第二条微博转发规模中等,参与用户超过 5000 个;第三条微博转发规模最大,参与用户达 50000个之多.网络图数据中每个节点代表转发微博的用户,边代表用户之间的转发关系,而节点的属性就是用户的个人资料信息,包括性别、年龄、职业、所在城市、和活跃度等 10 个属性.

图 4 展示的是加载微博转发数据集 I 后的系统效果图.其中图 4(a)反映了综合多属性后的聚类效果,并且右下角有个明显的规模较大的用户社团,根据颜色(灰色对应微博的评论字数)可知,该社团在属性评论数方面具有最佳的聚集特征,进一步探索得知这个用户群体在转发微博时大都添加了自己的观点和评论,讨论比较积极.类似的最上面还有一个品红色(品红色对应微博转发日期)的社团,可知该社团用户之间的微博转发大都发生在同一天,转发时间比较集中.此外,我们还发现很多绿色(绿色对应城市)的社团分布其中,可知来自不同城市的用户大多形成了各自地域内的转发群体.

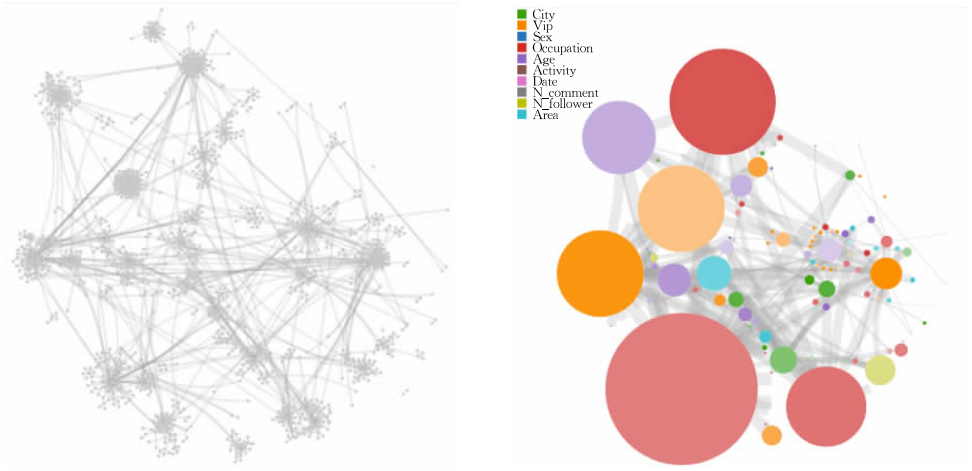
图 6 展示了不同单维属性上的聚类情况.其中图 6(a)中根据年龄聚类而成的社团规模较小,分布比较离散,说明年龄在社团形成中的内聚特性不明显;而图 6(b)中根据职业聚类而成的社团规模大小不一,间接反映了不同职业对社团形成的作用明显不同,有些职业的用户之间广泛联系,转发互动性强,因而容易形成较大规模的社团.而层次视图中高亮的聚类划分线和社团也可以同步印证以上结论.



(a) 年龄 (b) 职业

图 6 分别基于年龄和职业的社团聚类划分(微博转发 I)

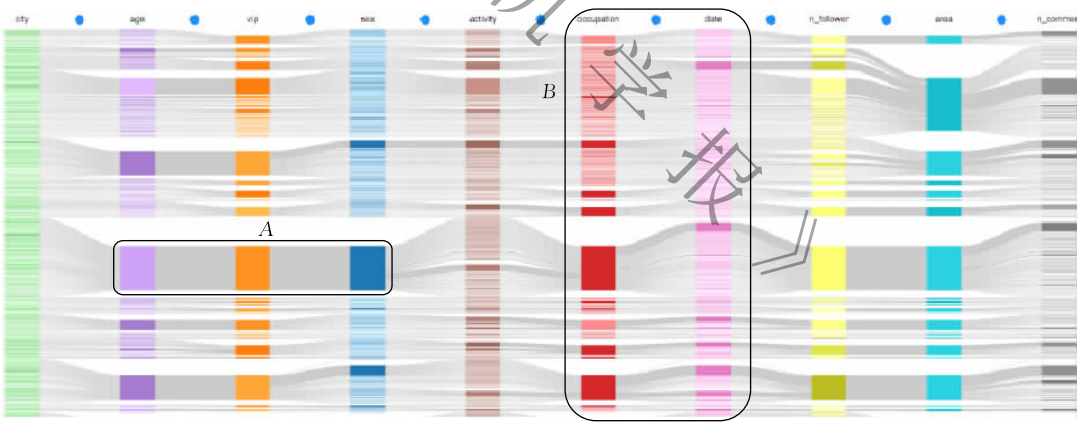
图 7 展示了原始网络图和加载微博转发数据集 II 后的多维属性聚类效果.通过对比发现,我们的方法可以将节点数量从 5000 有效降低到 100,边的数量也显著减少.用户通过宏观的社团结构可以大致了解原始网络的构成,社团节点的颜色和大小反映了该社团内部具有一定规模的关联密切的用户群体,并且这些群体大都具有相同的属性值.通过社团之间的位置关系和连边,也可以了解不同特定群体之间的相对互动关系.在这套数据中,我们发现微博转发日期这一属性被参数 ε_1 过滤掉,根据统计结果显示 99%的转发集中在初始微博发布后的数个小时内,只有少量转发出现在第二天,这样对所有微博进行转发日期聚集度的计算时,其信息熵值远小于参数 ε_1 的初始值 0.1,从而被忽略,避免影响其他对社团形成更具实际意义的属性检测.而最终的社团结构对参数 ε_2 较为敏感,取值较高时意味着社团属性聚集度要求较低,容易形成较多大尺度的社团,反之取值较小时会倾向形成众多小规模的社团.



(a) 原始网络图　　(b) 本文方法

图 7　原始图和多维属性聚类后的结果对比(微博转发 II)

我们还可以利用属性桑基图进一步对比不同属性之间聚类结构的关系. 如图 8(a)所示，区域 B 中相邻的两个属性是职业和转发日期，可以看到线条从职业过渡到转发日期时大都明显的分散开，说明职业这个属性在微博转发数据 II 中对社团形成的作用要强于转发日期，因为它所形成的社团规模更大. 属性桑基图也可以帮助用户查找一些特定的社团，比如区域 A 中的社团，它在年龄、VIP 等级和性别方面都有较强的聚集性，说明这三个属性在该社团的形成中都起了较强的作用，其所包含的用户是一群性别一致、年龄相仿以及 VIP 等级相同的微博用户. 当通过交互用 Treemap 的形式展示社团矩形块时(图 8(b))，用户可以详细了解其内部构成，包括层次结构和属性值分布.

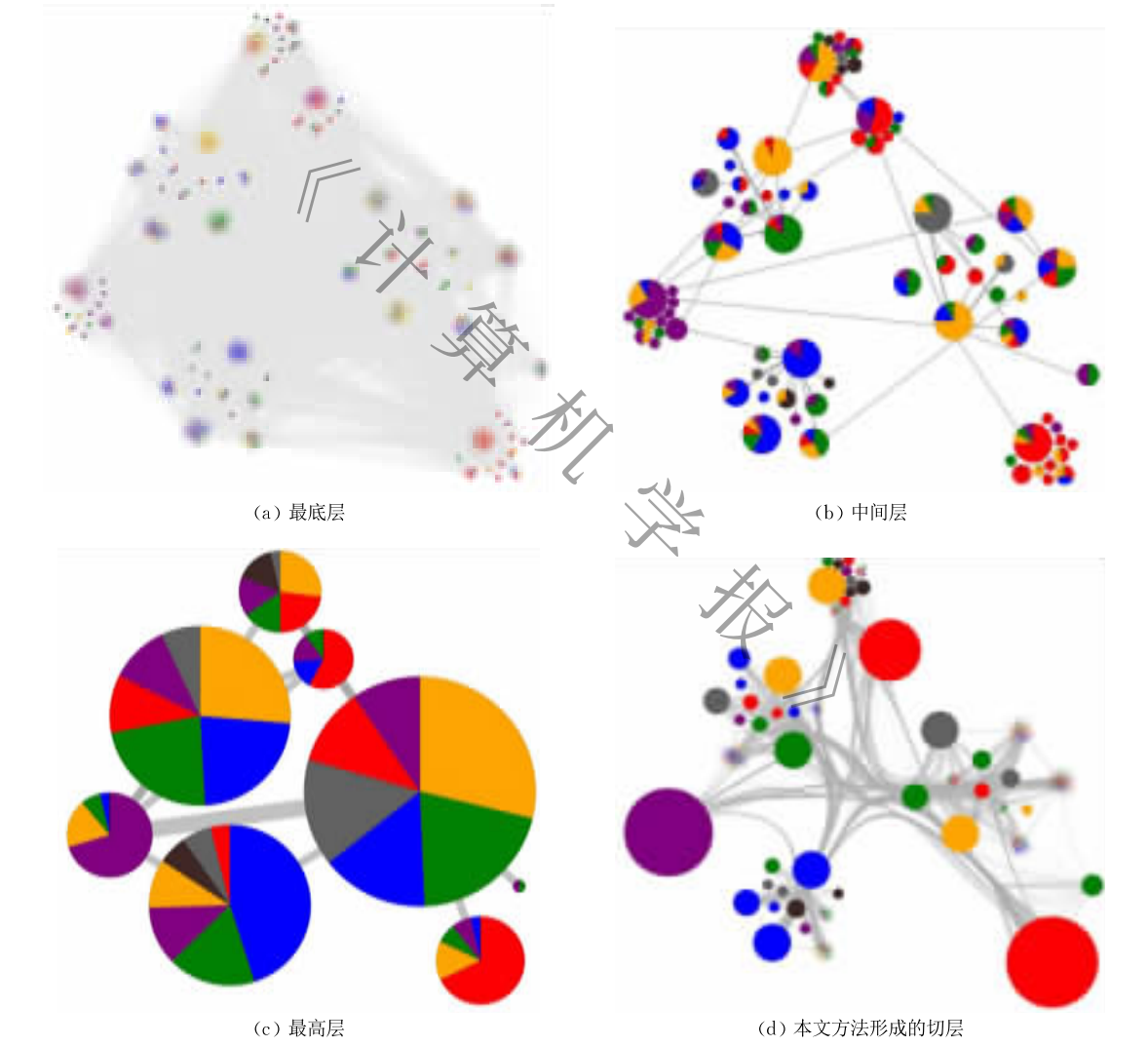


(a) 根据占比最大属性值填充的单一颜色

(b) Treemap

图 8　属性桑基图中矩形块的两种展示方式(微博转发 II)

图 9 分别展示了使用微博转发数据集 III 形成的传统层次化聚类方法和本文方法的效果对比，其中颜色代表用户所在省份，我们借此研究在该事件的传播中地域对社团形成的作用以及社团之间的亲疏远近关系. 其中图 9(a)展示了原始的最底层网络图，而图 9(b)和图 9(c)展示的是中间层和最高层的聚类结构，饼状图代表一个社团，其大小反映了社团的规模，内部扇形反映了社团中不同省份的用户比例. 虽然用户通过这种方式可以自底向上或者自顶向下逐级探索，但是在分析过程中会不可避免地出现以下问题. 当在接近底层细节时，由于用户只能聚焦于局部区域来分辨不同类别的用户在该群体的比例，然后依次探索其他区域，因此无法对网络社团及其属性信息形成整体连贯的心理感知地图. 而在较高层级时，用户可以通过饼状图直观了解社团的属性组成，但是在分析属性杂乱的社团时，其更深层次的成因往往不得而知. 有可能该社团内部是多个属性聚集但是取值各不相同的子社团，也可能所有子社团的属性组成都是杂乱无章的、语义特征不明确的，需要进一步往下拓展分析，从而导致用户的交互成本增加、探索时间过长.



(a) 最底层 (b) 中间层

(c) 最高层 (d) 本文方法形成的切层

图 9 层次化聚类方法与本文方法对比(微博转发 III)

而图 9(d)展示的是本文方法，避免了上下层级之间反复迭代的交互操作，直接将不同层级中具有最优属性值的社团筛选出来，同时这些社团属性聚集程度较高、语义特征明确，只用对应的主色显示即可，增强了用户对大规模多元图的视觉认知. 例如图中红色代表北京，可以快速观察到两个以北京用户为主体形成的社团，但是它们距离较远，并未有太多互动. 而图中紫色社团表示该微博转发群体以上海用户为主，且在该事件的传播中上海用户只形成了这一个有代表性的社团. 相比之下，其他地区(绿色、蓝色、黄色等)形成的社团规模相对较小，但是他们之间的互动更为频繁. 因此本文方法结合属性信息，

在研究社团形成规则、社团之间的分隔及连接张力时更为有效便捷.

6.2　论文引用

这套数据集包含约 2000 个论文节点，引用关系形成的边约 6000 条，这些论文分别来自数据挖掘、信息检索、网络服务、机器学习等 10 个研究领域. 我们重点利用本文算法分析研究领域这一属性在论文引用社团形成中的作用. 论文引用关系中，往往同一研究领域的论文之间引用关系较为频繁，但是交叉学科即来自不同领域的论文形成的引用社团也是引文网络研究者重点关注的现象. 因为交叉学科体现了科学向综合性发展的趋势，是学术研究的前沿阵地.

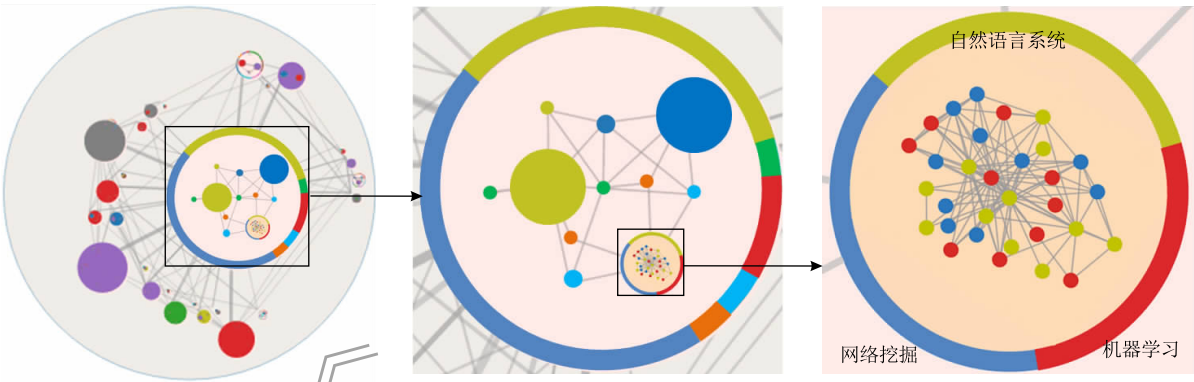


图 10　嵌套式的网络图展示方法

如图 10 所示，首先将网络视图切换到嵌套式的可视化效果进行展示，其中颜色表示论文的研究领域，可发现大部分社团都是由同领域的论文构成，但是中间很大的一个社团展示了内部层级关系，说明该社团包含的论文领域聚集性较弱，接着我们层层递进查看其内部细节，最终发现某个子社团（图 10 最右侧的图）所含的最底层论文节点主要来自三个研究方向，即自然语言系统、网络挖掘和机器学习，说明这些论文虽然来自不同的三个领域，但它们之间比较频繁地引用和借鉴了另一门或者两门学科的技术方法，创造出了跨学科的研究成果. 这可以帮助相关领域的研究人员进行交叉学科的论文检索与推荐. 这也反映了本文方法在多属性聚类和单属性聚类方面都有较强的适用性.

6.3　专家反馈和用户实验

我们邀请了来自引文网络推荐算法设计、社交网络分析和信息检索的三位专家，在向他们介绍我们的方法和系统后，让他们一一使用了我们的多元图可视分析系统来分析不同的数据集，之后我们记录了他们对该系统的点评. 引文网络推荐算法设计专家表示该系统提供了友好的交互接口，在减轻对大规模网络结构认知负担的同时，能够有效帮助他探索论文转发的关系结构和层次结构，并且轻松查找交叉领域的论文结果和所属领域. 而社交网络分析专家表示，“不同事件的微博转发网络有不同的特点，这个系统可以让用户方便的对比这些事件的网络结构，直观理解不同的用户属性在网络演化和社区形成中扮演的角色.”信息检索专家表示，“目前具有多维节点属性的网络无处不在，我们可以从交通网络、生物网络等其他网络提取出大量的节点信息，而该系统可以让不同领域的研究者深入探讨网络结构和语义信息之间的关联.”

最后专家也提到了一些需要改进的地方. 一个是连续变量的分组问题，因为不同的分组策略将影响最优社团结构的发现. 例如分组数量的多少以及均匀或者不均匀的分组都会对属性聚集的质量产生影响. 其次就是颜色的选择. 多元图往往具有较多属性，每个属性都有很多值，但是可使用颜色的数量有限而且颜色深度很难帮助用户区分不同的属性值，这会给用户的认知带来混乱，应用其他的编码元素，例如形状、纹理等或许是可行的方案. 第三，专家们认为，如果能够提供社团聚类层次结构的编辑功能以及社团的扩展和折叠功能，可以更方便地帮助用户探索大规模多元图. 此外，专家提到不同属性的取值空间差异较大，比如性别只有两种，而省份多达三十几个，那么在统一信息熵阈值比较的情况下，这些取值较多、更为离散的属性可能会被取值较少、更为集中的属性掩盖掉. 而单单通过层次树图中饼状图的颜色比例也很难让用户了解这些细节，因此如何更科学的量化和展示不同属性的聚集程度还需要进一步深化. 最后有专家反馈在交互探索第三个微博数据（5 万节点）时遇到卡顿和响应时间较长等问题. 目前我们的实验环境是单机，在进行更大规模的实验时发现运行几十万节点的多元网络数据就会出

现内存不足和浏览器加载崩溃等问题.从理论上讲本文算法可以支撑更大规模的数据集并且表现更好,但是受限于硬件设备,如果要处理百万级别甚至千万级别的网络数据则需要搭建 PowerGraph 这种分布式的大规模图数据并行处理框架和集群,这也是下一步我们要改进的方向.

7 总结与展望

本文提出一种针对大规模多元图的语义增强简化可视分析方法.首先,在借助模块度提取出网络聚类层次结构的基础上,通过引入阈值和多维属性信息熵的计算和比较分析进行网络层次结构的自适应划分;其次,针对筛选出的在结构和属性上都具有较好聚集特性的社团,分别设计可交互关联的网络视图、层次视图和属性桑基视图来分析社团之间的拓扑结构、层次关系和属性分布;最后,进一步整合上述方法,开发了语义增强的大规模多元图简化可视分析系统,通过实验微博转发数据和论文引用数据,从不同角度帮助用户分析了用户属性(职业、地域、年龄等)和论文属性(领域、发表时间等)在用户社团和论文社团形成中的作用.

参 考 文 献

[1] McPherson M, Cook S L M. Birds of a feather: Homophily in social networks. Annual Review of Sociology, 2001, 27: 415-444

[2] Yavas M, Yucel G. Impact of homophily on diffusion dynamics over social networks. Social Science Computer Review, 2014, 32(3): 354-372

[3] Theodosiou T. Protein—protein interaction predictions using text mining methods. Methods, 2015, 74: 47-53

[4] Rafiei D, Curial S. Effectively visualizing large networks through sampling//Proceedings of the 2005 IEEE Visualization. Minneapolis, USA, 2005: 375-382

[5] Leskovec J, Faloutsos C. Sampling from large graphs//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA, 2006: 631-636

[6] Stumpf M P, Wiuf C, May R M. Subnets of scale-free networks are not scale-free: Sampling properties of networks. Proceedings of the National Academy of Sciences of the United States of America, 2005, 102(12): 4221-4224

[7] Doerr C, Blenn N. Metric convergence in social network sampling//Proceedings of the 5th ACM Workshop on HotPlanet. Hong Kong, China, 2013: 45-50

[8] Leskovec J, Kleinberg J, Faloutsos C. Graphs over time: Densification laws, shrinking diameters and possible explanations//Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA, 2005: 177-187

[9] Lovász L, Lov L, Erdos O P. Random walks on graphs: A survey. Combinatorics, 1993, 8(4): 1-46

[10] Wu Y, Cao N, Archambault D, et al. Evaluation of graph sampling: A visualization perspective. IEEE Transactions on Visualization and Computer Graphics. 2016, 23(1): 401-410

[11] Holten D. Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. IEEE Transactions on Visualization and Computer Graphics, 2006, 12(5): 741-748

[12] Cui W, Zhou H, Qu H, et al. Geometry-based edge clustering for graph visualization. IEEE Transactions on Visualization and Computer Graphics, 2008, 14(6): 1277-1284

[13] Lambert A, Bourqui R, Auber D. Winding roads: Routing edges into bundles. Computer Graphics Forum, 2010, 29(3): 853-862

[14] Luo S J, Liu C L, Chen B Y, et al. Ambiguity-free edge-bundling for interactive graph visualization. IEEE Transactions on Visualization and Computer Graphics, 2012, 18(5): 810-821

[15] Telea A, Ersoy O. Image-based edge bundles: Simplified visualization of large graphs. Computer Graphics Forum, 2010, 29(3): 843-852

[16] Ersoy O, Hurter C, Paulovich F, et al. Skeleton-based edge bundling for graph visualization. IEEE Transactions on Visualization and Computer Graphics, 2011, 17(12): 2364-2373

[17] Holten D, Van Wijk J J. Force-directed edge bundling for graph visualization. Computer Graphics Forum, 2009, 28(3): 983-990

[18] Selassie D, Heller B, Heer J. Divided edge bundling for directional network data. IEEE Transactions on Visualization and Computer Graphics, 2011, 17(12): 2354-2363

[19] Gansner E R, Hu Y, North S, et al. Multilevel agglomerative edge bundling for visualizing large graphs//Proceedings of the Pacific Visualization Symposium. Hong Kong, China, 2011: 187-194

[20] Hurter C, Ersoy O, Telea A. Graph bundling by kernel density estimation. Computer Graphics Forum, 2012, 31(3): 865-874

[21] Peng D, Lu N, Chen W, et al. SideKnot: Revealing relation patterns for graph visualization//Proceedings of the 2012 IEEE Pacific Visualization Symposium. Songdo, Korea, 2012: 65-72

[22] Wang Y, Shen Q, Archambault D, et al. AmbiguityVis: Visualization of ambiguity in graph layouts. IEEE Transactions on Visualization and Computer Graphics, 2016, 22(1): 359-368

[23] Wattenberg M. Visual exploration of multivariate graphs//Proceedings of the 2006 Conference on Human Factors in Computing Systems. Quebec, Canada, 2006: 811-819

[24] Pretorius A J, Van Wijk J J. Visual analysis of multivariate state transition graphs. IEEE Transactions on Visualization and Computer Graphics, 2006, 12(5): 685-692

[25] Zhao J, Collins C, Chevalier F, Balakrishnan R. Interactive exploration of implicit and explicit relations in faceted datasets. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(12): 2080-2089

[26] Bezerianos A, Chevalier F, Dragicevic P, et al. GraphDice: A system for exploring multivariate social networks. Computer Graphics Forum, 2010, 29(3): 863-872

[27] Shneiderman B, Aris A. Network visualization by semantic substrates. IEEE Transactions on Visualization and Computer Graphics, 2006, 12(5): 733-740

[28] Shen Z, Ma K L, Eliassirad T. Visual analysis of large heterogeneous social networks by semantic and structural abstraction. IEEE Transactions on Visualization and Computer Graphics, 2006, 12(6): 1427

[29] Abello J, Ham F V, Krishnan N. ASK-GraphView: A large scale graph visualization system. IEEE Transactions on Visualization and Computer Graphics, 2006, 12(5): 669-676

[30] Archambault D, Munzner T, Auber D. Grouse: Feature-based, steerable graph hierarchy exploration//Proceedings of the EuroVis07: Joint Eurographics-IEEE VGTC Symposium on Visualization. Norrköping, Sweden, 2007: 67-74

[31] Vehlow C, Reinhardt T, Weiskopf D. Visualizing fuzzy overlapping communities in networks. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(12): 2486

[32] Batagelj V, Didimo W, Liotta G, et al. Visual analysis of large graphs using (x,y)-clustering and hybrid visualizations. IEEE Transactions on Visualization and Computer Graphics, 2011, 17(11): 1587-1598

[33] Vehlow C, Beck F, Weiskopf D. Visualizing the evolution of communities in dynamic graphs. Computer Graphics Forum, 2015, 34(1): 277-288

[34] Rieck B, Fugacci U, Lukasczyk J, Leitte H. Clique community persistence: A topological visual analysis approach for complex networks. IEEE Transactions on Visualization and Computer Graphics, 2018, 24(1): 822-831

[35] Itoh T, Muelder C, Ma K L, Sese J. A hybrid space-filling and force-directed layout method for visualizing multiple-category graphs//Proceedings of the IEEE Pacific Visualization Symposium. Beijing, China, 2009: 121-128

[36] Shi L, Liao Q, Tong H, et al. Hierarchical focus+context heterogeneous network visualization//Proceedings of the IEEE Pacific Visualization Symposium. Yokohama, Japan, 2014: 89-96

[37] Jusufi I, Kerren A, Zimmer B. Multivariate network exploration with JauntyNets//Proceedings of the 17th International Conference on Information Visualization. London, UK, 2013: 19-27

[38] Archambault D, Munzner T, Auber D. GrouseFlocks: Steerable exploration of graph hierarchy space. IEEE Transactions on Visualization and Computer Graphics, 2008, 14(4): 900-913

[39] Liu Y, Wang C, Ye P, Zhang K. HybridVis: An adaptive hybrid scale visualization of multivariate graphs. Journal of Visual Languages and Computing, 2017, 41: 100-110

[40] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008, 2008(10): P10008

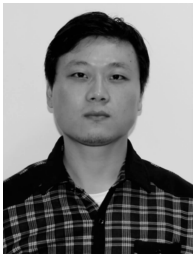
[41] Dhillon I S, Guan Y, Kulis B. Weighted graph cuts without eigenvectors a multilevel approach. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(11): 1944-1957

[42] Navlakha S, Rastogi R, Shrivastava N. Graph summarization with bounded error//Proceedings of the ACM SIGMOD International Conference on Management of Data. Vancouver, Canada, 2008: 419-432

[43] Newman M E J. Modularity and community structure in networks. Proceedings of the National Academy of Sciences of the United States of America, 2006, 103(23): 8577-8582

[44] Brun C, Chevenet F, Martin D, et al. Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. Genome Biology, 2003, 5(1): R6

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



LIU Yu-Hua, Ph. D., lecturer. His research interests include data visualization, visual analytics.

ZHANG Ru-Min, B. S. Her research interests include data visualization and visual analytics.

ZHANG Jing-Yu, B. S. His research interests include data visualization and visual analytics.

GAO Feng, B. S. His research interests include data visualization and visual analytics.

GAO Yuan, B. S. His research interests include data visualization and visual analytics.

ZHOU Zhi-Guang, Ph. D., associate professor. His research interests include data visualization, visual analytics.

Background

This research is related to the field of graph visualization. Previous works focus on how to visualize either the topology of a network or the multidimensional data that is associated with the nodes. However, both addressing topology-based tasks and attribute-based tasks at the same time is still an open research problem.

To solve the difficulty of visualizing such multivariate networks arising from two conflicting goals: visualizing topology and visualizing node attributes, we introduce a semantic-enhanced method in this paper for the visual abstraction of large-scale multivariate graphs. First, the hierarchal structure is extracted from original network dataset by means of a modularity-based graph clustering. Second, a graph cut scheme is designed to optimize the hierarchy of networks by comparing the information entropies of multi-dimensional attributes, so that the communities with obvious aggregating characteristics on one attribute will be displayed in the resultant simplified visualizations. Then, several coordinated views are also provided to visualize the relationships, hierarchies and attribute distributions of these communities, enabling users to further analyze the roles of different attributes in the network evolution and community formation from different aspects. A set of experimental results shows that our method can effectively simplify the visualizations of large-scale multivariate graphs in different fields and help users explore the topological and semantic structures of graphs.

The work of this paper is supported by the National Natural Science Foundation of China (Nos. 61872314, 61802339), the MOE (Ministry of Education in China) Project of Humanities and Social Sciences (No. 18JYC910017), the Natural Science Foundation of Zhejiang Province (No. LY18F020024), the Major Project of Humanities and Social Sciences of Higher Education of Zhejiang Province (No. 2018QN021), the Open Project Program of the State Key Lab of Zhejiang University CAD&CG (No. A1806). Our group has been working on graph visualization for years. Several research papers have been published on *IEEE Transactions on Visualization and Computer Graphics*, *IEEE Transactions on Human Machine Systems*, *Journal of Visualization* and *Journal of Visual Language and Computing*.