

基于证据理论的多关系网络重要节点挖掘方法

罗浩 闫光辉 张萌 包峻波 李俊成 刘婷

(兰州交通大学电子与信息工程学院 兰州 730070)

摘要 多关系网络作为现实世界建模的典型形式,已成为当前网络科学领域研究的热点.挖掘网络中的重要节点作为复杂网络分析领域的基本问题,是理解复杂网络结构和动力学特性的有效方式.关于单关系复杂网络中重要节点的研究已经形成了较为完备的方法框架,但对于多关系网络中重要节点研究缺乏系统性的研究成果,尝试推广已有的重要节点挖掘方法到多关系网络已成为当前研究的热点问题之一.本文综合考虑中心性和传递性对节点重要程度的影响,对无权无向多关系网络中的重要节点挖掘问题进行了系统研究.首先,通过建立多层网络模型对多关系网络进行描述,并构造多层邻接矩阵对其进行表示.其次,按照多层网络节点局部聚集系数的概念,给出了刻画网络传递性的多关系网络节点局部聚集系数的计算方法.在此基础上,结合多关系网络节点的度中心性,推广单关系网络的 ClusterRank 重要节点度量指标到多关系网络,提出了多重 ClusterRank (Multiplex ClusterRank, MCR) 指标.考虑到多关系网络耦合信息和传递机制的差异性会对节点重要程度的度量产生影响,多重 ClusterRank 指标在规模较大的网络中可能会存在计算误差,从而影响度量重要节点的准确性,因此引入 D-S 证据理论对多重 ClusterRank 进行改进,将节点的度中心性和局部聚集系数进行融合,提出了更为高效的多重证据中心性 (Multiplex Evidential Centrality, MEC) 重要节点挖掘方法.最后,在 4 个真实数据集上进行了大量实验,并从鲁棒性和脆弱性、传播动力学特性两方面对所提出的方法进行了评价.实验结果表明,多重证据中心性所得到重要节点度量结果要优于只关注网络结构的度中心性和特征向量中心性所得到的重要节点度量结果,而多重 ClusterRank 在规模较大的网络中确实会存在计算误差,难以准确挖掘网络中的重要节点.实验验证了多重证据中心性可以有效消除多关系网络耦合信息和传递机制在重要节点挖掘中的影响,能够更准确地挖掘网络中的重要节点.此外,通过分析各方法的时间复杂度并统计各方法的实际运行时间,验证了相对于多重 ClusterRank,改进后的多重证据中心性具有更低的时间复杂度.本文所做工作对多关系网络重要节点挖掘问题提供了新的思想方法,并进一步拓展了 D-S 证据理论的应用场景.

关键词 多关系网络;多重网络;重要节点;中心性;传递性;局部聚集系数;证据理论

中图法分类号 TP181

DOI号 10.11897/SP.J.1016.2020.02398

Identifying Important Nodes in Multi-Relational Networks Based on Evidence Theory

LUO Hao YAN Guang-Hui ZHANG Meng BAO Jun-Bo LI Jun-Cheng LIU Ting

(School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070)

Abstract As a typical model of the real world, the multi-relational network has become one of hot topics in the field of network science. Identifying important nodes is one of the fundamental problem of network analysis, which is essential to understand the structure and dynamic characteristics of the complex networks. Lots of excellent studies carried on the importance of nodes in a single-relational network, but the research on important nodes in multi-relational networks

收稿日期:2019-09-05;在线发布日期:2020-02-16. 本课题得到国家自然科学基金(62062049)、教育部人文社会科学研究青年基金(20YJCZH212)和甘肃省自然科学基金(20JR5RA390)资助. 罗浩,博士研究生,中国计算机学会(CCF)会员,主要研究方向为数据挖掘和复杂网络分析. E-mail: luoh382@foxmail.com. 闫光辉(通信作者),博士,教授,中国计算机学会(CCF)会员,主要研究领域为数据库理论与系统、物联网工程与应用、数据挖掘、复杂网络分析等. E-mail: yanghacademic@163.com. 张萌,硕士,主要研究方向为数据挖掘和复杂网络分析. 包峻波,硕士,主要研究方向为数据挖掘和复杂网络分析. 李俊成,硕士,主要研究方向为数据挖掘和复杂网络分析. 刘婷,硕士,主要研究方向为数据挖掘和信息安全.

lacks systematic research. Trying to explore and extend the existing methods in single networks to multi-relational networks has become one of the hot issues. For multi-relational networks, this paper focused on the study of the identifying importance nodes considering the influence of centrality and transitivity. We created the multilayer network model to describe a multi-relational network and used the multilayer adjacency matrix to represent it. To quantify the transitivity in multi-relational networks, we proposed a method for calculating the local clustering coefficient according to the notion of clustering coefficient for multilayer networks. Ulteriorly, we proposed multiplex ClusterRank combining the multiplex degree centrality and extending ClusterRank in single-relational networks to multi-relational networks. Considering the influence of coupling information and the difference of transmission mechanism for multi-relational network on vital information of nodes, multiplex ClusterRank may have calculation errors in large-scale multi-relational networks, which may affect the accuracy of measuring essential nodes. Therefore, in this work we improved multiplex ClusterRank and proposed another more efficient method to identify vital nodes for multi-relational networks named multiplex evidential centrality, which fused the node degree centrality and local clustering coefficient by the Dempster-Shafer evidence theory. Numerous experiments were carried out on four real networks. The proposed method is evaluated from two aspects: robustness and vulnerability, transmission dynamics characteristics. Experimental results show that multiplex evidential centrality can get better results than multiplex degree centrality and multiplex eigenvector centrality which only focus on the structure of multiplex networks and multiplex ClusterRank does have calculation errors in large-scale networks, making it difficult to identify important nodes. The experimental results verified multiplex evidential centrality can effectively eliminate the influence of the coupling information and the transmission mechanism in multi-relational networks, and has lower time complexity than multiplex ClusterRank. In this paper, we not only provide new ideas and methods for identifying critical nodes for multi-relational networks but also expands the application of information fusion technology.

Keywords multi-relational network; multiplex network; important nodes; centrality; transitivity; local clustering coefficient; evidence theory

1 引言

近年来,随着复杂网络分析技术的迅猛发展,人们对现实世界中广泛存在的复杂系统已经有了从宏观到微观的递进认识.节点作为微观层面的基本单元,在现实网络中往往扮演着不同角色^[1-2],挖掘网络中的重要节点已成为复杂网络分析的基本问题之一.迄今发展起来的理论框架主要是面向单个网络结构.单个网络结构是一种简化的网络科学研究模型,而现实世界中相互关联、彼此依存的复杂系统普遍存在,刻画此类复杂系统的网络结构称之为网络的网络(Networks of Networks, NoN).相伴大数据时代的到来, NoN 逐渐成为网络科学领域富有挑战性的前沿课题^[3].多关系网络(multi-relational networks)作为 NoN 的一种典型形态,从复杂网络

分析的角度出发,尝试发展理论框架来研究多关系网络,并大量推广已有复杂网络分析方法到多关系网络具有重大理论意义和现实意义.

现阶段,国内外学者针对多关系网络的研究进行了一些探索,对于多关系网络的建模和表示,相对理想的方法是建立多层网络模型.国外学者已针对多层网络模型,提出了相关度量指标和分析方法,形成了较为完整的理论框架,为多关系网络的研究奠定了理论基础.但在实际应用层面,面向多关系网络进行重要节点挖掘尚缺少成熟的方法框架.因此,针对多关系网络的结构和动力学特点,设计相应的重要节点挖掘方法已成为当前节点重要性研究的重要任务.

本文针对无权无向多关系网络重要节点挖掘问题,通过建立多层网络模型对其进行表示,综合考虑节点中心性和传递性,提出了多重 ClusterRank

(Multiplex ClusterRank, MCR) 指标. 考虑到多关系网络传递机制和关系间信息的差异性, 引入 D-S 证据理论将多关系网络节点的度中心性和局部聚集系数进行信息融合, 提出了多重证据中心性 (Multiplex Evidential Centrality, MEC). 通过 4 个真实网络的实验结果表明, 由于多层网络传递机制差异性较强, 节点的中心性和传递性不具备相关性, 采用传统 ClusterRank 方法简单融合的多重 ClusterRank, 会造成节点部分信息的损失. 而采用 D-S 证据理论进行融合的多重证据中心性可以有效消除节点传递机制对节点重要程度造成的影响, 相比于只考虑结构中心性的度中心性方法而言, 能够更加准确地挖掘多关系网络中的重要节点, 特别是在大规模网络的重要节点挖掘中, 其准确度高于其它方法.

本文的主要贡献如下:

(1) 参考多层网络局部聚集系数的概念和计算方法, 对多关系网络进行了抽象和简化, 给出了多关系网络局部聚集系数的计算方法;

(2) 针对无权无向多关系网络, 统筹考虑节点的中心性和传递性, 提出了多重 ClusterRank 度量方法, 并对多关系网络节点的度中心性和局部聚集系数的相关关系进行了研究, 验证了“由于受多层网络传递机制和层间信息的影响, 二者不具备相关性, 若直接扩展单层网络 ClusterRank 的思想方法到多关系网络, 会对重要节点度量指标的准确度造成影响.”的猜想;

(3) 针对多重 ClusterRank 存在的缺陷, 进一步引入 D-S 证据理论, 将节点的度中心性和局部聚集系数融合, 提出了节点多重证据中心性的度量方法, 实验结果表明, 该方法可以消除多层网络传递机制和层间信息差异性对测量节点重要程度的影响, 相比于单纯关注中心性的度中心性, 综合考虑中心性和传递性的多重证据中心性能够更加精准地挖掘多关系网络中的重要节点, 使得多层网络的度量指标和 D-S 证据理论在挖掘重要节点的实际应用中得以扩展.

2 相关工作

2.1 多关系网络的表示

对多关系网络开展研究, 首要问题是建立合适的模型和表示方法. 鉴于目前在单个网络上已经建立了较为成熟的分析框架, 对多关系网络的简单处

理思路是采用“降维”思想, 围绕特定领域网络, 运用社会学、可拓学等其它学科的理论, 通过构建加权网络的方式将多关系网络转化为单关系网络^[4-5]; 或是借助矩阵分解、多子网复合等理论方法对已有网络进行时间划分或快照处理, 对单个网络分别研究后再进行复合^[6-7]. 然而, 通过降维转化为单关系网络的处理方法, 会丧失原始数据中所含的特定结构信息, 以及不同类型关系之间的关联信息, 不利于后续研究的开展. 此外, 还有一些研究围绕“全连通网络”, 运用超图理论展开分析^[4], 该类研究虽能够对网络进行融合, 保留了信息的完整性, 但“超图”模型通常只适用于刻画“全连通网络”, 局限性较强.

在计算机科学和计算机线性代数领域, 研究者们运用张量分解 (tensor decomposition)^[8-9] 和多路数据 (multiway data) 分析^[10] 的方法研究了不同类型的多层网络. 在物理学领域, 物理学家们围绕多层网络 (multilayer networks)^[11-14]、网络的网络 (Networks of Networks, NoN)^[15-16] 和节点着色网络 (node-colored networks), 已经做了许多重要的基础性工作. Boccaletti 等人^[17] 于 2014 年首次提出了多层网络的基本数学定义, 该定义统一了迄今文献上关于多层网络的不同提法. 经过近几年的发展, 多层网络模型已逐渐成熟, 形成了一些符合复杂网络基本特征和动力学机制的理论方法, 为后续研究提供了理论工具. 针对多关系网络, 可将每一类关系映射到一个网络层上, 同时考虑各类关系间的交互信息, 构建多层网络开展研究. 因此, 多层网络模型是表示多关系网络的理想形式, 将多层网络模型运用到多关系网络中, 在应用层面进行拓展, 是值得深入研究的一个问题.

2.2 重要节点度量指标

结构中心性和传递性是衡量节点重要程度的重要因素. 在单个网络的重要节点研究方面, 已经有了较完备和丰富的度量指标方法挖掘网络中的重要节点^[18]. 根据网络的局部特性划分, 可分为全局中心性、局部中心性、半局部中心性 3 类. 全局中心性的经典指标有介数中心性 (Betweenness Centrality, BC)^[19]、接近中心性 (Closeness Centrality, CC)^[20]、特征向量中心性 (Eigenvector Centrality, EC)^[21] 等, 局部中心性有度中心性 (Degree Centrality, DC)^[21]. 为平衡全局中心性的复杂度和局部中心性因素的单一性, Chen 等人^[22] 提出了一种半局部中心性 (Semi-Local Centrality, SLC), 通过考虑节点

的 4 阶邻居信息来度量节点重要程度, Gao 等人^[23]在其基础上通过引入 D-S 证据理论, 提出了一种证据半局部中心性 (Evidential Semi-local Centrality, ESC), 但两者都只关注了中心性, 而忽略了传递性在度量重要节点中的关键作用. Chen 等人^[24]综合考虑了二阶邻居和局部集聚特性, 提出了 ClusterRank 指标, 该指标在运行效率和准确度上都要优于前者.

2.3 多层网络的度量指标

面向多层网络, 如何借用已有的物理量去描述其基本拓扑结构和性质还面临巨大的挑战, 将单个网络中已有的成熟工具和方法扩展到多层网络领域成为当前复杂网络研究的一大热门方向. 最初的做法是将不同网络层的节点和连边以不同权重的方式进行合并, 从而将多层网络转化成为单层网络, 再利用传统物理量分析网络的各种结构属性. 这种方法虽然简单, 但存在很多问题, 因为在将多层网络合并成单层网络时, 一些不同网络层之间的重要信息会全部丢失, 同时还存在如何设定各种权重的问题^[25]. 因此, 发现一套适合多层网络理论框架的必要性开始凸显, 国内外学者围绕设计多层网络的度量指标做了大量研究工作, 也取得了一些成果, 多层网络的中心性、路径长度、聚集系数等概念被相继提出, 为后续研究奠定了基础. 与单层网络的度量指标不同, 多层网络的一些度量指标是在特定条件和应用场景下提出的, 尚未形成像单层网络众多经典方法一样的公认指标. 值得一提的是, Battiston 等人^[26]在向量的基础上提出的多层网络节点度 (degree)、节点强度 (strength) 以及 Cozzo 等人^[27]基于多层网络的传递性提出的多层网络聚集系数的概念都是从多层网络的基本性质和动力学特性出发的, 具有一定的普适性, 成为目前比较经典的多层网络度量方法. 特别是在节点的传递性方面, 在实际数据集上的实验分析表明, 在社会网络和交通网络中层内聚集性和层间聚集性的差异是不同的, 这实际上反映了在这些网络中节点传递性的起源机制大不一样. 他们提供的聚集系数定义表达式为不同类型的网络传递性研究提供了新的思路.

3 基本概念

3.1 多关系网络的矩阵表示

按照 Boccaletti 等人^[17]提出的多层网络定义, 多关系网络可用一种特殊的多层网络进行构建, 称

之为多重网络 (Multiplex Network), 其唯一的层间连接模式是来自不同层对应节点之间的连接.

图 1 描述了具有 10 个节点, 3 个网络层的多层网络和多重网络示意图. 一般地, 对于 N 个节点, M 层的无权无向多关系网络, 可用 $MN \times MN$ 的邻接矩阵进行表示, 这对于计算算法的实现是十分有用的. 进一步, 可根据连接方式的不同将多关系网络分解为层内网络和层间网络, 分别用层内邻接矩阵和层间邻接矩阵进行表示^[28], 定义如下.

定义 1. 层内邻接矩阵. 多关系网络第 α 类关系所对应的层内连接可由一个邻接矩阵来表示, 记为 $\mathbf{A}^{[\alpha]} \in \mathbb{R}^{N \times N}$, 则多关系网络的层内邻接矩阵定义为

$$\mathbf{A}_{\text{MI}} = \bigoplus_{k=1}^M \mathbf{A}^{[\alpha]} \in \mathbb{R}^{MN \times MN} \quad (1)$$

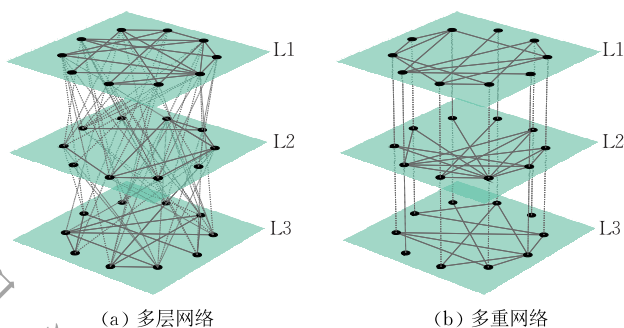


图 1 多层网络与多重网络原理图

定义 2. 层间邻接矩阵. 多关系网络所对应的层间连接可由一个加权邻接矩阵来表示, 记为 $\mathbf{W}_{\text{C}} \in \mathbb{R}^{M \times M}$, 取单位矩阵 $\mathbf{I}_{\text{MC}} \in \mathbb{R}^{N \times N}$, 则多关系网络的层间邻接矩阵定义为

$$\mathbf{W}_{\text{MC}} = \mathbf{W}_{\text{C}} \otimes \mathbf{I}_{\text{MC}} \in \mathbb{R}^{MN \times MN} \quad (2)$$

根据层内邻接矩阵和层间邻接矩阵, 可以很方便地得到多关系网络的邻接矩阵.

定义 3. 多关系网络的邻接矩阵. 令 $\mathbf{A}_{\text{MI}}$ 为多关系网络层内邻接矩阵, $\mathbf{W}_{\text{MC}}$ 为层间邻接矩阵, 则多关系网络的邻接矩阵:

$$\mathbf{A}_{\text{M}} = \mathbf{A}_{\text{MI}} + \mathbf{W}_{\text{MC}} \in \mathbb{R}^{MN \times MN} \quad (3)$$

3.2 多关系网络的度中心性

根据 Battiston 等人^[26]提出的多层网络节点度和节点强度的概念, 可得到多关系网络节点的度中心性 (Multiplex Degree Centrality, MDC) 定义.

定义 4. 多关系网络节点的度中心性. 假设多关系网络 $G_{\text{M}} = (V, E_{\text{M}})$ 包含 M 种关系, 节点数为 N , 则节点 $v_i \in V (i=1, 2, \dots, N)$ 的度中心性向量可表示为: $\mathbf{K}_i = (K_i^{[1]}, K_i^{[2]}, \dots, K_i^{[M]})$, 其中 $K_i^{[\alpha]}$ 为节点 v_i 在层 α 中的度, 定义为

$$K_i^{[\alpha]} = \sum_j \sum_{\beta} a_{ij}^{[\alpha, \beta]} \tag{4}$$

进一步,我们可以在多关系网络上得到相应的节点度、平均度、度的二阶矩的定义.

节点度(node degree):

$$K_i = \sum_{\alpha} K_i^{[\alpha]} \tag{5}$$

平均度(mean degree):

$$\langle K \rangle = \frac{\sum_i K_i}{N} \tag{6}$$

度的二阶矩(the second moment of the degree):

$$\langle K^2 \rangle = \frac{\sum_i (K_i)^2}{N} \tag{7}$$

3.3 多关系网络的局部聚集系数

Cozzo 等人^[27]从路径和环的概念出发,提出了

多层网络局部聚集系数和全局聚集系数的概念. 在单层网络中,可用邻接矩阵来计算网络中路径和环的数量^[29-30]. 类似地,将多层网络中的行走称之为“超行走(supra-walk)”,并通过层内邻接矩阵和层间邻接矩阵来计算多层网络的聚集系数. 按照 Cozzo 等人提出的多层网络聚集系数的计算方法,本文给出了多关系网络局部聚集系数(Multiplex Local Clustering Coefficient, MLCC)的概念和计算方法. 由 3.1 节可知,本文运用多重网络模型来表示多关系网络,则在多关系网络中,通过“超行走”得到的三元闭环有且仅有图 2 所示的 5 种基本形式,其中较大节点代表闭环的起始节点,层内路径用实线表示,层间路径用虚线表示,较粗实线代表层内第二步行走路径. 通过“超行走”所构成的三元闭环可通过将各网络层聚合到一层得到,由此可得到多关系网络节点聚集系数的定义.

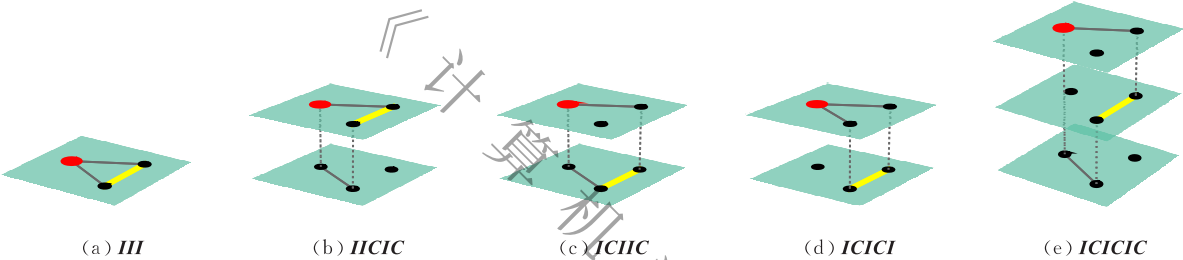


图 2 多关系网络基本闭环结构示意图

定义 5. 多关系网络的局部聚集系数. 令 M, I, C 分别表示多关系网络的邻接矩阵、层内邻接矩阵、层间邻接矩阵,

$$\Omega = \{III, IICIC, ICIC, ICICI, ICICIC\}$$

为多关系网络中基本三元闭环结构集合, ξ 为集合 Ω 中的元素,

$$\hat{\Omega} = \{IFI, IFCIC, ICFIC, ICFIC, ICFIC\}$$

为在构成基本三元闭环结构的“超行走”过程中,层内第 2 步发生在一个完全图上的闭环结构集合. 其中, $F = \bigoplus_{\alpha} F^{[\alpha]}, F^{[\alpha]} = J^{[\alpha]} - I^{[\alpha]}$ 为层 α 中一全连接网络的邻接矩阵, $\hat{\xi}$ 表示 $\hat{\Omega}$ 中的元素,则多关系网络节点 v_i 的局部聚集系数可定义为

$$C_{M,i} = \frac{\sum_{\xi \in \Omega} \omega_{\xi}(\xi)_{ii}}{\sum_{\hat{\xi} \in \hat{\Omega}} \omega_{\hat{\xi}}(\hat{\xi})_{ii}} \tag{8}$$

式(8)中, ω_{ξ} 和 $\omega_{\hat{\xi}}$ 代表“超行走”分别跨越 1 层、2 层和 3 层所占的贡献,在无特定背景情况下,一般设 $\omega_{\xi} = \omega_{\hat{\xi}} = 1/3$.

3.4 D-S 证据理论

Dempster 和 Shafer 建立的 D-S 证据理论,是

在概率论基础上扩展的 1 套数学理论,适用于人工智能、模式识别等领域的实际问题. D-S 证据理论涉及确定辨识框架(frame of discernment),计算基本概率指派(Basic Probability Assignment, BPA)以及证据合成等基本流程,本节简要介绍其基本概念和合成方法^[31].

定义 6. 辨识框架. 在 D-S 证据理论中,所考察判断的事物(或对象)的全体可用一个由 N 个两两互斥元素组成的有限完备集合来表示,称之为辨识框架(或称为假设空间). 记为

$$\Psi = \{\psi_1, \psi_2, \dots, \psi_N\}.$$

Ψ 的幂集 2^{Ψ} 记为

$$2^{\Psi} = \{\emptyset, \psi_1, \dots, \psi_N, \psi_1 \cup \psi_2, \dots, \psi_1 \cup \psi_2 \cup \psi_3, \dots, \Psi\}.$$

定义 7. 基本概率指派. 设 Ψ 为辨识框架, Ψ 的幂集 2^{Ψ} 构成命题集合,对于 $\forall A \subseteq \Psi$,若函数 $m: 2^{\Psi} \rightarrow [0, 1]$ 满足以下两个条件:

$$\begin{cases} m(\emptyset) = 0 \\ \sum_{A \subseteq \Psi} m(A) = 1 \end{cases} \tag{9}$$

则称 m 为基本概率指派, $m(A)$ 为命题 A 的基本概率数,被视为准确分配给 A 的信度.

定义 8. Dempster-Shafer 组合规则. 设 m_1 和 m_2 为 2 组基本概率指派, 对应的命题分别为 $A_1, A_2, \dots, A_k$ 和 $B_1, B_2, \dots, B_l$, 用 m 表示 m_1 和 m_2 组合后的新证据, 则 Dempster-Shafer 组合规则为

$$\begin{cases} m(\emptyset) = 0, \\ m(A) = \frac{1}{1-k} \sum_{A_i \cap B_j = A} m_1(A_i) m_2(B_j) \end{cases} \quad (10)$$

其中, $k = \sum_{A_i \cap B_j = \emptyset} m_1(A_i) m_2(B_j)$ 称为冲突系数, 用于衡量证据焦元间的冲突程度, k 越大, 则冲突越大. 由式(10)可知 $k \neq 1$, 当 $k = 1$ 时, 组合规则无法使用, 即 BPA 存在矛盾.

4 多重证据中心性

4.1 多重 ClusterRank 度量指标

中心性和传递性作为影响节点重要性的主要因素, 可分别用节点的度中心性和局部聚集系数进行度量, 其中节点的度中心性与节点的重要程度呈正相关、节点的局部聚集系数与其重要程度呈负相关. 在单个网络中, Chen 等人^[24]提出的 ClusterRank 算法同时考虑了中心性和传递性对节点的影响, 取得了较好的度量效果. 根据 ClusterRank 算法的思想, 在多关系网络上进行推广, 首先提出了多重 ClusterRank (Multiplex ClusterRank, MCR).

定义 9. 多重 ClusterRank. 在多关系网络 $G_M = (V, E_M)$ 中, 记 K_i 为节点 v_i 的度中心性、 $C_{M,i}$ 为节点 v_i 的局部聚集系数, 令 $f(C_{M,i}) = 10^{-C_{M,i}}$ 为 $C_{M,i}$ 的非线性负相关函数, 则节点 v_i 的多重 ClusterRank 度量值为

$$v_{MCR,i} = \frac{f(C_{M,i})}{K_i} \sum_{\alpha} R_i^{[\alpha]} \quad (11)$$

其中,

$$R_i^{[\alpha]} = \sum_{j \in \Gamma_i} [(K_i^{[\alpha]} + 1)] \quad (12)$$

为层 α 内, 节点 v_i 的二阶邻居节点的连接强度, Γ_i 为节点 v_i 二阶邻居节点的集合.

4.2 多重证据中心性

考虑到多关系网络传递机制和层间信息的差异, 猜想节点的中心性和传递性之间可能不具备相关性, 会对多重 ClusterRank 的计算结果造成误差, 从而降低评估节点重要程度的准确性.

鉴于此, 运用 D-S 证据理论^[23,32]对多重 ClusterRank 指标进行改进, 对节点的度中心性和局部聚集系数进行信息融合, 进而得到多重证据中心性

(Multiplex Evidential Centrality, MEC). 根据 3.4 节 D-S 证据理论基本流程, 给出多重证据中心性的计算步骤:

① 计算相关指标

计算节点度中心性向量 $\mathbf{K}$ 和局部聚集系数向量 $\mathbf{C}_M$, 记 K_i 和 $C_{M,i}$ 分别为 $\mathbf{K}$ 和 $\mathbf{C}_M$ 中对应第 i 个位置的元素值, 即节点 v_i 的度中心性和局部聚集系数. 分别计算 $f(C_{M,i}) = 10^{-C_{M,i}}$, 所得结果向量记为 $\mathbf{F}$. 统计 $\mathbf{K}$ 和 $\mathbf{F}$ 的最大值和最小值, 分别为 K_M, K_m, F_M, F_m . 通常, $K_M \neq K_m, F_M \neq F_m$.

② 确定节点的辨识框架 Ψ

简单而言, 网络中的节点要么重要, 要么不重要, 可定义辨识框架 $\Psi = \{I, NI\}$, 其中 I 表示节点重要, NI 表示节点不重要, 二者互斥且 Ψ 为有限完备集合.

③ 构造基本概率指派函数

令 $m_{K,I}(i)$ 和 $m_{K,NI}(i)$ 分别为在 K_i 下节点 v_i 重要和不重要的概率, $m_{F,I}(i)$ 和 $m_{F,NI}(i)$ 分别代表在 $f(C_{M,i})$ 下节点 v_i 重要和不重要的概率. 在无特定条件下, 通常约定该概率指派服从均匀分布, 则 BPA 函数可分别定义为

$$m_{K,I}(i) = \frac{|K_i - K_m|}{K_M - K_m + 2\mu} \quad (13)$$

$$m_{K,NI}(i) = \frac{|K_i - K_M|}{K_M - K_m + 2\mu} \quad (14)$$

$$m_{F,I}(i) = \frac{|F_i - F_m|}{F_M - F_m + 2\lambda} \quad (15)$$

$$m_{F,NI}(i) = \frac{|F_i - F_M|}{F_M - F_m + 2\lambda} \quad (16)$$

μ 和 λ 的取值对节点排序没有影响, 通常取 $\mu, \lambda \in (0, 1)$.

④ 进行 D-S 证据组合

首先定义一个新的指标, 记为

$$M(i) = \{m_1(i), m_{NI}(i), m_{\Psi}(i)\} \quad (17)$$

根据式(10), 可分别计算 $m_1(i), m_{NI}(i), m_{\Psi}(i)$. 其中 $m_1(i)$ 表示节点重要的概率, $m_{NI}(i)$ 表示节点不重要的概率, $m_{\Psi}(i)$ 表示节点重要或者不重要的概率.

⑤ 计算 MEC.

根据贝叶斯理论, 在无其它先验知识的情况下, 将 $m_{\Psi}(i)$ 的值等可能分配给 $m_1(i)$ 和 $m_{NI}(i)$, 进一步修正节点 v_i 重要的概率 $M_I(i)$ 和不重要的概率 $M_{NI}(i)$, 即

$$M_I(i) = m_1(i) + \frac{m_{\Psi}(i)}{2} \quad (18)$$

$$M_{NI}(i)=m_{NI}(i)+\frac{m_{\Psi}(i)}{2}$$

(19)

由式(18)和式(19)可得, $M_1(i)$ 越大, 或者 $M_{NI}(i)$ 越小, 均说明节点 v_i 越重要. 鉴于此, 定义节点 v_i 的多重证据中心性为

$$\begin{aligned} v_{MEC,i} &= M_1(i) - M_{NI}(i) \\ &= m_1(i) - m_{NI}(i) \end{aligned}$$

(20)

进一步归一化处理可得:

$$v_{MEC,i}^* = \frac{|v_{MEC,\min}| + v_{MEC,i}}{\sum_{i=1}^N \{|v_{MEC,\min}| + v_{MEC,i}\}}$$

(21)

4.3 时空复杂度分析

多关系网络中, 记节点数为 n , 关系数为 m .

(1) 时间复杂度分析

多关系网络度中心性的时间复杂度为 $O(mn)$, 节点的局部聚集系数时间复杂度为 $O(m^3n^3)$. 在已知节点度中心性和局部聚集系数的情况下, 多重 ClusterRank 和多重证据中心性由于综合考虑了中心性和传递性, 故二者的时间复杂度均依赖于度中心性和局部聚集系数, 其中多重 ClusterRank 由于考虑了节点的一阶邻居信息, 其时间复杂度为 $O(mn\langle K \rangle)$, 而多重证据中心性将层间信息通过 D-S 证据理论进行了信息融合, 仅关注了节点本身, 故其时间复杂度为 $O(n)$. 综合考虑节点度中心性和节点聚集系数的时间复杂度后, 二者的时间复杂度均为 $O(m^3n^3)$.

(2) 空间复杂度分析

鉴于通过构建 $mn \times mn$ 的邻接矩阵对多关系网络进行表征, 故其空间复杂度为 $O(mn \times mn)$. 在具体实现过程中, 考虑到邻接矩阵的稀疏性, 采用稀疏矩阵对其进行处理, 可以有效降低多关系网络的空间复杂度.

5 实验与讨论

相比于单关系网络, 目前适用于无权无向多关系网络的重要节点挖掘方法相对较少, 多层网络的度中心性是目前公认度比较高的适用于无权无向多关系网络重要节点的经典度量指标.

为从运行时间和重要节点度量准确性两方面评测所提出的方法, 验证 4.2 节提出的“由于受多关系网络传递机制的差异, 节点的中心性和传递性间不具备相关性, 会造成多重 ClusterRank 的计算误差, 从而降低评估节点重要程度的准确性.”的猜想, 以及与只关注节点结构的度中心性相比, 融合节点传递性的多重证据中心性更能准确识别网络中的重要节点, 选取了网络类别、网络规模、关系数量、关系类别各不相同的 4 个真实网络数据集进行了 5 组不同实验, 运用重要节点挖掘研究中最常采用的 SIR 传播模型以及鲁棒性和脆弱性方法进行评价. 同时, 除选择多层度中心性作为对比方法外, 另外选取了适用于无权无向网络的多重特征向量中心性 (Multiplex Eigenvector Centrality, MEVC) 作为对比, 用以说明提出算法的有效性. 多重特征向量中心性定义为^[33]

$$M_{j\beta}^{ia} \boldsymbol{\Theta}_{ia} = \lambda_1 \boldsymbol{\Theta}_{j\beta}$$

(22)

$$v_{MEVC,i} = \boldsymbol{\Theta}_{ia} \boldsymbol{u}^a$$

(23)

其中 $M_{j\beta}^{ia}$ 为多层邻接张量, λ_1 为最大特征值, $\boldsymbol{\Theta}_{ia}$ 为对应的特征向量, $\boldsymbol{u}^a$ 为元素全为 1 的向量.

实验平台为 Intel Core i9-9900K 3.6 GHz, 64 GB RAM, 操作系统为 Windows 10, 代码由 Python 实现. 选取数据集分别为 CKM Physicians Innovation^[34]、CS Aarhus^[35]、EU Air Transportation^[36]、Moscow Athletics 2013^[37], 数据集的基本统计信息如表 1 所示.

表 1 实验数据集基本统计信息

数据集名称	层数	节点数	边数	网络类型	平均度	度的二阶距
CKM Physicians Innovation	3	246	1551	社会网络	18.60	405.40
CS Aarhus	5	61	620	社会网络	40.33	1752.20
Eu Air Transportation	37	450	3588	交通网络	1347.94	1.8×10^6
Moscow Athletics 2013	3	88804	210250	社会网络	10.69	3286.23

5.1 运行时间评价

通过统计 4 个真实网络数据集上各算法的实际运行时间, 来比较提出方法的运行速率. 记 $t_K, t_L, t_{MCR}, t_{MEC}, t_E$ 分别为多关系网络中计算度中心性、局部聚集系数、多重 ClusterRank、多重证据中心性和特征向量中心性的运行时间. 根据 4.3 节对各算法

的时间复杂度分析可知, 多重 ClusterRank 和多重证据中心性的运行时间依赖于 t_K 和 t_L , 即

$$t_{MCR} = t_K + t_L + t_{MCR}^*$$

(24)

$$t_{MEC} = t_K + t_L + t_{MEC}^*$$

(25)

其中, t_{MCR}^* 和 t_{MEC}^* 分别为单独计算多重 ClusterRank 和多重证据中心性的运行时间.

由于 4 个网络规模不同,我们通过统计各方法运行时间占总运行时间的比重,来说明算法运行时间的长短,即

$$P(t)=\frac{t_i}{\sum_j t_j}$$

(26)

5 种方法在 4 个网络上的运行时间统计结果如图 3 所示,可以看出,在整个运行过程中,度中心性和特征向量中心性的计算时间很短,基本可以忽略不计,计算局部聚集系数的运行时间相对较长,在计算多重 ClusterRank 和多重证据中心性的过程中,主要时间消耗在计算局部聚集系数的环节上,而总

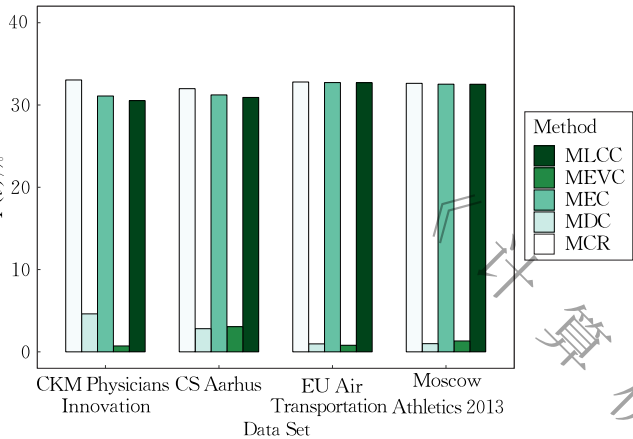


图 3 各方法运行时间统计图

体运行时间和局部聚集系数的运算时间差异不大。

图 4 显示了 4 个网络中 t_{MCR}^* 和 t_{MEC}^* 的运行时间统计,可以看到,在单独计算二者的运行时间上,多重证据中心性要比多重 ClusterRank 用时少得多。综上,实验结果与 4.2 节中时间复杂度的分析结论是一致的,在运行时间上,多重证据中心性要优于多重 ClusterRank。

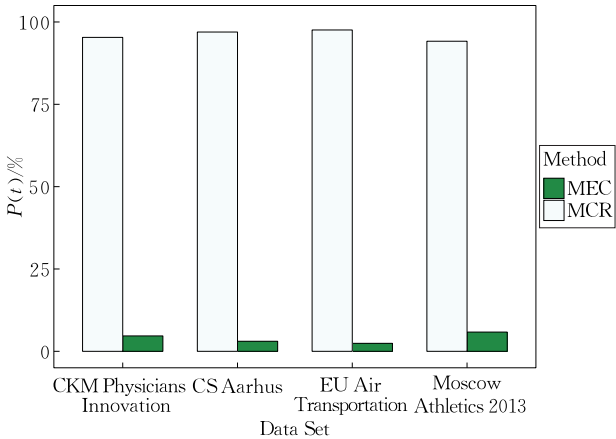


图 4 MCR 和 MEC 运行时间统计图

5.2 度量准确性评价

表 2 列出了在 4 个真实网络数据集中,运用 4 种方法分别对节点重要程度进行度量,各自得到重要程度最高的 10 个节点的标号。

表 2 4 种方法得到前 10 个重要节点标号

CKM Physicians Innovation				CS Aarhus				EU Air Transportation				Moscow Athletics 2013			
MDC	MEVC	MCR	MEC	MDC	MEVC	MCR	MEC	MDC	MEVC	MCR	MEC	MDC	MEVC	MCR	MEC
v122	v41	v15	v122	v44	v51	v7	v44	v15	v1	v50	v15	v22	v29604	v13892	v22
v29	v47	v98	v29	v51	v19	v44	v51	v50	v26	v15	v50	v727	v29616	v12858	v727
v141	v123	v93	v141	v23	v53	v34	v23	v38	v2	v40	v38	v7	v29605	v74738	v7
v36	v49	v29	v36	v34	v56	v21	v7	v40	v14	v38	v40	v42	v29609	v25936	v42
v216	v129	v36	v55	v7	v58	v51	v34	v2	v50	v2	v2	v605	v8	v17817	v605
v100	v205	v8	v100	v46	v50	v26	v46	v252	v74	v83	v252	v17	v29611	v19867	v17
v46	v131	v24	v46	v26	v7	v31	v26	v64	v366	v7	v12	v9	v29878	v12942	v9
v191	v211	v46	v31	v31	v18	v19	v31	v83	v293	v166	v64	v810	v29803	v19882	v810
v55	v125	v11	v94	v33	v54	v13	v13	v12	v62	v64	v83	v315	v243	v46251	v315
v231	v44	v28	v27	v13	v6	v46	v8	v7	v439	v12	v7	v327	v30047	v51112	v327

为验证 4.2 节中的猜想,首先对度中心性和局部聚集系数的相关性进行了讨论,通过计算 4 个网络中度中心性向量和局部聚集系数向量的 Kendall 相关系数^[22]来说明二者的相关性,计算结果统一取 4 位有效数字,如表 3 所示。可以看出,在 Moscow Athletics 2013 中,度中心性与局部聚集系数之间呈现出较弱相关性外,在其它 3 个数据集上,二者均未呈现出明显的相关关系,所得结论与 4.2 节中猜想一致。

表 3 度与局部系数相关性

数据集	Kendall 相关系数
CKM Physicians Innovation	0. 0175
CS Aarhus	−0. 2650
EU Air Transportation	0. 2958
Moscow Athletics 2013	0. 5278

通常,对重要节点挖掘方法的评价一般采用基于网络的传播动力学模型和基于网络的鲁棒性和脆弱性 2 种方法^[39],分别在 4 个数据集上进行了 5 组不同实验,从网络的传播特性和鲁棒性及脆弱性

2 个方面对所提出的方法进行了全面评价.

前 4 组实验从网络的传播特性方面对所提出方法进行了评价. 基于网络的传播动力学模型对重要节点度量方法评价, 常采用 SIR 传播模型. 在 SIR 模型^[17, 38-41]中, 对重要节点的评判标准一般有 2 项: 一种为传播速率, 可从传播初始阶段感染态 $I(t)$ 和恢复态 $R(t)$ 的数量和达到稳态时所用时间两方面来评判, 数量越大、所用时间越小, 说明该节点越重要; 另一种为节点的平均传播范围, 即达到稳态时感染态 $I(t)$ 和恢复态 $R(t)$ 的数量之和, 数量越大, 说明该节点越重要. 为方便说明, 令第 t 个时间迭代步时, 感染态和恢复态的数量之和记为 $F(t)$, 感染态的平均数量记为 $F_i(t)$, 达到稳态时的数量记为 $F(t_c)$.

由于网络规模不同, SIR 模型的传染概率和恢复概率依赖于各层层内网络的拓扑结构. 通常, 传染概率的设置依赖于各层内网络的渐近传播阈值 $\beta_c^{[a]}$. 渐进传播阈值为传播过程中应取的最小传播概率^[38], 定义为

$$\beta_c^{[a]} = \frac{\langle K^{[a]} \rangle}{\langle (K^{[a]})^2 \rangle - \langle K^{[a]} \rangle} \quad (27)$$

根据贝叶斯理论和条件概率, 各层传染概率设置为

$$p_c^{[a]} = \frac{\lambda_i \beta_c^{[a]}}{N_R} \quad (28)$$

其中 λ_i 为放大系数, N_R 为多关系网络的关系数. 恢复概率通常依赖于各层的平均度, 类似地, 各层恢复概率设置为

$$p_r^{[a]} = \frac{\lambda_r}{N_R \langle K^{[a]} \rangle} \quad (29)$$

由于传播过程是依概率进行感染传播, 即使 2 组实验条件完全相同, 最终得到的结果也会存在一定差异. 为此, 特选取 100 次独立实验所得结果的期望值作为评价结果.

实验 1.

本实验主要通过观测各节点传播初期的 $F(t)$ 与各方方法测量值之间的相关关系, 来说明节点的传播能力. 本实验设置 $t=10$. 在理想状态下, $F(t)$ 与对应的测量值应呈现正相关关系.

在 4 个数据集上的实验结果如图 5 所示. 图 5

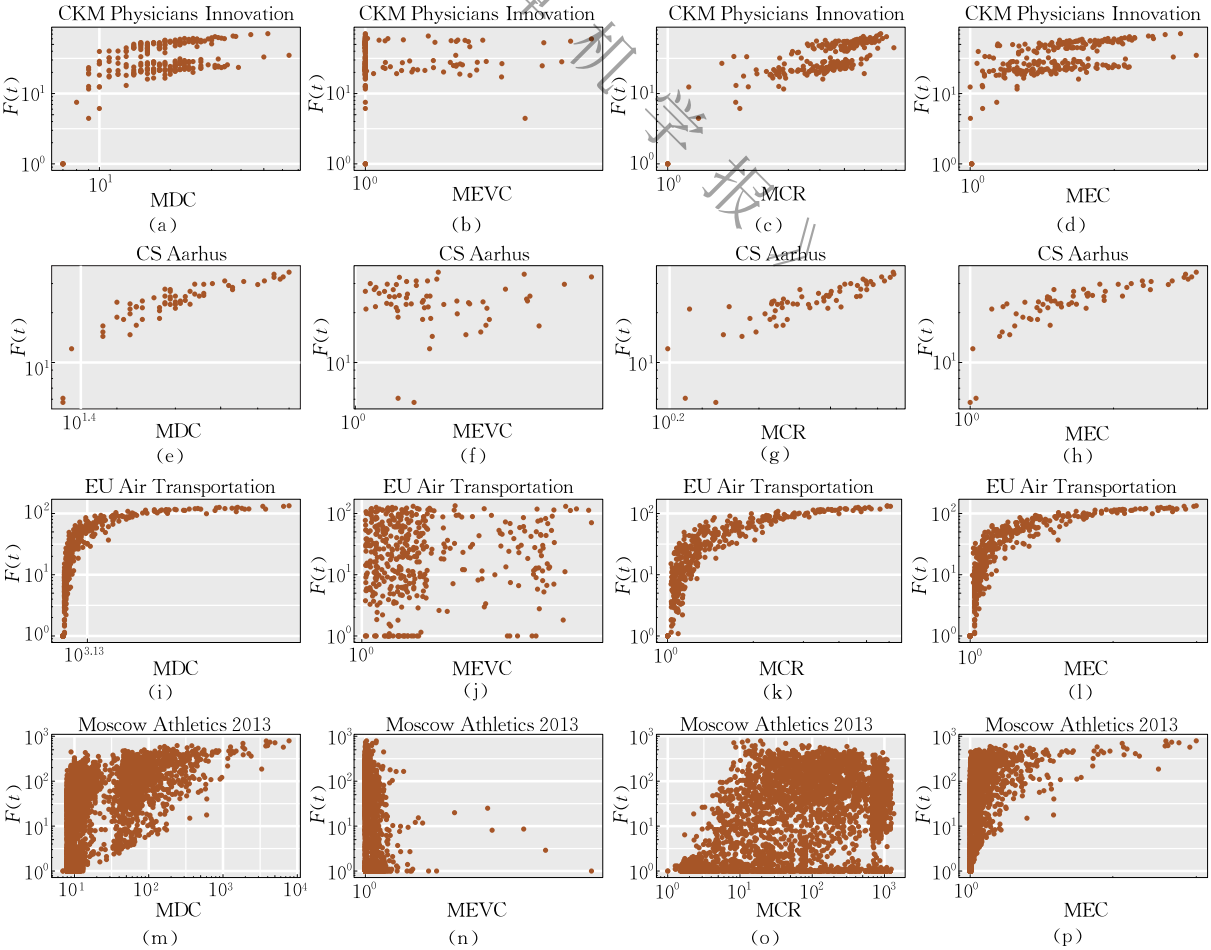


图 5 $t=10$ 时, $F(t)$ 与各方方法测量值间的相关关系

中,每行子图分别对应同一数据集下的不同方法,坐标轴采用双对数坐标,横轴表示每一种方法下各节点的测量值,纵轴表示各节点作为传染源节点时,所对应的 $F(t) (t=10)$. Moscow Athletics 2013 由于规模较大,故选取各方法得到的前 1000 个节点作为实验节点.

实验结果表明:①在 4 个网络中,多层度中心和多重证据中心性都能呈现出较好的正相关关系,多重证据中心性要优于多层度中心性,特别是在规模较大的 Moscow Athletics 2013 中,优势更加明显;②多重 ClusterRank 在规模较小的 3 个网络中,能够呈现出较好的正相关关系,但在规模较大的 Moscow Athletics 2013 网络中, $F(t)$ 与多重 ClusterRank 没有呈现出相关关系,一些测量值较低的节点反而 $F(t)$ 越大;③相比于其它 3 种方法,

多层特征向量中心性总体表现最差,除在 CS Aarhus 和 CKM Physicians Innovation 中呈现出较弱的正相关关系外,在其它网络中均没有呈现出明显的相关关系.

实验 2.

为了进一步验证各方法的性能,本实验研究了按不同度量方法对节点的重要程度进行排序后,节点排序与所对应的平均感染数量 $\langle F(t) \rangle (t=10)$ 之间的关系.性能较好的度量指标,其图像应该呈现递增减关系,即随着节点重要程度的降低,对应的受感染节点的平均数会减少.为了对比,本实验特意增加了随机排序这种极端情况.实验结果如图 6 所示,其中横轴代表各节点排序,纵轴代表对应的受感染节点平均数量.类似地,在 Moscow Athletics 2013 中,分别选取排序前 1000 名的节点作为实验节点.

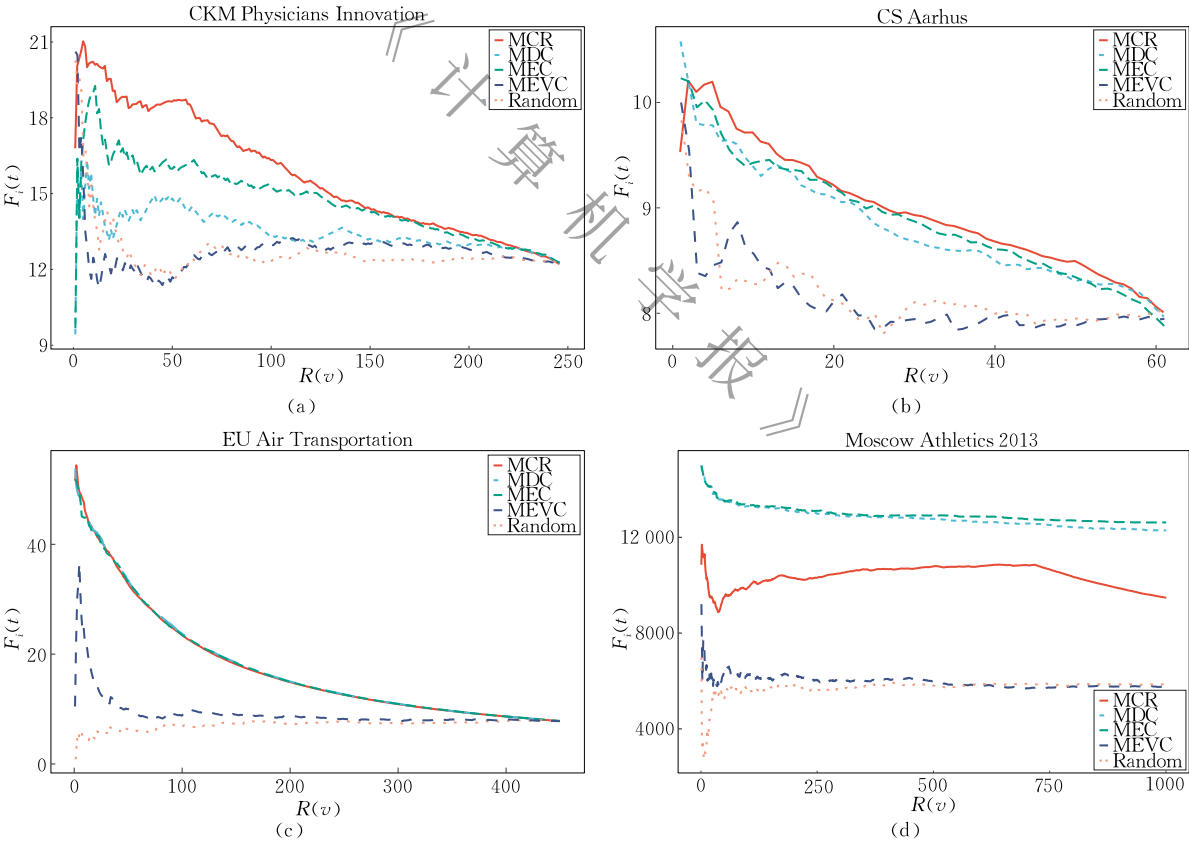


图 6 通过 5 种方法得到的节点排序对应 $t=10$ 时的平均感染度

显然,随机排序的情况最差.在 CS Aarhus 和 CKM Physicians Innovation 中,4 种方法均能呈现出递增减趋势,从传播范围上分析,多重 ClusterRank 得到的结果要优于其它 3 种方法,特别是在 CKM Physicians Innovation 中,其它 3 种方法都表现出了幅度较大的波动,而多重 ClusterRank 呈现出的波动

较小.在 Moscow Athletics 2013 中,多重证据中心性和多层度中心性能够呈现出较理想的递增减趋势,而其它两种方法表现较差.从传播范围上分析,多重证据中心性要优于多层度中心性.在 EU Air Transportation 中,多层特征向量中心性表现最差,其它三种方法表现相当,都能够呈现出较好的递增减趋势.

实验 3.

本实验通过观测传播达到稳态时所用的时间和 $F(t_c)$, 从传播速率和传播范围两方面来评测各方法的测量性能. 在 4 个网络中, 按照各方法对节点的重要程度进行排序, 分别选取前 1% 的节点作为初始传染源节点, 统计 100 个时间迭代步的 $F(t)$, 并绘制 SIR 时间增长曲线, 其中横轴代表时间迭代步数, 纵轴代表对应的 $F(t)$, 结果如图 7 所示. 4 个网络上的

实验结果显示: ① 无论是传播时间还是传播范围, MEC 都能表现出较好结果, 特别是在大规模网络中尤为明显; ② MCR 在规模较小的网络中表现较为理想, 但在规模较大的 Moscow Athletics 2013 数据集中, 其得到的结果最差, 从而进一步验证了之前的猜想; ③ 多层度中心性在规模较小的网络中所得到的结果较好, 但在大规模网络中, 与 MEC 的差距明显增大, 从而说明考虑网络传递性因素是非常必要的.

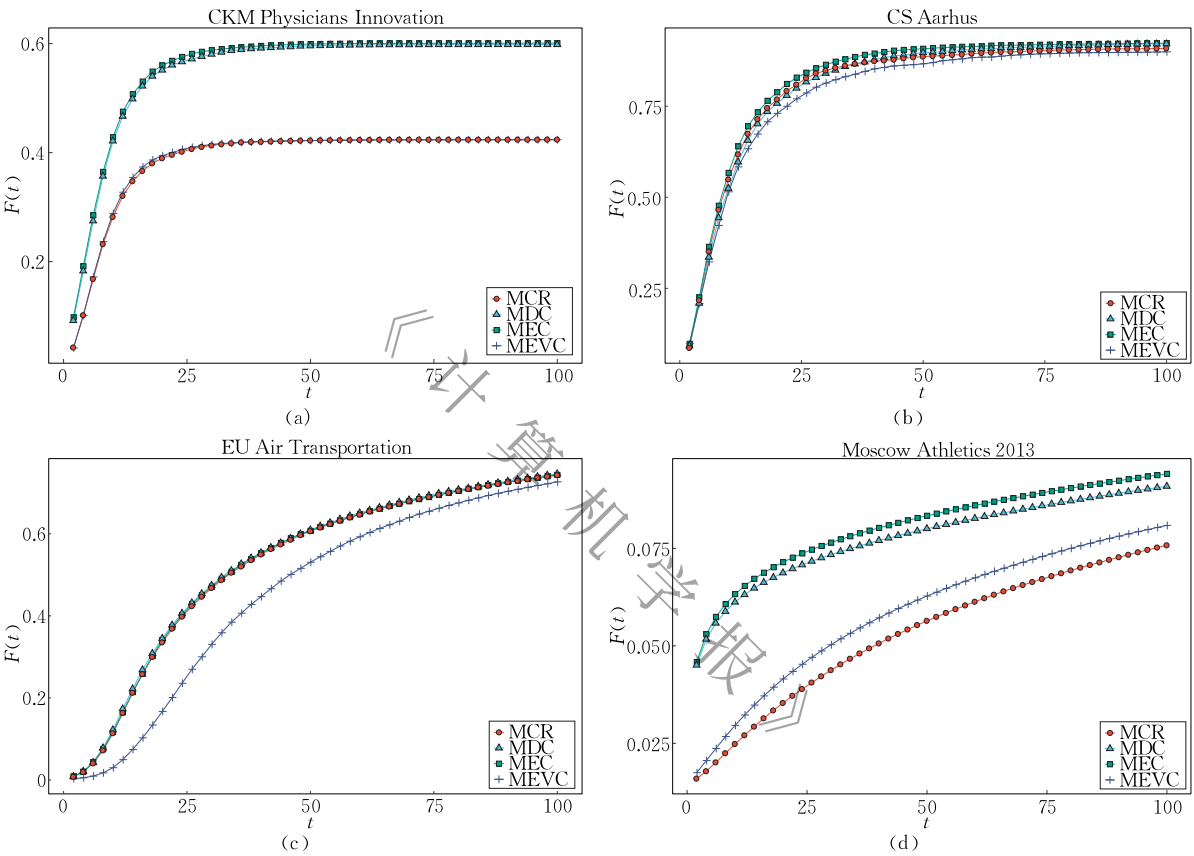


图 7 前 1% 的节点作为初始传染源节点时的 SIR 传播曲线

实验 4.

本实验通过比较节点在 SIR 模型上的传播能力与各测量值之间的相关关系, 来评测各方法的性能. 我们对各节点在 SIR 模型中得到的稳态值 $F(t_c)$ 进行排序, 并计算了其与各方法之间的肯德尔相关系数 (τ). 图 8 描述了在 4 个网络中, 各节点在 SIR 模型中的传播能力与 4 种重要节点测量方法间的相关关系, 图中每一行子图对应同一数据集的不同方法, 坐标轴采用双对数坐标, 横轴表示各节点在对应方法下的测量值, 纵轴表示各节点在 SIR 模型中得到的稳态值 $F(t_c)$, 每一个数据点代表网络中的

一个节点, 节点的灰度表示在传播过程中, $t=10$ 时的传播能力, 即 $F(10)$, 图中标识的 τ 值即为相应的肯德尔相关系数值. 在 CS Aarhus 和 CKM Physicians Innovation 中, 四种方法与节点传播能力之间表现出的相关性都较弱, 多重 ClusterRank 和多重证据中心性要优于其它 2 种方法. 在 Moscow Athletics 2013 中, 多重证据中心性和多层度中心性能够呈现出较好的相关关系, 多重证据中心性仍然表现最好, 而其它两种方法呈现出的相关性较弱. 在 EU Air Transportation 中, 除多层特征中心性外, 其它 3 种方法都能呈现出较强的相关性.

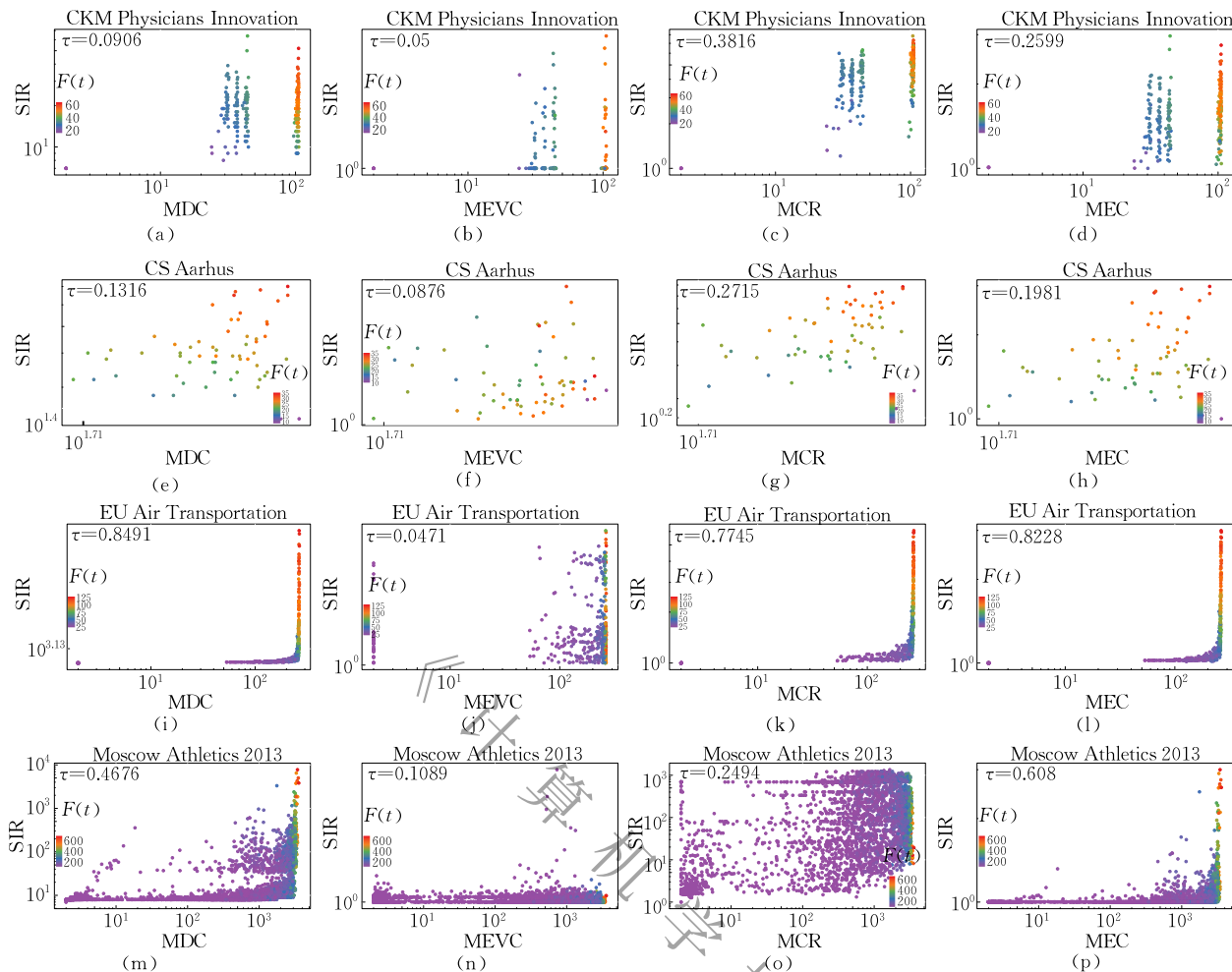


图 8 节点测量值与 SIR 传播能力间的相关关系

实验 5.

本实验从网络的鲁棒性和脆弱性方面对所提方法进行了评价。着重考察网络中一部分节点被移除后,网络结构发生变化情况,变化越大,说明移除的节点越重要。用各重要节点挖掘方法将网络中所有节点按重要性进行降序排序,将一部分节点从网络中移除,用 $C(i/N)$ 表示移除 i/N 比例的节点后,网络中最大连通集 (largest component) 的节点数目的比例,通过绘制 “ $i/N-C(i/N)$ ” 曲线,对节点移除的影响进行分析。图 9 显示了在 4 个数据集上使用 4 种重要节点挖掘方法对节点排序后,移除一定比例重要节点后,对网络最大连通集的影响。实验表明,在 4 个数据集上,通过多重证据中心性和多重度中心性得到的节点排序,当移除部分节点后,网络结构变化最大,特别是在移除前 1/4 后,

多重证据中心性呈现出的变化趋势要比多重度中心性更为明显。而多重 ClusterRank 方法,在前 3 个规模较小的数据集中,也能呈现出明显的变化趋势,但在规模较大的 Moscow Athletics 2013 数据集中,变化趋势缓慢,而多重证据中心性和多重度中心性呈现出瞬间下降的趋势。图 10 显示了在 Moscow Athletics 2013 数据集中,移除前 1000 个节点,网络最大连通集的变化趋势。结果表明,在大规模网络中,通过多重证据中心性和多重度中心移除极少数重要节点,便能引起网络结构的分裂,多重证据中心性所得到的结果要优于多重度中心性,而多重 ClusterRank 和多重特征向量中心性所得到的结果变化趋势不明显,进一步说明了多重证据中心性的有效性和多重 ClusterRank 方法不适用于大规模网络的局限性。

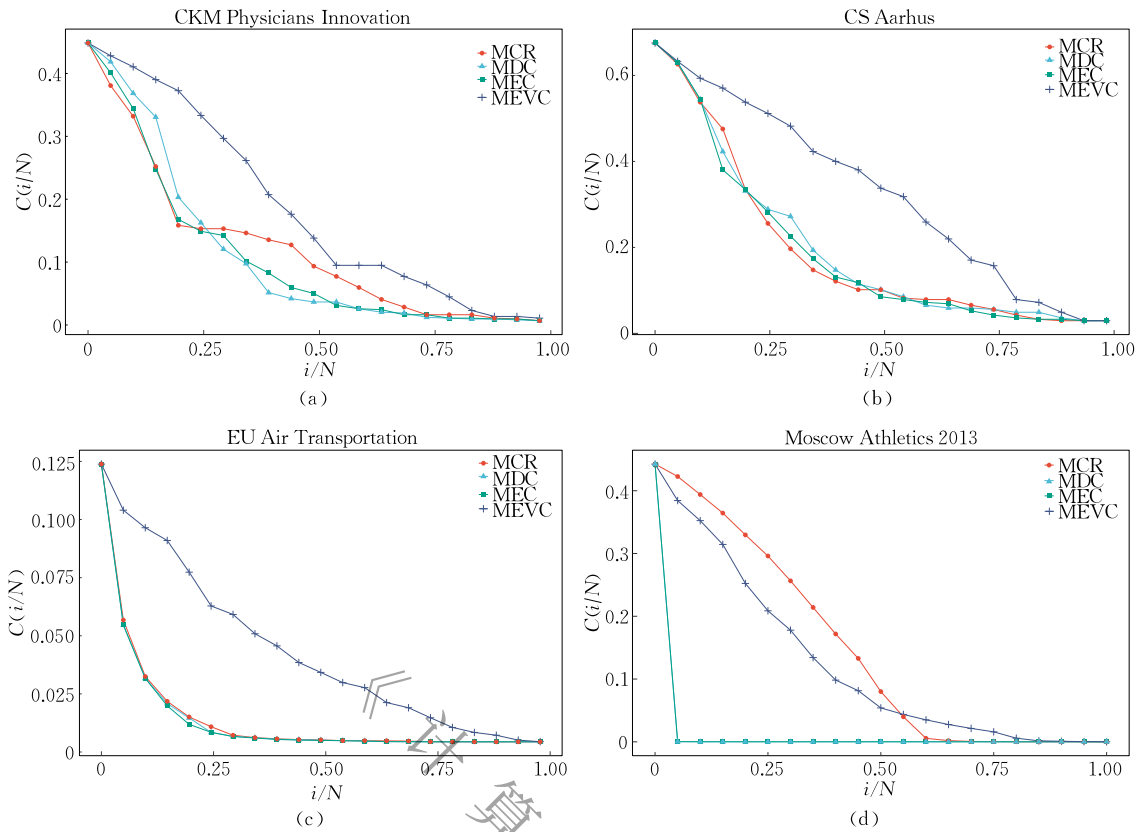


图 9 移除节点比例和网络最大连通集规模的关系

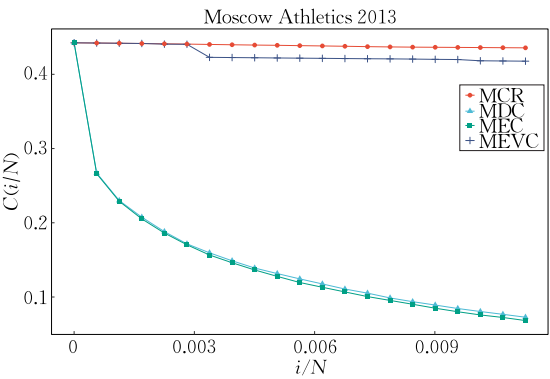


图 10 移除前 1000 个节点的网络连通变化趋势

通过上述 5 组实验可以得出,由于 D-S 证据理论是基于概率论的处理不确定问题的信息融合方法,在融合节点中心性和传递性信息的过程中,用概率来刻画节点重要程度的大小,能够有效消除由于多关系网络耦合信息和传递机制的差异性对度量节点重要信息的影响,所提出的多重证据中心性能够更加精准有效地挖掘网络中的重要节点,而多重 ClusterRank 采用的是相关性融合的确定方法,但由于多关系网络耦合信息和传递机制的差异性,节点的传递性和中心性之间不具备相关性,因此无法准确挖掘网络中的重要节点.特别是在大规模网络

中,网络中心性和传递性间的矛盾更加突出,导致多重 ClusterRank 度量重要节点有效性的降低.多重度中心性和多重特征向量中心性由于只考虑了单方面因素,虽然在算法复杂度上具有优势,但相比于综合考虑中心性和传递性的多重证据中心性,在度量效果上缺乏精准性.

6 结束语

本文针对无权无向多关系网络重要节点挖掘问题进行了研究分析,综合考虑多关系网络的中心性和传递性对节点重要程度的影响,提出了多重 ClusterRank 重要节点度量指标,并针对其存在问题引入 D-S 证据理论进行改进,进一步提出了多重证据中心性重要节点挖掘方法.在 4 个真实网络数据集上的实验结果表明:

(1) 在多关系网络中,由于受耦合信息和传递机制差异性的影响,节点的度中心性和局部聚集系数不具备相关性,对二者进行直接融合的多重 ClusterRank 会造成计算结果误差,降低重要节点挖掘的准确性,特别是在大规模网络中,无法有效识别网络中的重要节点;

(2) 通过引入 D-S 证据理论进行指标融合的多重证据中心性,可以有效消除多关系网络耦合信息和传递机制的差异性对节点重要信息造成的影响,相比于只关注网络结构的多层度中心性和多层特征向量中心性,能够更加准确地挖掘网络中的重要节点,特别是在大规模网络中优势更为明显;

(3) 相比于多重 ClusterRank,改进的多重证据中心性具有更低的时间复杂度。

综上,本文所做工作不仅针对无权无向的多关系网络的重要节点挖掘方法提供了新的思路方法,也进一步拓展了 D-S 证据理论的应用场景。但是,本文的工作也不够完美,受实验环境的限制,尚不适用于超大规模的多关系网络。如何对超大规模网络进行表示,设计合理方法挖掘网络中的重要节点,在保证准确性的基础上有效解决时空复杂度较高的问题,仍是多关系网络重要节点挖掘研究所面临的难点问题,也是今后研究工作的重点之

参 考 文 献

- [1] Cohen R, Erez K, Benavraham D, et al. Breakdown of the Internet under intentional attack. *Physical Review Letters*, 2001, 86(16): 3682-3685
- [2] Albert R, Jeong H, Barabási A. Error and attack tolerance of complex networks. *Nature*, 2001, 406: 378-382
- [3] Fang Jin-Qing. From a single network to “network of networks” development process: Some discussions on the exploration of multilayer supernetwork models and challenges. *Complex Systems and Complexity Science*, 2016, 13(1): 40-47 (in Chinese)
(方锦清. 从单一网络向《网络的网络》的转变进程——略论多层次超网络模型的探索与挑战. *复杂系统与复杂性科学*, 2016, 13(1): 40-47)
- [4] Wang Na-Na, Gao Hong, Liu Wei. Research on a heterogeneous edge multi-graph network model. *CAAI Transactions on Intelligent Systems*, 2017, 12(4): 475-481(in Chinese)
(王娜娜, 高红, 刘巍. 异质边多重图网络模型研究. *智能系统学报*, 2017, 12(4): 475-481)
- [5] Wang Na-Na, Gao Hong, Li Shan-Shan, et al. A research on hypergraph of heterogeneous edge. *Journal of Guangdong University of Technology*, 2017, 34(1): 6-10(in Chinese)
(王娜娜, 高红, 李珊珊等. 基于异质超边的超图. *广东工业大学学报*, 2017, 34(1): 6-10)
- [6] Huang Rui-Yang, Wu Qi, Zhu Yu-Hang. Community detection in dynamic heterogeneous network with joint nonnegative matrix factorization. *Application Research of Computers*, 2017, 34(10): 2989-2992(in Chinese)
(黄瑞阳, 吴奇, 朱宇航. 基于联合矩阵分解的动态异质网络社区发现方法. *计算机应用研究*. 2017, 34(10): 2989-2992)
- [7] Bin Sheng, Sun Geng-Xin. Important node detection algorithm for multiple relationships on line social network based on multi-subnet composited complex network model. *Journal of Nanjing University: Natural Sciences*, 2017, 53(2): 378-385 (in Chinese)
(宾晟, 孙更新. 基于多子网复合复杂网络模型的多关系社交网络重要节点发现算法. *南京大学学报: 自然科学*, 2017, 53(2): 378-385)
- [8] Kolda T G, Bader B W. Tensor decompositions and applications. *SIAM Review*, 2009, 51(3): 455-500
- [9] Jo H, Moon H, Ki Baek S. Immunization dynamics on a 2-layer network model. *Physica A: Statistical Mechanics and Its Applications*, 2006, 361(2): 534-542
- [10] Acar E, Yener B. Unsupervised multiway data analysis: A literature survey. *IEEE Transactions on Knowledge and Data Engineering*, 2009, 21(1): 6-20
- [11] Kurant M, Thiran P, Hagmann P. Error and attack tolerance of layered complex networks. *Physical Review E: Statistical Nonlinear & Soft Matter Physics*, 2006, 76(2): 26103
- [12] Kurant M, Thiran P. Layered complex networks. *Physical Review Letters*, 2006, 96(13): 138701
- [13] Wei X, Chen S, Wu X, et al. A unified framework of interplay between two spreading processes in multiplex networks. *EPL (Europhysics Letters)*, 2016, 114(2): 26006
- [14] Mei G, Wu X, Wang Y, et al. Compressive-sensing-based structure identification for multilayer networks. *IEEE Transactions on Cybernetics*, 2018, 48(2): 754-764
- [15] Zhou C, Zemanová L, Zamora-López G, et al. Structure-function relationship in complex brain networks expressed by hierarchical synchronization. *New Journal of Physics*, 2007, 9(6): 178
- [16] Zamora-López G, Zhou C, Kurths J. Structural characterization of networks using the cat cortex as an example//*Lectures in Supercomputational Neurosciences: Dynamics in Complex Brain Networks*. Berlin: Springer, 2008: 77-106
- [17] Boccaletti S, Bianconi G, Criado R, et al. The structure and dynamics of multilayer networks. *Physics Reports*, 2014, 544(1): 1-122
- [18] Lü L, Chen D, Ren X, et al. Vital nodes identification in complex networks. *Physics Reports*, 2016, 650: 1-63
- [19] Freeman L C. A set of measures of centrality based on betweenness. *Sociometry*, 1977, 40(1): 35-41
- [20] Freeman L C. Centrality in social networks conceptual clarification. *Social Networks*, 1978, 1(3): 215-239
- [21] Bonacich P. Factoring and weighting approaches to status scores and clique identification. *The Journal of Mathematical Sociology*, 1972, 2(1): 113-120
- [22] Chen D, Lü L, Shang M, et al. Identifying influential nodes in complex networks. *Physica A: Statistical Mechanics and Its Applications*, 2012, 391(4): 1777-1787

- [23] Gao C, Wei D, Hu Y, et al. A modified evidential methodology of identifying influential nodes in weighted networks. *Physica A: Statistical Mechanics and Its Applications*, 2013, 392(21): 5490-5500
- [24] Chen D, Gao H, Lü L, et al. Identifying influential nodes in large-scale directed networks: The role of clustering. *PLoS One*, 2013, 8(10): e77455
- [25] Zhang Xin. Multilayer networks science: Concepts, theories and data. *Complex Systems and Complexity Science*, 2015, 12(2): 103-107(in Chinese)
(张欣. 多层复杂网络理论研究进展: 概念、理论和数据. *复杂系统与复杂性科学*, 2015, 12(2): 103-107)
- [26] Battiston F, Nicosia V, Latora V. Structural measures for multiplex networks. *Physical Review E*, 2014, 89(3): 32804
- [27] Cozzo E, Kivela M, Domenico M D, et al. Structure of triadic relations in multiplex networks. *New Journal of Physics*, 2015, 17(7): 73029
- [28] Solé-Ribalta A, De Domenico M, Kouvaris N E, et al. Spectral properties of the Laplacian of multiplex networks. *Physical Review E*, 2013, 88(3): 32807
- [29] Newman M E. Properties of highly clustered networks. *Physical Review E*, 2003, 68(2): 26-121
- [30] Newman M E. *Networks: An Introduction*. Oxford, England: Oxford University Press, 2010
- [31] Han Chong-Zhao, Zhu Hong-Yan, Duan Zhan-Sheng. *Multi-Source Information Fusion*. 2nd Edition. Beijing: Tsinghua University Press, 2010(in Chinese)
(韩崇昭, 朱洪艳, 段战胜. *多源信息融合*. 第 2 版. 北京: 清华大学出版社, 2010)
- [32] Wei D, Deng X, Zhang X, et al. Identifying influential nodes in weighted networks based on evidence theory. *Physica A*, 2013, 392(10): 2564-2575
- [33] De Domenico M, Solé-Ribalta A, Omodei E, et al. Centrality in interconnected multilayer networks. *Physica D*, 2013, 6(6): 7868
- [34] Coleman J, Katz E, Menzel H. The diffusion of an innovation among physicians. *Social Networks*, 1957, 20(4): 253-270
- [35] Magnani M, Micenkova B, Rossi L. Combinatorial analysis of multiple networks. *arXiv e-prints*, 2013, arXiv:1303.4986: 1-17
- [36] Cardillo A, Zanin M, Romance M, et al. Emergence of network features from multiplexity. *Scientific Reports*, 2012, 3(2): 1344
- [37] Omodei E, De Domenico M D, Arenas A. Characterizing interactions in online social networks during exceptional events. *Frontiers in Physics*, 2015, 3(15): 59
- [38] Ren Xiao-Long, Lü Lin-Yuan. Review of ranking nodes in complex networks. *Chinese Science Bulletin*, 2014, 59(13): 1175-1197(in Chinese)
(任晓龙, 吕琳媛. 网络重要节点排序方法综述. *科学通报*, 2014, 59(13): 1175-1197)
- [39] Yagan O, Qian D, Zhang J, et al. Conjoining speeds up information diffusion in overlaying social-physical networks. *IEEE Journal on Selected Areas in Communications*, 2013, 31(6): 1038-1048
- [40] Salehi M, Sharma R, Marzolla M, et al. Spreading processes in multilayer networks. *IEEE Transactions on Network Science and Engineering*, 2015, 2(2): 65-83
- [41] Zheng C, Xia C, Guo Q, et al. Interplay between SIR-based disease spreading and awareness diffusion on multiplex networks. *Journal of Parallel and Distributed Computing*, 2018, 115: 20-28



LUO Hao, Ph. D. candidate. His current research interests include data mining and complex network analysis.

YAN Guang-Hui, Ph.D., professor. His current research interests include database theory and system, Internet of Things engineering and application, data mining, complex

network analysis, etc.

ZHANG Meng, M. S. Her current research interests include data mining and complex network analysis.

BAO Jun-Bo, M. S. His current research interests include data mining and complex network analysis.

LI Jun-Cheng, M. S. His current research interests include data mining and complex network analysis.

LIU Ting, M. S. Her current research interests include data mining and information security.

Background

Identifying important nodes is one of the principal tasks for complex network analysis, and it is essential to understand the structure and dynamic characteristics of the complex networks, which has been widely used social networks,

transportation networks, financial networks, etc. The important nodes in complex networks are the extraordinary nodes which play more significant role than other nodes on the structure and function of networks.

In recent years, the researches of identifying important nodes are mainly based on single-relational networks. There are already more complete and rich indicators and methods to measure important nodes in single networks. According to the local property, it can divide into three categories: global centrality, local centrality, and semi-local centrality. Classical methods of global centrality include closeness centrality (CC), eigenvector centrality (EC), and betweenness centrality (BC). Degree Centrality (DC) is a typical method of local centrality. In order to balance the complexity of global centrality and the local centrality, semi-local centrality (SLC) is proposed to measure the importance of nodes by considering the fourth-order neighbor information of nodes while evidential semi-local centrality (ESC) was based on the evidence theory. But both measures focus on the centrality and ignore the critical role of transitivity in essential nodes measurement. Chen et al. proposed the ClusterRank considering the nearest neighbors and the local clustering coefficient of a node comprehensively, which is superior to the former one in terms of operational efficiency and accuracy. However, A single network could not satisfy the demand of accurately description about large number of interconnected and interdependent complex systems. Therefore, the study of Network of Networks (NoN) has gradually become one of the hot topics in complex network analysis. Multi-relational networks, as a typical category of NoN, exist widely in the real world. In the study of node importance, it is of great theoretical and practical significance to explore and extend the existing methods in single networks to multi-relational networks.

In this work, we focused on the study of identifying important nodes for multi-relational networks. Firstly, we created the multilayer network model to describe a multi-relational network and used the multilayer adjacency matrix to represent it. To quantify the transitivity for multi-relational networks, we proposed a method for calculating the local clustering coefficient according to the notion of clustering coefficient for multilayer networks. Ulteriorly, we proposed multiplex ClusterRank combining the multiplex degree

centrality and extending ClusterRank to multi-relational networks. Considering the influence of coupling information and the difference of transmission mechanism for multi-relational network on vital information of nodes, multiplex ClusterRank may have calculation errors in large-scale multi-relational networks, which may affect the accuracy of measuring essential nodes. Therefore, in this work we improved multiplex ClusterRank and proposed another more efficient method to identify vital nodes for multi-relational networks named multiplex evidential centrality, which fused the node degree centrality and local clustering coefficient by the Dempster-Shafer evidence theory. Numerous experiments were carried out on four real networks. The proposed method is evaluated from two aspects: robustness and vulnerability, transmission dynamics characteristics. Experimental results show that multiplex evidential centrality can get better results than multiplex degree centrality and multiplex eigenvector centrality which only focus on the structure of multiplex networks and multiplex ClusterRank does have calculation errors in large-scale networks, making it difficult to identify important nodes. The experimental results verified multiplex evidential centrality can effectively eliminate the influence of the coupling information and the transmission mechanism in multi-relational networks, and has lower time complexity than multiplex ClusterRank. In this paper, we not only provide new ideas and methods for identifying critical nodes for multi-relational networks but also expands the application of information fusion technology.

In the further research of identifying important nodes, we will focus on how to establish a reasonable model to represent the network and design an effective method for large-scale multi-relational networks.

This paper is supported by the National Natural Science Foundation of China (Grant No. 62062049), the Humanity and Social Science Youth Foundation of Ministry of Education of China (Grant No. 20YJCZH212), and the Natural Science Foundation of Gansu Province, China (Grant No. 62062049).