

基于半局部结构的异常连边识别算法

石灏苒¹, 吉立新², 刘树新², 王庚润²

(1. 信息工程大学, 河南 郑州 450001;
2. 国家数字交换系统工程技术研究中心, 河南 郑州 450002)

摘要: 随着网络科学领域研究的进展, 所涉及的真实网络类型愈加广泛。复杂系统中存在的冗余错误关系, 或出于异常目的刻意发生的行为, 如网页错误点击、电信网刺探呼叫等, 都对基于网络结构的分析工作造成了重大影响。复杂网络异常连边识别作为图异常检测重要分支, 旨在识别网络结构中由于人为制造或数据收集错误所产生的异常连边。现有方法主要从结构相似性角度出发, 利用节点间连通结构评估连边异常程度, 易导致网络结构分解, 且检测精度受网络类型影响较大。针对这一问题, 提出了一种 CNSCL 算法, 在半局部结构尺度下计算节点重要性, 分析不同类型局部结构, 在不同结构中根据半局部中心性量化连边对网络整体连通性贡献, 结合节点结构相似性差异量化连边可信程度。由于计算过程中需去掉连边以衡量对网络整体连通性影响, 存在节点重要性需重复计算问题。因此在计算过程中, 所提算法还设计了一种动态更新方法以降低算法计算复杂度, 降低了算法计算复杂度, 使其可推广应用至大规模网络。在 7 种具有不同结构紧密程度的真实网络上与现有方法进行对比, 实验结果表明, 在 AUC 衡量标准下, 该方法较基准方法具有更高的检测精度, 且在网络稀疏或缺失条件下, 仍能保持较为稳定的识别精度。

关键词: 复杂网络; 图异常检测; 异常连边识别; 鲁棒性

中图分类号: TP393

文献标志码: A

doi: 10.11959/j.issn.2096-109x.2021040

Abnormal link detection algorithm based on semi-local structure

SHI Haoran¹, JI Lixin², LIU Shuxin², WANG Gengrun²

1. Information Engineering University, Zhengzhou 450001, China
2. National Digital Switching System Engineering Technology Research Center, Zhengzhou 450002, China

Abstract: With the research in network science, real networks involved are becoming more and more extensive. Redundant error relationships in complex systems, or behaviors that occur deliberately for unusual purposes, such as wrong clicks on webpages, telecommunication network spying calls, have a significant impact on the analysis work based on network structure. As an important branch of graph anomaly detection, anomalous edge recognition in

收稿日期: 2021-01-07; 修回日期: 2021-03-06

通信作者: 石灏苒, shihaoran@126.com

基金项目: 国家自然科学基金 (61803384)

Foundation Item: The National Natural Science Foundation of China (61803384)

引用格式: 石灏苒, 吉立新, 刘树新, 等. 基于半局部结构的异常连边识别算法[J]. 网络与信息安全学报, 2022, 8(1): 63-72.

Citation Format: SHI H R, JI L X, LIU S X, et al. Abnormal link detection algorithm based on semi-local structure[J]. Chinese Journal of Network and Information Security, 2022, 8(1): 63-72.

complex networks aims to identify abnormal edges in network structures caused by human fabrication or data collection errors. Existing methods mainly start from the perspective of structural similarity, and use the connected structure between nodes to evaluate the abnormal degree of edge connection, which easily leads to the decomposition of the network structure, and the detection accuracy is greatly affected by the network type. In response to this problem, a CNSCL algorithm was proposed, which calculated the node importance at the semi-local structure scale, analyzed different types of local structures, and quantified the contribution of edges to the overall network connectivity according to the semi-local centrality in different structures, and quantified the reliability of the edge connection by combining with the difference of node structure similarity. Since the connected edges need to be removed in the calculation process to measure the impact on the overall connectivity of the network, there was a problem that the importance of nodes needed to be repeatedly calculated. Therefore, in the calculation process, the proposed algorithm also designs a dynamic update method to reduce the computational complexity of the algorithm, so that it could be applied to large-scale networks. Compared with the existing methods on 7 real networks with different structural tightness, the experimental results show that the method has higher detection accuracy than the benchmark method under the AUC measure, and under the condition of network sparse or missing, It can still maintain a relatively stable recognition accuracy.

Keywords: complex network, graph-based anomaly detection, abnormal link detection, robustness

0 引言

随着网络科学领域研究的进展,所涉及的真实网络类型愈加广泛。与个体特征相比,关系特征具有更易获得、更可信且可视化更强的优点。对于社交网络^[1]、电力网络^[2]、通信网络^[3]、交通运输网络^[4]等而言,数据间通联关系比数据本身蕴含更多有价值的信息。图异常检测作为复杂网络的主要研究方向,旨在利用所观察到的网络结构特征识别网络中由于人为制造或数据收集错误所产生的异常连边,相比基于个体特征的异常检测方法,具有特征不易伪造、简单高效的特点而在近年来备受关注,在网络入侵检测、电信诈骗检测、社交网络恶意用户识别等多个领域均有广泛应用^[5]。

基于图结构的异常检测方法将个体通联关系抽象为复杂网络结构进行分析,根据从不同层面提取的拓扑结构特征可以将其分为基于连边、基于子图、基于社团和基于分解 4 种方法^[6]。基于连边的异常检测方法假设节点越不相似,则存在于节点间连边越可能是异常连边。基于子图特征的方法主要通过提取节点邻域特征进行概率分布统计以识别相关异常连边所构成的子图,其中以节点自我中心网络特征为出发点已有较多研究,Oddball 方法通过统计自我中心网络连边与节点等关系以达到识别异常子图目的^[7],并在此基

础上衍生出了各类改进方法^[8],其对于集体异常连边具有较高识别精度,但容易忽略网络结构中的单点异常连边。基于社团方法通过划分社团可以较精确地识别跨社团异常连边,但对于社团内存在的异常连边检测精度较低^[9]。基于分解的方法从网络整体规律性出发,通过优化网络整体规律性以去除不符合网络生成机制的异常连边或子结构^[10],如 Tong 等^[11]在非负矩阵分解方法基础上提出了一种 NrMF 方法,在识别不符合网络结构规律的异常结构同时具有较好的解释性。该方法虽具有较高的精度,但较依赖于网络结构规律性且计算复杂度较高,不适用于大型网络。

由于从连边层面提取特征的异常检测方法对网络中存在的异常连边相关性等特征要求较低且具有较高的算法鲁棒性。因此,本文从连边层面出发,从结构相似性与网络连通度两方面对连边可信度进行衡量,提出了基于半局部结构信息的异常连边识别算法。由于在去除连边以衡量其对网络结构连通性影响的过程中将导致半局部结构改变,使部分节点重要性变化。为减少时间复杂度,本文分析了连边去除后对不同半局部结构的影响,利用动态更新方法避免了节点重要性的重复计算;在存在不同比例异常连边的多个真实网络结构中进行了实验,并考虑了网络真实结构不完整的影响,证明了该算法对于网络结构稀疏及真实网络结构缺失等影响均具有较高鲁棒性。

1 相关工作

1.1 相关工作介绍

基于连边层面的异常检测方法主要利用节点间拓扑结构相似程度对连边可信度进行衡量。节点相似性越高，则该节点对间连边可信度越高，根据所利用信息可分为基于局部结构、基于半局部结构和基于全局结构 3 种方法^[12]。其中，基于局部结构的方法通过节点一阶邻居结构衡量节点间相似性，如共同邻居指标^[13]（CN, common neighbors）、资源分配指标^[14]（RA, resource allocation）等，其具有计算复杂度低的优点，但同时其计算精度相比其他方法较低。基于全局的方法则利用全局结构对节点相似性进行衡量，如平均通勤时间^[15]（ACT, average commute time），其利用节点对间存在的所有连通结构进行相似度衡量，具有较高的识别精度，但考虑了过多无关信息导致计算复杂度过高，不适用于大规模网络。而基于半局部结构对前两种方法进行了折中，如 Zhou 等^[16]考虑了影响节点相似性的主要结构进而提出的局部路径（LP, local path）指标，其在局部结构基础上考虑了三阶路径的影响，具有与全局方法相近精度的同时降低了计算复杂度。

Zhang 等^[17]在真实网络中对比了包含以上算法在内的 18 种利用不同拓扑结构的相似性异常连边识别算法检测精度与鲁棒性并进行了总结。Zeng 等^[18]将 CN 算法与边介数相结合，提出了一种 Hybrid 异常连边识别算法，在消除异常连边同时保证了网络的连通性。以下对部分效果较好的相关指标进行简要介绍。

1) CN 指标：通过节点 x 与 y 共同连接的节点数量来衡量节点对间相似度，其中 $\Gamma(x)$ 代表 x 的邻居集合：

$$S_{xy}^{CN} = |\Gamma(x) \cap \Gamma(y)| \quad (1)$$

2) AA 指标（adamic-adar index）^[19]：在 CN 基础上考虑了节点重要性差异，用度对数的倒数加权表示这种共同邻居差异化。

$$S_{xy}^{AA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\lg k_z} \quad (2)$$

3) RA 指标：在 AA 指标基础上利用资源分配原理进行改进，用共同邻居度倒数代替 RA 中

对数的倒数进行加权。

$$S_{xy}^{RA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z} \quad (3)$$

4) HPI 指标（大度节点有利指标，hub promoted index）^[20]：在 CN 指标基础上考虑节点度对节点对相似性的影响，大度节点与其他节点具有更高相似性。

$$S_{xy}^{HPI} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\min\{k_x, k_y\}} \quad (4)$$

5) LP 指标：在共同邻居的基础上考虑三阶路径的影响，提出了基于局部路径的相似性指标，其定义为

$$S_{xy}^{LP} = A^2 + \alpha A^3 \quad (5)$$

其中 α 为衰减因子， A 代表网络邻接矩阵， A^3 代表网络的三阶路径矩阵。

6) 偏好性连接（PA, preference attachment）指标^[21]：从双方节点度考虑连边相似性，节点对间连边产生概率正比于两节点度的乘积。

$$S_{xy}^{PA} = k_x k_y \quad (6)$$

7) ACT 指标：该指标基于随机游走方法提出，将节点对 x 与 y 间相似度定义为一个随机游走粒子从 x 到 y 所需的平均步数，表示为

$$S_{xy}^{ACT} = \frac{1}{l_{xx}^+ + l_{yy}^+ - 2l_{xy}^+} \quad (7)$$

其中， l_{xy}^+ 代表拉普拉斯矩阵的伪逆 L^+ 中对应的元素值。

8) Hybrid 指标^[18]：该方法将 CN 指标与边介数 EB 分别归一化后加权线性相加表示节点对间连边产生概率，认为节点对间连边概率正比于两节点间共同邻居数量以及该连边的边介数中心性。

$$S_{xy}^{EB} = \sum_{m>n} \frac{C_{mn}^{xy}}{C_{mn}^{EB}} \quad (8)$$

$$S_{xy}^{Hyb} = \lambda \frac{S_{xy}^{CN}}{\max_{mn}(S_{mn}^{CN})} + (1-\lambda) \frac{S_{xy}^{EB}}{\max_{mn}(S_{mn}^{EB})}$$

其中， C_{mn}^{xy} 表示节点对 m, n 间所有最短路径中经过连边 l_{xy} 的数量， C_{xy}^{EB} 为对应连边的边介数中心性。

1.2 问题描述

对于数据错误收集导致的虚假噪声或是人为制造的异常行为，如跨网络异常访问等，这类行

为在网络结构中常以单点异常连边的形式存在,其主要特征是无规则分布且产生连边的双方节点结构相似性较低。在该类异常连边检测问题中,可定义一个无向无权网络为 $G(V,E)$, 其中, V 为网络上节点集合, E 为网络的连边集合, 节点总数为 $|V|=N$, 连边总数为 $M=N(N-1)/2$, 即全集 U 。假设观察到的网络连边集为 E^O , 异常连边组成的连边集合为 E^P , 则有 $E^O=E+E^P$, 且 $E \cap E^P = \emptyset$ 。给定一个相似度定义方法, 对网络上每对节点对 $(x,y) \in U$ 赋予一个分数值 S_{xy} , 然后进行降序排列, 排名越靠前的连边可信度越高, 而排名越靠后连边存在的概率则越小, 越可能是异常连边。

为了对检测精确程度进行评估, 从整体角度出发采用 AUC 对算法精度进行衡量。通过在不存在连边集 $U-E$ 中选取一定数量连边构成异常连边集作为测试集 E^P , 将 E^P 加入真实连边集 E 得到训练集 $E^{T[22]}$ 。显然 E^T 即为观察到的网络结构 E^O , 而异常连边排序越靠后, 则方法识别精度越高。此时的 AUC 可以理解在测试集中的边分数值比在真实连边集 E 中随机选择一条连边分数值低的概率, 每次随机从真实连边集 E 中选择一条连边与 E^P 中随机一条连边分值进行比较, 若 E^P 中分值较低则加 1 分, 若相等则加 0.5 分^[23]。独立比较 n 次, n' 代表分值较低的次数, n'' 代表相等的次数, 则 AUC 可以定义为

$$AUC = \frac{n' + 0.5n''}{n} \quad (9)$$

2 基于半局部结构的异常连边识别算法

网络中一条连边的异常程度可以从局部与全局两方面进行衡量; 从局部角度, 产生连边的节点双方越相似, 则越可能产生联系; 从全局角度, 一条连边若删除后对网络结构完整性影响越大, 其可信度越高。根据这一定义, 本文将半局部中心性与闭合三元组理论相结合, 提出了一种对观察网络连边可信度度量的异常连边识别算法 CNSCL, 该算法首先计算删除连边后其导致网络结构分解的可能性, 若删除该连边后网络分解可能性越高, 则其被认定为网络结构中的重要连边, 具有高可信度, 否则进一步从局部角度分析其与所连节点的相似程度。以此从两方面计算该节点

对间连边可信程度 S_{xy} , 当 S_{xy} 越小时, 该节点对间连边越可能是异常连边。算法如下。

算法 1 CNSCL 算法

输入 网络结构 $G(V,E)$, 测试集 E^P

输出 排序结果 L , 识别精度 AUC

1) $O=[]$; %用于存放节点的半局部中心性

2) for $i=1$ to $|V|$

3) $O(i)=SCL(i)$;

4) end

5) for $E_{xy} \in E$

6) if $\Gamma(x) \cap \Gamma(y) = \emptyset$

7) $E=E/E_{xy}$

8) $SCL(i)=Update(SCL(i))$

9) $S_{xy}=CNSCL(SCL(x), SCL(y))$

10) $E=E \cup E_{xy}$

11) else

12) $S_{xy}=CNSCL(\Gamma(x) \cap \Gamma(y))$

13) end

14) end

15) 对 S_{xy} 降序排列, 得到最终排序结果 L

16) 计算 AUC 识别精度

2.1 半局部中心性

节点在网络中位置不同, 使节点对于网络结构的影响有所差异, 为了衡量节点在网络中的重要性, 近年来提出了多种经典的节点重要性排序方法。其中, 度中心性指标通过节点度进行衡量, 计算最为简便, 但这种基于局部结构的排序方法无法精确表示节点在网络中所处位置, 故效果欠佳。而基于全局信息的方法, 如介数中心性与接近中心性, 虽然具有较好的节点重要性刻画能力, 但过高的计算复杂度难以在大规模网络上应用^[22]。Chen 等^[23]将基于局部和全局信息的节点重要性排序方法相结合, 提出了一种半局部中心性排序方法 (SCL, semi-local centrality), 可在大规模网络上有效衡量网络节点重要性。

该方法在衡量节点重要性过程中同时考虑了邻居节点以及多阶邻居节点的影响, 用 $N(w)$ 表示节点 w 的两步路径以内可达节点数, 则节点 v 的半局部中心性可以由式(10)定义:

$$Q(u) = \sum_{w \in \Gamma(u)} N(w) \quad (10)$$

$$SCL(v) = \sum_{u \in \Gamma(v)} Q(u) \quad (11)$$

其中, $\Gamma(u)$ 是节点 u 的邻居节点集合, $Q(u)$ 代表节点 u 邻居节点集的两步以内节点数之和, 则最终半局部中心性 SCL 可由邻居的 Q 值相加所得, 该方法涉及了节点的四阶邻居信息, 在降低计算复杂度的同时可有效衡量节点重要性。该方法与采用 Floyd 算法的介数中心性复杂度 $O(n^3)$ 相比较, 在具有好于介数中心性排序结果的同时, 计算复杂度仅为 $O(nk^2)$ 。

2.2 基于半局部结构的异常连边识别算法

如何定义节点间相似度, 使其可以基于现有观察到的网络结构识别出异常连边, 是异常连边识别的核心问题。图 1 将网络局部结构抽象为 3 种情形, 图 1(a) 表示网络社团内部存在共同邻居的紧密局部结构; 图 1(b) 表示节点对双方不存在共同邻居的局部结构, 该结构常见于社团间或社团内连接较不紧密的节点对; 图 1(c) 是当图 1(b) 中某一节点仅存在一条连边的特殊情况, 表示边缘节点的局部结构。显然在单个完整网络结构中, 由于网络完全连通, 对于情形 3 而言, 边缘节点 i 所拥有的连边一定是真实连边, 否则就是独立于网络的孤立节点。同样, 在情形 2 中若节点 i 与 j 之间不存在除连边 l_{ij} 外其他路径, 则该连边也一定是真实连边。

基于对以上 3 种局部结构的分析可知, 若仅基于闭合三元组或局部路径信息 (如 CN、LP 等) 的指标进行异常连边识别, 虽可对情形 1 中的节点对相似程度进行精准定义, 但对于情形 2 与情形 3 而言, 该类连边由于不存在共同邻居, 且不一定存在除连边 l_{ij} 外其他路径, 无法对其相似度进行准确刻画。而这类连边常常是稀疏网络结构中占比较大的连边, 对这类连边的错误识别可能导致网络的大面积分解。而同时考虑边介数与共同邻居影响的 Hybrid 异常连边识别方法^[18]虽可在网络结构稀疏影响下保证处理后网络结构的完整性, 且对情形 2 这类边介数较大的社团间连边进行精准识别, 但无法正确处理情形 3 中的边介数较小且不存在共同邻居的边缘节点连边。为有效处理情形 2、情形 3 中混合的异常连边, 利用半局部中心性衡量连边对网络连通性贡献, 从网

络连通性与节点相似性两方面定义相关连边可信度衡量方法。

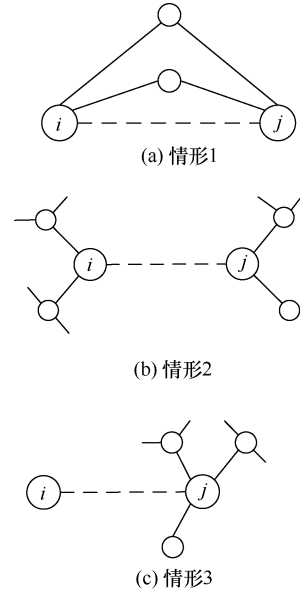


图 1 局部拓扑结构抽象
Figure 1 Abstraction of local topology

定义 1 基于半局部结构的异常连边识别。

由于半局部中心性通过节点四阶邻居信息来衡量节点重要性, 对于给定存在连边 l_{xy} 的节点对 (V_x, V_y) , 若其符合情形 2 中的局部结构特征, 则当其对应半局部中心性 $SCL(x)$ 、 $SCL(y)$ 越大, 双方拥有的路径越多, 节点对间存在连通路经的概率越大, 相应去除连边 l_{xy} 后网络分解可能性越低。而存在连通路经概率以及数量主要由连边数较少的节点决定, 因此连边可信程度可定义如下。

$$S_{xy} = \frac{1}{\min(SCL(x), SCL(y)) + 1} \quad (12)$$

式(12)中节点 V_x 与 V_y 对应的半局部中心性 $SCL(x)$ 、 $SCL(y)$ 为去掉连边 l_{xy} 后计算所得, 目的是避免异常连边对节点重要性计算产生影响并以此衡量连边去除后其对连通性影响。情形 3 作为情形 2 的一种特殊情况, 连边 l_{xy} 作为边缘节点与网络连接的唯一连边, 去掉连边后边缘节点 x 将成为孤立节点, 此时 $SCL(x)=0$, 则 $\min(SCL(x), SCL(y))=0$, 其连边可信度 $S_{xy}=1$, 证明其必然是真实存在连边。

对于情形 1 中异常连边而言^[24], 去除连边 l_{xy} 后对网络连通性没有影响。实验证明, 在连边较

稠密且共同邻居较多的网络上, CN 指标与 AA、RA 都具有较高的识别精度和鲁棒性, 且效果基本相同, 说明闭合三元组理论可以较好地处理情形 1 结构中的异常连边。因此, 本文结合 CN 与以上对于情形 2、情形 3 的异常连边处理方法, 提出了 CNSCL 模型, 具体方法如下。

$$S_{xy} = \begin{cases} |\Gamma(x) \cap \Gamma(y)|, & |\Gamma(x) \cap \Gamma(y)| \neq \emptyset \\ \frac{1}{\min(\text{SCL}(x), \text{SCL}(y)) + 1}, & |\Gamma(x) \cap \Gamma(y)| = \emptyset \end{cases} \quad (13)$$

该模型将连边进行分类处理, 对于网络结构稀疏的鲁棒性较强, 针对边缘节点连边以及社团间连边的有效识别可使其在保持较高准确度的同时保证连边识别处理后网络的结构完整性。

2.3 半局部中心性动态更新方法

由于以上方法在对每条观察到的连边 l_{xy} 计算相似度时都需要先去掉该连边再计算节点半局部中心性, 因此半局部中心性是不断变化的。若每次都对网络上所有节点的二阶邻居数 $N(w)$ 重新计算, 则该方法进行连边识别的时间复杂度为 $O(mnk^2)$, 若推广到所有节点间相似度计算, 则时间复杂度为 $O(n^3k^2)$, 显然是难以接受的。而若在每次去掉 l_{xy} 后仅计算节点 x 与 y 的四阶路径信息来构成其半局部中心性, 会使大部分节点的二阶邻居数被多次冗余地重复计算。本文从局部结构出发, 在观察到的网络结构上预先计算每个节点的半局部中心性, 之后不再重复计算 $N(w)$ 与 $Q(u)$, 而是根据节点 x, y 的二阶邻居信息对其半局部中心性直接进行动态更新, 使其时间复杂度降低到 $O((n+m)k^2)$, 其中 m 为不存在共同邻居的连边数, k 为网络上节点平均度。

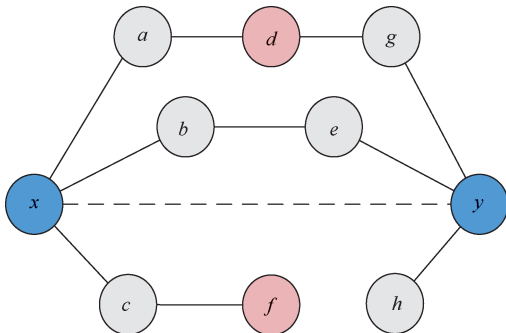


图 2 半局部中心性动态更新示意
Figure 2 Dynamic update of semi-local centrality

如图 2 所示, 当节点 x 与 y 不存在共同邻居时, 去除连边 l_{xy} 后, 对于节点 x, y 而言, 其自身及其一阶邻居节点的二阶邻居集 $N(w)$ 会发生改变, 但对二阶邻居没有影响。其可总结为以下两类节点: ① 去除连边 l_{xy} 的所连节点对 x, y ; ② x, y 的一阶邻居节点, 且该邻居节点不存在 x, y 之间的任何三阶路径上。

定义 2 半局部中心性动态更新。对于给定存在连边但不存在共同邻居的节点对 (x, y) , 当去掉连边 l_{xy} 后, 对于节点 y 而言, 其对应的 $N'(w)$ 更新规则如下, 其中, w_x 代表节点 x 的二阶邻居集合, w_y 代表节点 y 的二阶邻居集合。

$$\begin{aligned} N'(u_i) &= N(u_i) - 1, u_i \in \Gamma(y) \quad \Gamma(u_i) \cap \Gamma(x) = \emptyset \\ N'(y) &= N(y) - (|\Gamma(x) + 1| - |\Gamma(x) \cap w_y|) \end{aligned} \quad (14)$$

根据以上 $N(w)$ 更新公式, 对应的 $Q'(u)$ 为

$$\begin{aligned} Q'(u_i) &= \sum_{w \in \Gamma(u_i)} N'(w), \\ &= Q(u_i) - (|\Gamma(x) + 1| - |\Gamma(x) \cap w_y|) - \\ &\quad (|\Gamma(u_i) \cap \Gamma(y)| - |\Gamma(u_i) \cap \Gamma(y) \cap w_x|) \end{aligned} \quad (15)$$

其中, $u_i \in \Gamma(y) \quad i=1, 2, \dots, k_y$ 。

对于删除连边 l_{xy} 后节点 y 的半局部中心性 $\text{SCL}'(y)$ 可以通过以下计算直接进行更新。

$$\begin{aligned} \text{SCL}'(y) &= \sum_{u_i \in \Gamma(y)} Q'(u_i) = \\ &= \sum_{u_i \in \Gamma(y)} Q(u_i) - Q(x) - k_y (|\Gamma(x) + 1| - |\Gamma(x) \cap w_y|) - \\ &\quad \sum_{u_i \in \Gamma(y)} |\Gamma(u_i) \cap \Gamma(y)| - |\Gamma(u_i) \cap \Gamma(y) \cap w_x| \end{aligned} \quad (16)$$

节点 x 与以上 y 更新方法一致, 由以上分析可知, 对于连边 l_{xy} , 仅需节点 x 与 y 的二阶邻居信息就可以直接对半局部中心性进行更新并完成相似度计算。

3 仿真结果及分析

为验证算法有效性, 本文在 7 个典型的真实网络上分别进行实验, 并从以下几个方面对实验进行详细介绍。

3.1 实验数据

为了验证所提算法在不同结构网络上的识别精度鲁棒性, 实验选取了 7 个真实网络, 由 5 个结构稀疏网络和 2 个结构相对稠密的网络组成。5

个结构稀疏网络如下。① **Water**: 意大利北部城市摩德纳供水网络, 节点代表水站或用水用户, 连边代表输水管道。② **Euroroad(ER)**: 欧洲国际公路网, 节点代表城市, 连边代表公路。③ **Power Chilean(PC)**: 智利核心电网, 节点代表变电站或换流站, 连边为高压线路。④ **Power**: 美国电力网络拓扑结构。⑤ **Router**: 路由器层面的互联网拓扑结构网络。两个结构相对稠密的网络如下。

① **Politicalblogs(PB)**: 美国某论坛的博客首页间通过超链接所构成网络。② **Jazz**: 爵士乐手网络。表 1 展示了这些网络的数据特征。

3.2 实验结果与分析

为了体现所提 CNSCL 方法的性能优劣, 本文采用基于局部结构信息的相似度指标 CN、AA、RA、HDI、LP 和基于随机游走的相似度指标 ACT, 以及文献[18]提出的将介数中心性与 CN 结合的 Hybrid 方法等作为基本指标进行对比。实验为了模拟实际问题中的情况, 设置了以下两种环境以验证本文提出的方法不仅可以对仅存在异常连边的网络有较高检测精度, 且在结构缺失干扰条件下具有较强的鲁棒性: 只存在异常连边的网络结构; 存在异常连边且部分真实连边缺失的网络结构。

由于异常连边由随机选取生成, 为了避免影响实验结果准确性, 本文实验在 7 个真实网络上分别独立进行了 100 次实验, 且每次实验都重新随机生成异常连边, 最终取平均值作为实验结果; 并在 Power 与 PB 网络上分别验证了在上述两种环境下, 随着异常连边与缺失连边比例增加, 各指标的识别精度鲁棒性。

本节首先在仅存在异常连边的网络环境下通过 AUC 衡量了 CNSCL 的识别精度, 并与各基准方法进行对比, 如表 2 所示。由实验结果可以看出, 各基准指标在社交网络这类平均最短距离较小、聚类系数较高的网络上具有较强的异常连边识别能力, 但在面对平均最短距离较大、聚类系数较低, 邻居节点较少的结构稀疏网络, 其检测能力显著降低且表现不稳定。Hybrid 方法虽然拥有较稳定的识别能力, 但精度普遍不高。而本文提出的 CNSCL 方法在不同结构紧密程度的网络上均表现出较为稳定的检测精度, 且可较好抑制真实网络结构缺失的影响。

由表 2、表 3 实验结果对比发现, 在网络真实结构缺失且存在异常连边和仅存在异常连边两种情况中, 各类算法对于网络结构稀疏影响的鲁

表 1 7 个真实网络数据特征
Table 1 Features of 7 real network data

网络	网络节点数 (N)	网络连边数 (E)	节点平均度 ($\langle k \rangle$)	平均最短路径长度 ($\langle d \rangle$)	聚类系数 (C)	同配系数 (r)
Water	272	317	2.33	14.04	0.005	-0.031
ER	1 174	1 417	2.41	14.41	0.017	0.127
PC	347	444	2.56	8.16	0.087	-0.077
Power	4941	6 594	2.67	18.99	0.080	0.004
Router	5 022	6 258	2.49	6.45	0.012	-0.138
PB	1 222	16 717	27.36	2.74	0.320	-0.221
Jazz	198	5 484	27.70	2.24	0.618	0.0202

表 2 仅存在异常连边条件下各指标在 7 个真实网络上 AUC 衡量下的识别精度
Table 2 Measurement on 7 real networks existing abnormal edges by AUC

网络	基准指标								
	CN	AA	RA	HPI	PA	LP	ACT	Hybrid	CNSCL
Water	0.4981	0.4982	0.498	0.498	0.1757	0.1389	0.2392	0.1815	0.7701
ER	0.528	0.5298	0.5286	0.5282	0.337	0.3839	0.4338	0.2559	0.6509
PC	0.6035	0.6046	0.6049	0.6031	0.4958	0.5922	0.497	0.4746	0.7403
Power	0.6029	0.6033	0.6033	0.6034	0.4689	0.568	0.4827	0.4901	0.7381
Router	0.5645	0.5649	0.5651	0.5647	0.7393	0.7879	0.3789	0.6575	0.8194
PB	0.9232	0.9252	0.9231	0.9494	0.9049	0.9313	0.8876	0.8797	0.9233
Jazz	0.9461	0.9511	0.958	0.8921	0.7671	0.9424	0.7868	0.9451	0.945

表 3 异常与缺失连边比例为 0.1 条件下各指标在 7 个真实网络上 AUC 衡量下的识别精度
Table 3 AUC of 7 real networks existing abnormal edges which missing 10% edges

网络	基准指标								
	CN	AA	RA	HPI	PA	LP	ACT	Hybrid	CNSCL
Water	0.4997	0.4996	0.4995	0.4995	0.156	0.1793	0.2438	0.2152	0.6949
ER	0.5209	0.5202	0.5204	0.5206	0.3008	0.3276	0.3435	0.2546	0.6261
PC	0.574	0.5744	0.5741	0.5735	0.4417	0.5088	0.3691	0.4067	0.7064
Power	0.5817	0.582	0.5814	0.5817	0.4336	0.5037	0.3928	0.5827	0.6912
Router	0.5434	0.5428	0.5427	0.5425	0.7229	0.7432	0.318	0.5021	0.7846
PB	0.9183	0.9203	0.9174	0.8858	0.9041	0.929	0.885	0.9195	0.919
Jazz	0.9401	0.9448	0.9518	0.9436	0.762	0.9355	0.7799	0.9381	0.9415

棒性普遍较低, 且网络真实结构缺失对其识别精度具有较大影响。为验证以上各类方法在存在不同比例异常连边情况下对于网络结构稀疏性影响的鲁棒性, 实验进一步在不同异常连边比例下对以上指标在 **Power** 和 **PB** 网络上的识别精度进行了计算, 以对比随着异常连边比例增加, 各类方法在稀疏网络以及连边相对稠密网络上识别精度以及算法处理后网络连通度的变化。对于真实网络 A^T , 在保证网络完全连通基础上, 先在其连边集 E^T 中按比例 **fm** 去掉一部分连边 E^M , 以模拟数据缺失导致的真实网络结构缺失; 后在此基础上

按比例 **fs** 添加一部分连边 E^S , 模拟数据收集错误或人为刻意制造的异常连边, 最终得到观察网络 A^O , 其对应连边集为 E^O 。显然有 $E^O = E^T - E^M + E^S$, 由表 1 各真实网络数据特征可知, 结构稀疏网络连边仅略多于节点数, 因此缺失连边噪声比例 **fm** 应控制在一定范围内以保证 A^O 的初始连通性。为方便比较, 图 3 从以上方法中选取了一部分具有代表性的指标在 **fm=0** 和 **fm=0.1** 两种情况下, 验证了随异常连边比例 **fs** 增加, 在 **Power** 和 **PB** 网络上的识别精度变化, 图 4 验证了以上实验中异常连边识别算法处理后网络连通性变化。

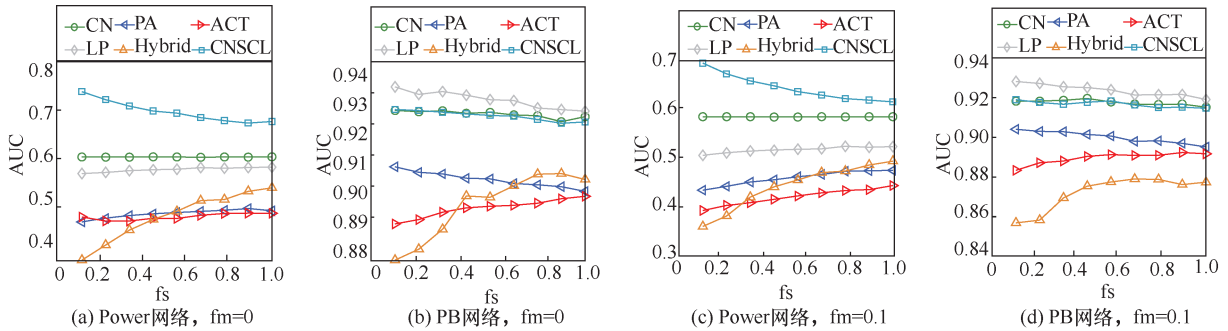


图 3 各指标在 **Power** 与 **PB** 网络上的噪声检测精度
Figure 3 Noise detection accuracy of indexes on **Power** and **PB**

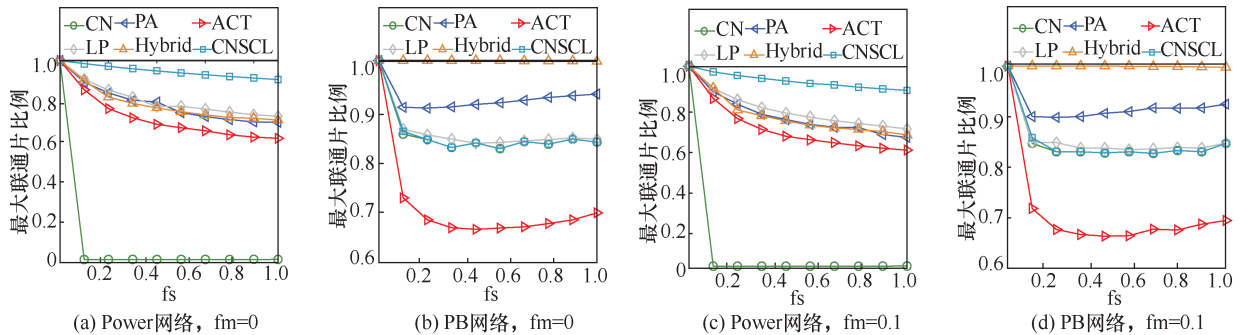


图 4 各指标在 **Power** 与 **PB** 网络上的噪声检测精度
Figure 4 Giant component of indexes on **Power** and **PB**

由图 3、图 4 实验结果可以发现, 各基准方法在 PB 社交网络上普遍具有较高识别精度与网络连通度, 但在 Power 网络上识别精度显著降低, 且无法保证网络具有较高连通度。出现这类现象的原因可能是 Power 网络中连边稀疏, 网络平均度较低, 且边缘节点较多, 社团间连边较少。没有共同邻居且连边双方路径较少, 使这类连边呈现的性质与异常连边特征较为相似。因此仅通过 CN、LP、ACT 这类通过路径计算相似度的方法难以进行区分, 而 CNSCL 方法从网络连通性出发衡量该类连边可以进行相对准确的识别。

在异常连边识别过程中, 需对网络结构中观察到的 m 条连边进行计算。Hybrid 方法由于需要求边介数, 因此其时间复杂度为 $O(n^2m)$, 可见其虽然可以保证处理后网络的完整性, 但计算太过复杂, 无法推广至大规模复杂网络中。而本文提出的 CNSCL 方法时间复杂度为 $O((n+m_1)k^2 + (m-m_1)k)$, 其中 m_1 为无共同邻居的连边数量。推广至对所有节点间相似度计算问题时, 其时间复杂度仅为 $O((n+m_1)k^2 + (n^2-m_1)k)$, 由此可见 CNSCL 方法在保证识别精度稳定、异常连边处理后网络结构连通的同时, 其计算复杂度使其可以推广至大规模网络。

4 结束语

在图异常检测领域中, 对呈现异常性质的重要节点研究较为广泛, 但针对异常连边识别的相关研究工作较少。除网络中广泛存在的欺诈入侵等异常行为外, 近年来针对对抗链路预测所提出的链路隐匿方法的研究逐渐兴起^[25], 该类方法通过在网络上增减连边制造异常连边以达到隐匿目标链路的目的, 将使各类异常行为难以被检测。而基于复杂网络结构的异常连边识别方法可以有效抑制这类影响, 具有重要应用价值。本文从节点重要性角度出发, 通过对网络局部结构进行抽象, 将相似性方法与半局部中心性相结合, 分析了网络上社团间连边或边缘连边的特点, 通过连边对网络连通的重要性来衡量其真实性, 显著降低该类连边被错误识别的概率的同时提高总体检测精度。在 7 个具有不同结构紧密程度的真实网络上的对比实验结果表明, 在 AUC 衡量标准下, 该方法在稀疏网络上检测

效果显著且同样适用于正常网络, 在网络真实结构缺失时仍能保持较为稳定的识别精度。此外, 由于该方法时间复杂度较低, 可推广应用于存在异常连边的大规模复杂网络。

参考文献:

- [1] 程晓涛, 刘彩霞, 刘树新. 基于局域信息的社交网络信息传播模型[J]. 计算机应用, 2015, 35(2): 322-331.
CHENG X T, LIU C X, LIU S X. Information propagation model for social network based on local information[J]. Journal of Computer Applications, 2015, 35(2): 322-331.
- [2] 刘树新, 季新生, 刘彩霞, 等. 一种信息传播促进网络增长的网络演化模型[J]. 物理学报, 2014, 63(15): 429-439.
LIU S X, JI X S, LIU C X, et al. A complex network evolution model for network growth promoted by information transmission[J]. Journal of Physics, 2014, 63(15): 429-439.
- [3] 刘树新, 季新生, 刘彩霞, 等. 局部拓扑信息耦合促进网络演化[J]. 电子与信息学报, 2016, 38(9): 2180-2187.
LIU S X, JI X S, LIU C X, et al. Information coupling of local topology promoting the network evolution[J]. Journal of Electronics & Information Technology, 2016, 38(9): 2180-2187.
- [4] 刘树新, 李星, 陈鸿昶, 等. 基于资源传输匹配度的复杂网络链路预测方法[J]. 通信学报, 2020, 41(6): 70-79.
LIU S X, LI X, CHEN H C, et al. Link prediction method based on matching degree of resource transmission for complex network[J]. Journal of Communications, 2020, 41(6): 70-79.
- [5] POURHABIBI T, ONG K L, KAM B H, et al. Fraud detection: a systematic literature review of graph-based anomaly detection approaches[J]. Decision Support Systems, 2020, 133: 113303.
- [6] AKOGLU L, TONG H, KOUTRA D. Graph-based anomaly detection and description: a survey[J]. Data Mining and Knowledge Discovery, 2015, 29(3): 626-688.
- [7] AKOGLU L, MCGLOHON M, FALOUTSOS C. OddBall: spotting anomalies in weighted graphs[J]. Lecture Notes in Computer Science, 2010, 6119(2): 410-421.
- [8] LALEH N, CARMINATI B, FERRARI E. Graph based local risk estimation in large scale online social networks[C]//IEEE International Conference on Smart City/Socialcom/Sustaincom. 2016.
- [9] DINI F, KATENKA N, BARFORD P, et al. Intrusion as (anti) social communication: characterization and detection[C]// Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2012.
- [10] WU T, QIAO S J, XIAN X P, et al. Network reconstruction and controlling based on structural regularity analysis[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 29: 1-14.
- [11] TONG H, LIN C Y. Non-negative residual matrix factorization with application to graph anomaly detection[C]//Proceedings of the Ele-

- venth SIAM International Conference on Data Mining. 2011.
- [12] MARTÍNEZ V, BERZAL F, CUBERO J C. A survey of link prediction in complex networks[J]. ACM Computing Surveys. 2016, 49(4): 1-33.
- [13] LORRAIN F, WHITE H C. Structural equivalence of individuals in social networks[J]. The Journal of Mathematical Sociology, 1971, 1(1): 49-80.
- [14] ZHOU T, LÜ L Y, ZHANG Y C. Predicting missing links via local information[J]. The European Physical Journal B, 2009, 71(4): 623-630.
- [15] KLEIN D J, RANDI Ć M. Resistance distance[J]. Journal of Mathematical Chemistry, 1993, 12(1): 81-95.
- [16] LÜ L Y, JIN C H, ZHOU T. Similarity index based on local paths for link prediction of complex networks[J]. Physical Review E, 2009, 80(4): 046122.
- [17] ZHANG P, WANG X, WANG F, et al. Measuring the robustness of link prediction algorithms under noisy environment[J]. Scientific Reports, 2016, 6(1): 18881.
- [18] ZENG A, CIMINI G. Removing spurious interactions in complex networks[J]. Physical Review E, 2012, 85(3): 036101.
- [19] CHEN D B, LYU L Y, SHANG M S, et al. Identifying influential nodes in complex networks[J]. Physica A: Statistical Mechanics and its Applications, 2012, 391(4): 1777-1787.
- [20] RAVASZ E, SOMERA A L, MONGRU D A, et al. Hierarchical organization of modularity in metabolic networks[J]. Science, 2002, 297(5586): 1551-1555.
- [21] XIE Y B, ZHOU T, WANG B H. Scale-free networks without growth[J]. Physica A: Statistical Mechanics and its Applications, 2008, 387(7): 1683.
- [22] 任晓龙, 吕琳媛. 网络重要节点排序方法综述[J]. 科学通报, 2014, 59(13): 1175-1197.
- REN X L, LYU L Y. Summary of the sorting methods of important network nodes[J]. Chinese Science Bulletin, 2014, 59(13): 1175-1197.
- [23] CHEN D B, LYU L Y, SHANG M S, et al, 2012. Identifying influential nodes in complex networks[J]. Physica A: Statistical Mechanics and its Applications, 2012, 391(4): 1777-1787.
- [24] 王翔. 链路预测算法鲁棒性的研究[D]. 北京: 北京邮电大学, 2016.
- WANG X. Research on Robustness of Link Prediction Algorithm[D]. Beijing: Beijing University of Posts and Telecommunications, 2016.
- [25] ZHOU K, TOMASZ P M, MARCIN W, et al. Attacking similarity-based link prediction in social networks[C]//Proceedings of the 18th International Conference on Autonomous Agents and Multi-Agent Systems. 2019: 305-313.

[作者简介]



石灏苒（1997—），男，重庆人，信息工程大学硕士生，主要研究方向为网络科学、图异常检测。



吉立新（1969—），男，江苏淮安人，国家数字交换系统工程技术研究中心研究员、博士生导师，主要研究方向为电信网安全、数据挖掘。



刘树新（1987—），男，山东潍坊人，博士，国家数字交换系统工程技术研究中心助理研究员，主要研究方向为链路预测、通信网络安全。



王庚润（1987—），男，安徽蒙城人，博士，国家数字交换系统工程技术研究中心助理研究员，主要研究方向为电信网安全、数据处理。