基于邻层传播的相对重要节点挖掘方法

赵 娜[1,2,3]，李 杰[1,2]，王 剑[4]，彭西阳[1]，景 铭[1,2]，聂永杰[3]，郁 湧[1,2*]

(1. 云南大学软件学院 昆明 650091；2. 云南大学云南省软件工程重点实验室 昆明 650091；
3. 云南电网有限责任公司电力科学研究院 昆明 650217；4. 昆明理工大学信息工程及自动化学院 昆明 650504)

【摘要】目前针对复杂网络中相对重要节点的挖掘方法已有一些成果，但方法的效率和准确性仍有待提高。该文基于如下假设——如果一个节点具有某种特征的邻居节点越多，则该节点具有此特征的可能性越大——提出了一种基于邻层传播(NLD)的相对重要节点挖掘算法，并通过实验比较与分析，验证了该方法的准确性与适用性。

关 键 词 复杂网络； 邻层传播； 相对重要性； 相对重要节点
中图分类号 TP301 **文献标志码** A doi:10.12178/1001-0548.2020283

Relatively Important Nodes Mining Method Based on Neighbor Layer Diffuse

ZHAO Na[1,2,3], LI Jie[1,2], WANG Jian[4], PENG Xi-yang[1], JING Ming[1,2], NIE Yong-jie[3], and YU Yong[1,2*]

(1. School of Software, Yunnan University Kunming 650091; 2. Key Laboratory in Software Engineering of Yunnan Province Yunnan University Kunming 650091; 3. Electric Power Research Institute of Yunnan Power Grid Kunming 650217; 4. College of Information Engineering and Automation, Kunming University of Science and Technology Kunming 650504)

Abstract At present, there have been some achievements in mining methods for relatively important nodes in complex networks, but the efficiency and accuracy of the methods still need to be improved. Based on the assumption that if a node has more neighbor nodes with certain characteristics, the more likely this node has such characteristics. This paper proposes a relatively important node mining algorithm based on neighbor layer diffuse (NLD), and verifies the accuracy and applicability of the method through experimental comparison and analysis.

Key words complex network; neighbor layer diffuse; relative importance; relatively important nodes

分析复杂网络中的节点重要性，是一个被广泛关注且具有重要意义的研究方向。目前，节点重要性的研究方法主要是针对网络中的所有节点做全局排序，以判断节点重要性[1-3]。然而，“相对于一个或一组特定的节点，网络中哪些节点是最重要的?”这类问题显示了节点的相对重要性、局部重要性同样具有较强的现实意义，尤其是当网络的规模非常大的时候。解决这类问题的一种典型办法就是先量化一个节点相对于一个已知重要节点的重要性 (称为相对重要性，有时也称为接近性或者相似性)，再计算一个节点相对已知的重要节点集的重要性，从而找到相对重要节点，即相对重要节点挖掘[4]。如在罪犯关系网络中，通过已知罪犯查找其余罪犯[5-7]；在蛋白质网络中，通过已知致病基因查找未知致病基因[8]，或通过已知染病节点查找或预测风险节点[9]；在传染病网络中，可针对已知感染人员，优先找出易感人群进行治疗、隔离，有效防止病毒的传播和扩散；在电力网络中，通过已知重要断路器或发电单元找出相对重要的断路器、发电单元等进行保护，可有效防止由相继故障引起的大范围停电。可见，挖掘网络中的相对重要节点具有重要的应用价值。

1 相对重要节点挖掘算法研究现状

相对重要节点挖掘的工作可以追溯到文献 [10] 在 2000 年提出的一种个性化的变种 HITS 算法研

收稿日期：2020－07－10；修回日期：2020－10－14
基金项目：国家重点研发计划 (2018YFB2100100)；国家自然科学基金 (62066048)；中国博士后科学基金 (2020M673312)；云南省科技厅面上项目 (202001BB050063)；云南省教育厅科学研究基金 (2019J0010，2019J0008)
作者简介：赵娜 (1982－)，女，博士，副教授，主要从事复杂性科学、软件工程及数据挖掘方面的研究.
通信作者：郁湧，E-mail：yuy1219@163.com

究工作。其后，文献 [11-12] 分别提出了个性化变种 PageRank 算法，开始更多地考虑网络中节点的相对重要性。2003 年，文献 [13] 定义了相对重要节点挖掘算法的通用框架，还明确提出了网络中节点的相对重要性是相对于一个或一组指定的特定节点集的重要性。随后相对重要节点的算法不断被提出。文献 [14] 提出了路径概率求和的方法，该方法将节点 s 相对于最近邻居节点s'的重要性定义为随机游走过程中从节点 s 跳到节点s'的概率。文献 [15] 提出集群粒子传播法来评价节点的相对重要性。文献 [16] 采用最短距离本身作为相对重要性衡量指标，文献 [17] 采用最短路径距离的 P 范数的倒数作为相对重要性衡量指标。文献 [18] 在介数中心性的基础上提出了信赖值的方法查找犯罪网络中的其余罪犯。尽管相对重要性挖掘方法已有一些成果，但仍存在诸多问题，如方法的效率和准确性有待提高；在不同网络上，方法的参数也需要明确如何选取等。

2 基于邻层传播的相对重要节点挖掘方法

在现实生活中，人类遗传病通常由多种致病基因 (含未知和已知的基因) 引起，由于它们导致相同或相似的疾病表型，因此未知的致病基因与已知的致病基因功能相关性越强，即某一基因与已知致病基因有越多的关联，该基因为致病基因的概率越大[19]。基于此思想，本文提出一个新的基于邻层传播 (neighbor layout diffuse, NLD) 的方法来挖掘相对重要节点，该算法的核心思想是：在网络中，与已知重要节点连接越多的节点，该节点与已知重要节点具有的共同特征越多，其为已知重要节点的相对重要节点的概率越大。

2.1 问题定义

对于由N个节点构成的网络$G(V,E)$，其中，N_1个节点构成重要节点集V_1，N_2个节点构成非重要节点集V_2。重要节点集$V_1 = R \cup U$，其中R是已知重要节点集，U是未知重要节点集。

本文的工作是：基于已知重要节点集R，分析目标节点集$T = V - R$中任意节点t相对于已知重要节点集R的重要性，最终找出目标节点集T中 top-k 个相对重要节点，并对结果进行 AUC、准确率和召回率分析。

2.2 NLD 算法中的分层

NLD 算法分为分层和传播两个步骤，传播是在分层的基础上进行的。分层是将已知重要节点作为第一层，放入集合L_1，将第一层节点的邻居节点作为第二层，放入集合L_2，将第二层节点的邻居节点 (不在第一层) 作为第三层，放入集合L_3，以此类推，直到将所有节点进行分层，得到分层结果 $\text{Layer} = \{L_1, L_2, \cdots, L_m\}$。

2.3 NLD 算法的传播

在上述分层过程完成后，算法进入传播步骤。NLD 算法的传播采用相对重要性分数值来量化节点重要性的传播情况。分层后，只有第一层节点为已知重要节点，且相邻两层的节点间有链接。设第一层的相对重要性分数值为 1，其他节点的值为 0。第一次传播时，由于第二层节点与第一层节点有链接，因此将第一层节点的相对重要性分数值通过链接传给第二层，即将第一层节点的重要性传给第二层，第二层任意节点与第一层节点联系越多，被传给的值就越多，即被传给的重要性越大。第二次传播时，把第二层节点作为传播源，第二层节点通过链接会传给相邻的第三层，同时会回传给相邻的第一层，以此类推。需要说明的是回传影响的仅是上一层，而下一层的节点值往往小于上一层的，所以回传的值会更小，回传不会影响整体的排序结果，并且能对上一层节点的相对重要性进行区分。NLD 算法描述如下：令传播轮数为S，初始时，$S = 0$，令已知重要节点集 R 中任意节点 r 的相对重要性分数值$Z_r(0) = 1$，目标节点集 T 中的任意节点t的相对重要性分数值$Z_t(0) = 0$。

第一轮传播时，$S = 1$，对于第一层中任意节点 i，其相对重要性分数$Z_i(0) \neq 0$，将$Z_i(0)/k$传递给邻居，k为节点度数；第二轮传播时，$S = 2$，对于第二层中任意节点i，其相对重要性分数$Z_i(1) \neq 0$，将$Z_i(1)/k$传递给邻居，以此类推，当 S=m−1 时，网络中任意节点i的相对重要分数$Z_i(S) \neq 0$，停止传播。对于任意节点i，相对重要性分数值$Z_i(S)$的公式表示如下：

$$Z_i(S) = Z_i(S-1) + \sum_{r \in L_s} A_{ri} \frac{Z_r(S-1)}{k_r} \tag{1}$$

式中，S表示传播轮数；$\boldsymbol{A}$表示邻接矩阵，A_{ri}根据节点r与节点i是否有连边取 1 或 0，k_r表示节点r的度数。L_s表示第S层节点的集合。NLD 算法使用伪代码表示如下：

输出：相对重要性分数值Z

初始化 $m = 1$，$L = L_1 = R$，$S = 0$；

for $i \in V$ do:

if $i \in R$ then:
$Z_i(S) \leftarrow 1$;
else $Z_i(S) \leftarrow 0$;
while $V \neq L$ do:
for 节点i是L_m中节点的邻居且$i \notin L$ do:
将节点i放入集合L_{m+1};
end for;
将集合L_{m+1}中的元素加入集合L;
$m \leftarrow m+1$;
end while;
while 存在任意节点v的$Z_v(S)=0$ do:
for $i \in V$ do:
$Z_i(S+1) \leftarrow Z_i(S)$;
for 节点$i \in L_s$ do:
for 节点i的邻居节点j do:
$Z_j(S+1) \leftarrow Z_j(S)+Z_i(S)/k_i$;
end for;
end for;
$S \leftarrow S+1$;
end while;
return Z;

传播结束后，比较所有节点的相对重要性分数值，相对重要性分数值越大的越可能是相对重要节点。

2.4　方法示例

以图 1 所示的网络为例，说明邻层传播算法的计算过程。

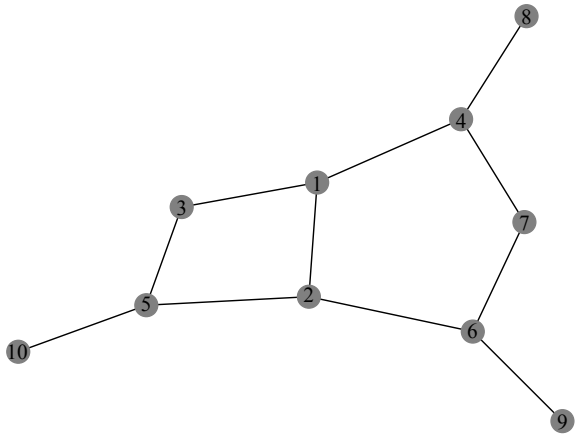

图 1　示例网络

把图中节点按顺序编号为 [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]。设节点 1 为已知重要节点。首先根据已知重要节点对网络进行分层。节点 1 属于第一层，节点 2, 3, 4 属于第二层，节点 5, 6, 7, 8 属于第三层，节点 9, 10 属于第四层。

初始时，根据节点编号顺序，节点对应的相对重要性分数值为 [1, 0, 0, 0, 0, 0, 0, 0, 0, 0]。经过第一轮传播后，各节点对应的相对重要性分数值为 [1, 0.333 3, 0.333 3, 0.333 3, 0, 0, 0, 0, 0, 0]。

经过第二轮传播后，各节点对应的相对重要性分数值为 [1.388 9, 0.333 3, 0.333 3, 0.333 3, 0.277 8, 0.111 1, 0.111 1, 0.111 1, 0, 0]。

经过第三轮传播后，各节点对应的相对重要性分数值为 [1.388 9, 0.463 0, 0.425 9, 0.5, 0.277 8, 0.166 7, 0.148 1, 0.111 1, 0.037 0, 0.092 6]。

此时，所有节点都有了相对重要性分数值，传播结束。然后，根据相对重要性分数值对节点进行排序。从示例网络的计算结果可得，对于节点 1，相对重要节点可能性的顺序应为 4, 2, 3, 5, 6, 7, 8, 10, 9。

3　实　验

3.1　数据集

实验采用了 4 个真实网络数据集，忽略网络连边的权重与方向，分别是：

1) 国际航空网络[20]，节点为国家，边代表两个国家有航线，SARS 病毒爆发早期传播到的国家为重要节点集。

2) 人类蛋白质相互作用网络 PPI[21]，节点代表蛋白质，边代表蛋白质之间存在相互作用，重要节点集是心脏病基因翻译的蛋白质。

3) Human 人类蛋白质相互作用网络[22]，节点代表人类蛋白质，边代表蛋白质之间存在相互作用，人类蛋白激酶这类蛋白质为重要节点集。

4) Mouse 小鼠蛋白质相互作用网络[22]，节点代表小鼠蛋白质，边代表蛋白质之间存在相互作用，小鼠蛋白激酶这类蛋白质为重要节点集。

以上网络数据集的基本拓扑特征如表 1 所示。

表 1　网络基本拓扑特征

网络	基本拓扑特征				
	N	N'	M	K	C
SARS	224	18	2 247	20.06	0.65
PPI	9 642	284	40 513	8.40	0.12
Human	3 574	186	6 002	3.36	0.15
Mouse	1 187	67	1 557	2.62	0.09

其中N表示网络节点个数，N'表示网络中重要节点个数，M表示网络的边数，K表示网络平均度，C表示网络平均聚类系数。

3.2 评价指标

本文使用 AUC、准确率和召回率对相对重要节点挖掘结果进行定量的准确性评价。

AUC 用于从整体上衡量算法的准确度，表示为：

$$\mathrm{AUC}=\frac{0.5n_1+n_2}{n} \tag{2}$$

具体计算过程为：每次从未知重要节点集和非重要节点集中各选择一个节点，比较两个节点的相对重要性分数，如果相等，则记 0.5 分，如果未知相对重要节点集中选择的节点相对重要性分数大于非重要节点集中选择的节点，则记 1 分。独立比较 n次 (n为遍历比较的次数)，其中有n_1次得 0.5 分，n_2次得 1 分。

准确率用于判断预测前L位的节点是否预测准确，定义为预测的前L个节点中预测准确的比例，公式如下：

$$\mathrm{precision}=\frac{N_r}{L} \tag{3}$$

式中，N_r为预测的前 L 个节点在未知重要节点集中出现的次数。

召回率用于判断预测的前L个节点中未知重要节点个数n_r与未知重要节点集U中节点个数的比值，具体可表示为：

$$\mathrm{recall}=\frac{n_r}{|U|} \tag{4}$$

3.3 实证比较

对每个网络进行 9 轮实验，每轮实验已知重要节点比率p依次为 10%、20%、30%、40%、50%、60%、70%、80%、90%。每轮实验中，用不同的方法计算节点的相对重要性，然后根据节点重要性进行排序。每轮进行 20 次独立实验，最后对排序结果计算 AUC 值。同时，取排序结果的前$N'/2$个节点作为预测的未知重要节点，比较准确率与召回率。

对比方法选用较为常用的 PPR[23]、KSmar[13] 和 PHITS[13]，3 种方法在 4 个网络上将各种方法的参数调至接近最优，具体取值如下：PPR、PHITS 方法分别取$s=0.75$、$s=0.1$，KSMar 方法中取$K=3$。比较结果如图 2、表 2 和表 3 所示。图 2 中，X 轴表示已知重要节点占重要节点集的百分比 P，Y 轴表示平均 AUC 值。

从实验结果可以看出，在 AUC 的评价下，无论已知重要节点比率多少，NLD 方法在 SARS、PPI 和 Mouse 网络上表现最好，在 Human 网络上表现较好。在准确率和召回率的评价下，NLD 在 Human 和 Mouse 网络上表现最好，在 SARS 和 PPI 网络上表现次好。

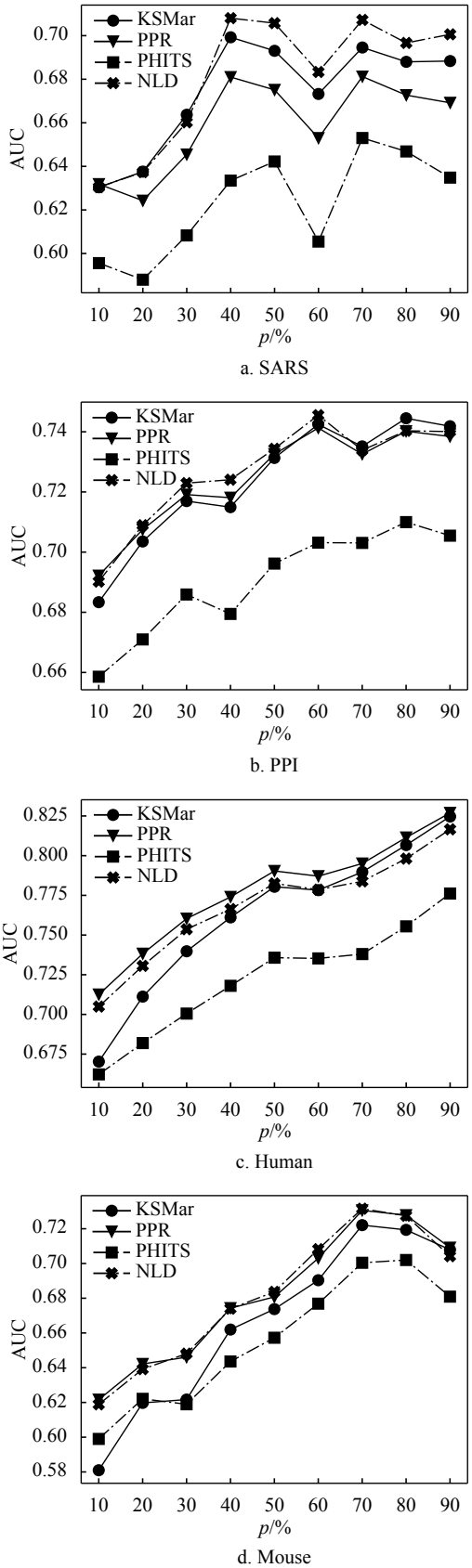

图 2 网络的 AUC 结果

对比不同的算法可知，KSmar 的结果会把距离所有已知重要节点更近的节点的相对重要性分数调高。通过 Ksmar 的公式 $KS = [TP + T^2P + \cdots + T^kP]$ 也可看出，转移概率矩阵会给k步内已知重要节点所能到达的节点赋予重要分数，k越小时，所获得的相对重要分数越高。PPR 是 PageRank 的改进方法，在无向图中，对于度大的节点往往会获得更多的相对重要性分数。PHITS 是 HITS 算法的改进，在无向图中，已知重要节点是权威节点也是中心节点，因此它往往给已知重要节点的邻居更高的重要性分数。NLD 则会给每层共同邻居越多的节点更多的重要性分数。在 SARS 网络中，SARS 爆发的国家与某个国家有越多的航班，该国家成为下一个爆发的国家的可能性更大，这符合 NLD 的思想，因此 NLD 在 SARS 网络上表现的较好。对于 PPI 网络，某一基因与已知致病基因有越多的关联，该基因为致病基因的概率越大[19]，符合 NLD 的思想，因此在 PPI 网络上表现也较好。

表 2　不同的方法在 4 个网络的平均准确率

网络	方法			
	KSmar	PPR	PHITS	NLD
SARS	0.183	0.175	0.179	0.183
PPI	0.110	0.106	0.127	0.109
Human	0.338	0.342	0.279	0.353
Mouse	0.280	0.274	0.261	0.284

表 3　不同的方法在 4 个网络的平均召回率

网络	方法			
	KSmar	PPR	PHITS	NLD
SARS	0.175	0.159	0.215	0.178
PPI	0.050	0.051	0.059	0.051
Human	0.083	0.088	0.075	0.092
Mouse	0.068	0.065	0.067	0.072

从实验的数据集来看，总体而言，NLD 方法优于其他方法，这说明 NLD 在相对重要节点挖掘方面具备准确性与适用性。由于这些方法的优劣一定程度上取决于网络结构本身，根据网络结构确定研究方法这也是今后研究的一个问题。

4　结　束　语

本文提出了一种挖掘相对重要节点的方法——NLD，该方法基于与越多已知重要节点关联，其为相对重要节点的概率越大的假设。本文将 NLD 与已有的挖掘相对重要节点较好的方法 KSmar、PPR、PHITS 进行对比，实验结果证明 NLD 在一定程度优于这些方法。同时，NLD 方法也为网络信息挖掘提供了新思路。在今后的工作中仍然有很多问题值得深入研究：1) 现有的各种度量网络中节点相对重要性的指标，比如路径长度的倒数、介数等，它们之间是否具有一定的联系。2) 现实世界中虽然很多网络都可以抽象为复杂网络，但针对不同网络设计其适用的挖掘算法仍是亟待研究的。

本文研究工作得到昆明市卫健委项目 (2020-09-04-112) 的资助，在此表示感谢。

参　考　文　献

[1] 赫南, 李德毅, 淦文燕, 等. 复杂网络中重要性节点发掘综述[J]. 计算机科学, 2007(12): 1-5.
HE Nan, LI De-Yi, GAN Wen-Yan, et al. Mining vital nodes in complex networks[J]. Computer Science, 2007(12): 1-5.

[2] CHEN D, LÜ L, SHANG M S, et al. Identifying influential nodes in complex networks[J]. Physica A, 2012, 391(4): 1777-1787.

[3] LÜ L Y, CHEN D B, REN X L, et al. Vital nodes identification-in complex networks[J]. Phys Rep, 2016, 650: 1-63.

[4] 朱军芳, 陈端兵, 周涛, 等. 网络科学中相对重要节点挖掘方法综述[J]. 电子科技大学学报, 2019, 48(4): 595-603.
ZHU Jun-fang, CHEN Duan-bing, ZHOU Tao, et al. A survey on mining relatively important nodes in network science[J]. Journal of University of Electronic Science and Technology of China, 2019, 48(4): 595-603.

[5] ALZAABI M. CISRI: A crime investigation system using the relative importance of information spreaders in networks depicting criminals communications[J]. IEEE T Inf Foren Sec, 2015, 10(2): 2196-2211.

[6] MAGALINGAM P, DAVID S, RAO A. Ranking the importance level of intermediaries to a criminal using a reliance measure[EB/OL]. (2015-07-07). https://arxiv.org/abs/1506.06221v3.

[7] MAGALINGAM P. Complex network tools to enable identification of a criminal community[J]. Bull Aust Math Soc, 2016, 94: 350-352.

[8] 赵静, 林丽梅. 基于分子网络的疾病基因预测方法综述[J]. 电子科技大学学报, 2017, 46(5): 755-765.
ZHAO Jing, LIN Li-mei. A survey of disease gene prediction methods based on molecular networks[J]. Journal of University of Electronic Science and Technology of China, 2017, 46(5): 755-765.

[9] 周涛, 汪秉宏, 韩晓璞, 等. 社会网络分析及其在舆情和疫情防控中的应用[J]. 系统工程学报, 2010, 25(6): 742-754.
ZHOU Tao, WANG Bing-hong, HAN Xiao-pu, et al. Social network analysis and its application in the prevention and control of propagation for public opinion and the epidemic[J]. Journal of Systems Engineering, 2010, 25(6): 742-754.

[10] CHANG H, COHN D, MCCALLUM A. Learning to create customized authority lists[C]//Proceedings of the 17th International Conference on Machine Learning. [S.l.]: ACM, 2000: 127-134.
[11] HAVELIWALA, TAHER H. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search[C]//IEEE Transactions on Knowledge and Data Engineering. [S.l.]: IEEE, 2003, 15(4): 784-796.
[12] JEH G, WIDOM J. Scaling personalized web search[C]//Proceedings of the 12th International Conference on World Wide Web. [S.l.]: ACM, 2003: 271-279.
[13] WHITE S, SMYTH P. Algorithms for estimating relative importance in networks[C]//The 3th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington DC, USA: ACM, 2003: 266-275.
[14] WANG H, CHANG C K, YANG H I, et al. Estimating the relative importance of nodes in social networks[J]. J Inf Process, 2013, 21(3): 414-422.
[15] RODRIGUEZ M A, BOLLEN J. An algorithm to determine peer-reviewers[C]//The 17th ACM Conference on Information and Knowledge Management. Napa Valley: ACM, 2008: 319-328.
[16] MAGALINGAM P, DAVIS S, RAO A. Using shortest path to discover criminal community[J]. Digital Investigate, 2015, 15: 1-17.
[17] LANGOHAR L. Methods for finding interesting nodes in weighted graphs[D]. Finland: University of Helsinki, 2014.
[18] MAGALINGAM P, DAVID S, RAO A. Ranking the importance level of intermediaries to a criminal using a reliance measure[EB/OL]. [2015-07-07]. https://arxiv.org/abs/1506.06221v3.
[19] TIV M, SNEL B, HUYNEN M A, et al. Predicting disease genes using protein-protein interactions[J]. Journal of Medical Genetics, 2006, 43(8): 691-698.
[20] JANI P. Airport, airline and route data[DB/OL]. (2017-01-02). https://openflights.org/data.html.
[21] PRASAD T S. Human protein reference database [DB/OL]. [2020-05-13]. http://www.hprd.org/sentDataRequest.
[22] XENARIOS I, RICE D W, SALWINSKI L, et al. DIP: The data-base of interacting proteins[J]. Nucleic Acids Research, 2000, 32(1): 289-291.
[23] HAVELIWALA T H. Topic-sensitive pagerank: A context-sensitive ranking algorithm for Web search[J]. IEEE Trans Knowl Data Eng, 2003, 15(4): 784-796.

编 辑 蒋 晓