一种复杂网络中节点安全重要性排序的度量方法

张子超[1,2], 郝蔚琳[1,2], 张伊凡[1,2]

[1] 北京大学信息科学技术学院 北京 100871
[2] 高可信软件技术教育部重点实验室 北京 100871

摘要 近年来, 网络空间安全成为信息安全中的热门领域之一, 随着复杂网络的研究日渐深入, 网络空间安全与复杂网络的结合也变得日益密切。网络的整体安全性依赖于网络中具体节点的安全性, 因此, 对网络节点的安全重要程度进行有效排序变得极为关键, 良好的排序方法应当将越重要的节点排在越靠前的位置。本文从网络的拓扑结构入手, 研究了网络节点的局部关键性, 在传统基础上考虑了相邻节点及次相邻节点的拓扑结构影响。同时, 由于传统方法很少引入动态因素, 因此本文引入了网络节点实时流量向量, 算法既包含网络拓扑结构, 又使用了不同时刻的节点流量, 采用了静态与动态相结合的方式。实验结果表明, 在破坏排序结果前 top-n 个节点时, 与传统方法相比, 本文算法在排序结果上具有更好的效果。

关键词 网络空间安全; 复杂网络; 节点重要性
中图法分类号 TP301 DOI 号 10.19363/J.cnki.cn10-1380/tn.2019.01.07

A measure approach for ranking the security importance of node security importance in complex network

ZHANG Zichao[1,2], HAO Weilin[1,2], ZHANG Yifan[1,2]

[1] School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China
[2] Key Lab of High Confidence Software Technologies(Peking University), Ministry of Education, Beijing 100871, China

Abstract In recent years, cyber security has become one of the hot fields in information security. With the deepening of research on complex networks, the association of network space security and complex networks has become increasingly close. The overall security of the network depends on the security of specific nodes in the network. Therefore, it is crucial to effectively rank the security importance of network nodes. A good ranking method should rank the more important nodes higher. This paper starts with the topology structure of the network and studies the local key information of the nearest nodes and next nearest nodes. At the same time, because traditional methods lack dynamic factors, real-time flow vector of network nodes is introduced in this paper. The algorithm not only contains network topology, but also node flow at different time points, adopting the combination of static and dynamic factors. Experimental results show that, compared with traditional methods, the algorithm presented in this paper has a better effect on ranking results when top-n nodes are destroyed.

Key words Cyber Security; complex network; node importance

1 引言

1.1 介绍

伴随着信息技术的迅猛发展, 我们的生活被各种类型的网络包围着。现实生活中的许多系统都可以通过网络的形式加以描述, 如人际关系网络[1]、交通网络[2]、科研引文合作网络[3]、新陈代谢网络[4]等等。近几年, 理解和研究网络的结构、功能和不同节点的关系成为了研究热点[1], 众所周知, 许多机制(如信息传播等)都受到较重要的节点的高度影响。

复杂网络是对复杂系统的抽象, 如果将复杂系统内部的各个元素抽象为节点, 元素之间的关系视为连接, 那么就构成了一个具有复杂连接关系的网络。因发现网络功能的正常运行极大地依赖若干重要节点, 近年来复杂网络中关键节点的研究引起了国内外的广泛关注, 如人际关系中地位较高的节点所处的位置, 交通网络中关键交通枢纽的车辆流量, 通信网络中核心节点对信息传播能力的影响和控制

通讯作者: 张子超, 学士, zhangzch@pku.edu.cn。
本课题得到国家重点研发计划“网络系统安全度量方法与指标体系”项目(No.2016YFB0800700)资助。
收稿日期: 2018-09-30; 修改日期: 2018-12-02; 定稿日期: 2018-12-09

等等实例。

大量的研究表明, 从各种现实网络中抽象出来的大规模复杂网络有三个基本统计特征:

(1) 小世界现象[5](Small World)。即网络中节点对之间的平均距离较短。网络的平均路径长度越短, 小世界现象越明显。

比如, 相对于传统的复杂网络而言, 社交网络[6](Social Network)具有更短的平均路径长度和有效直径, 其有效直径远小于 Web, 平均路径长度只有 Web 的三分之一。图 1 所示为韩国最大社交网络 Cyworld(L 为节点间最短路径平均长度, P(L)为最短路径长度为 L 的比例), 其中大部分节点间最短路径长度为 4~5, 超过 90%的节点间的平均路径长度小于 6。

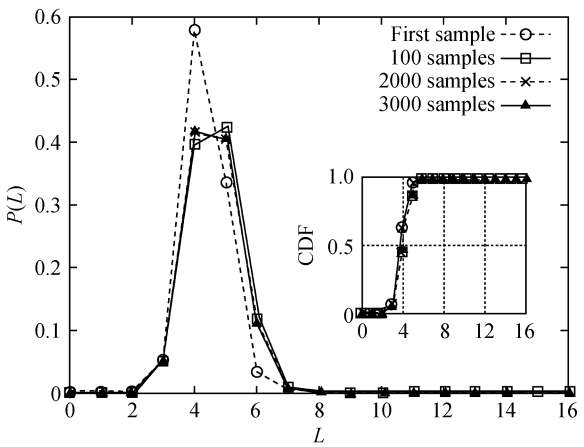


图 1 韩国 Cyworld 在线社交网络用户平均路径长度分布[7]

Figure 1 The average path length distribution of Cyworld online social network users in South Korea[1]

除了社交网络之外, 大量的实验研究表明, 真实网络几乎都具有小世界现象, 甚至存在“六度分隔”效应。

(2) 无标度特性[8](Scale-free)。无标度网络具有严重的异质性, 其各节点之间的连接状况(度数)具有严重的不均匀分布性, 即网络中少数节点拥有极其多的连接, 而大多数节点只有很少量的连接。无标度网络与随机网络不同, 见图 2 所示。

(3) 集聚性[8](Clustering)。在图论中, 集聚系数是用来描述一个图中的顶点之间结集成团的程度的系数。具体来说, 是一个点的邻接点之间相互连接的程度, 由集聚系数来刻画。

正是由于上述三个基本统计特征, 才使得研究节点重要性程度对于网络空间安全具有重大的意义。小世界现象导致了一旦关键节点被破坏或非法操纵, 如木马病毒传播, 那么造成的传播速度将会

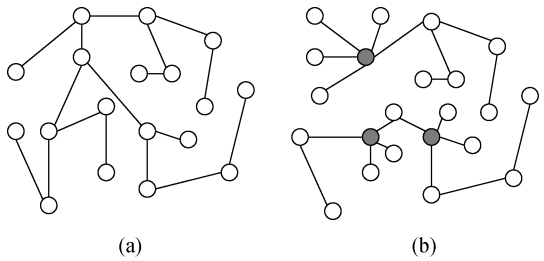


图 2 随机网络(a)与无标度网络(b)

Figure 2 Random network(a) and scale-free network(b)

非常快速的。无标度特性和集聚特性使得攻击后果成倍地放大, 在蓄意攻击下, 无标度网络显得异常脆弱[9], 如 DDoS 攻击[10]等。因此, 有必要依据网络的拓扑结构与实时流量, 结合静态与动态因素, 基于真实网络的测试环境, 通过合理的方法开展节点安全重要性排序工作。

本文首先介绍节点重要性排序问题的形式化描述, 然后介绍评价排序算法效果的方法。本文第二部分简要地回顾了复杂网络节点重要性排序方法的研究现状, 包括网络的局部特征、全局特征、随机游走及其他方法四个方面。随后的第三部分讲述本文提出的复杂网络中节点安全重要性排序方法, 结合网络局部拓扑结构特征和动态流量变化因素来说明我们提出的方法。第四部分为整体算法流程, 阐述本文的算法模型; 第五、六部分分别为实验与总结展望工作。

1.2 问题描述及评价方法

对于网络 $G=(V,E)$, 该网络是由 $|V|$ 个节点和 $|E|$ 条边所组成的一个无向网络。节点重要性排序是选择一种算法 A, 作用于图 G 后得到节点的有序排列 $A(G)$, 该结果中按照算法规定的节点重要性排序, 算法认定的越重要的节点处于 $A(G)$ 中越靠前的位置。

评价网络节点重要性排序算法的方法可以采用传播动力学进行度量, 即一般以网络节点作为传播源, 利用传播动力学模型进行仿真, 通过计算特定步骤后选定的网络目标节点的影响范围, 从而评价排序算法的优劣。而常用的度量方法是破坏性评价, 即删除算法选出的重要节点, 考察网络中该节点集 V_d 被破坏后网络 $G-V_d$ 连通状况的变化情况。一个网络中的节点被删除后, 与该节点相关联的所有边都会被删除, 从而会使网络的连通性变差, 一个节点越重要, 则被删除后网络的连通性会变得越差。通过这一评价方法, 越能破坏网络连通性的排序方法,

越具有准确的效果。

1.3 本文主要贡献

传统方法大多都是基于网络静态拓扑结构进行排序的, 而本文提出了一种结合网络静态拓扑结构与网络节点动态流量的排序度量方法。实验数据显示, 在与传统方法相比较时, 本文方法具有更好的排序效果。

2 相关研究

对于网络中一个节点的重要性没有特别精确的定义。一般情况下, 复杂网络中节点的重要性可以是该节点的影响力, 也可以是该节点承载的流量能力, 或者是一些其他的综合因素[11]。最初研究节点重要性是从网络的拓扑结构入手进行的, 国内外已经存在大量的相关研究[12], 并且存在一系列的度量指标。本小节从网络的局部特征、全局特征、随机游走及其他方法四个方面出发, 介绍网络节点排序方法的相关研究情况, 统计情况见表 1[13]。

2.1 网络局部特征指标

从网络局部特征入手的研究大多将关注点集中在节点相邻信息或者节点自身的信息之上, 这些做法的优点就是简单、直观, 并且由于局部信息量不大, 所以此类指标或算法的时间复杂度都不高, 可以适用于大规模数据网络。

最为直接和传统的做法就是使用节点的度指标, 节点的度定义为该节点的相邻节点数目, 即 $k(i)=\sum_{j\in V}a_{ij}$。在网络规模很大的情况下, 度甚至度分布已经能够对节点情况进行很好的刻画, 但是其缺点在于过于简单。

王建伟[14]等人认为连接节点的边越重要, 节点越重要是分析网络中节点的重要性与节点所连边的重要性的一个基本公理, 重新定义边的权值, 依据网络局部特征来评判网络中节点的重要性, 避免了对网络全局架构的了解。

Chen[12]等人为了使排序算法更加高效, 提出了一种权衡后的局部特征方法, 既保证了纳入充足局部信息, 又降低了时间复杂度, 其考虑了两层节点信息, 定义了局部集中度:

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

$$C_L(v)=\sum_{u\in\Gamma(v)}Q(u)$$

其中的 $\Gamma(u)$ 为节点 u 的相邻节点集合, $\Gamma(v)$ 为节点 v 的相邻节点集合, $N(w)$ 为节点 w 的最近相邻节点数和次相邻节点数之和。

Centola[15]研究社交网络的行为传播, 发现传播行为在高集聚(Clustering)类网络传播的更快, 节点的传播重要性与该节点的集聚性有关。

2.2 网络全局特征指标

基于网络全局信息的指标在计算节点重要性排序上会将网络全部信息包含在内, 由于包含整体信息, 因此此类指标的排序结果往往较为准确, 但很明显的缺点就是此类做法的算法时间复杂度较高, 对于数据规模较大的网络, 采用全局特征指标的做法时空开销巨大。

比较经典的两个指标是接近中心性(closeness centrality)和介数中心性[16](betweenness centrality)。接近中心性用来度量节点对其他节点施加影响作用的能力, 该指标认为一个节点的接近中心性数值越大, 则该节点越处于中心地位; 介数中心性度量节点在最短路径中的重要程度。关于上述两个指标的具体介绍详见本文后续的 4.1 节。

Stephenson 和 Zelen[18]认为度指标把相邻节点的重要程度一视同仁是过于简单的, 而节点之间的重要性是不同的, 例如一个节点的邻居节点如果很重要的话, 那么这个节点也应当被视为很重要; 反之同理。 所以需要将节点的相邻节点也作为一个反馈环节, 于是使用特征向量(eigenvector centrality)来度量节点重要程度, 此处的特征向量为网络邻接矩阵的最大特征值所对应的特征向量:

$$E_\lambda(i)=\lambda^{-1}\sum_{j=1}^{n}a_{ij}e_j$$

其中 λ 是邻接矩阵的最大特征值, $e=(e_1,e_2,\cdots,e_n)^{\mathrm{T}}$ 是邻接矩阵对应最大特征值所对应的特征向量。

与特征向量类似, Katz 指标[19]也将邻居节点的影响考虑在内, 同时引入了权重衰减因子。该做法是对较短路径赋予高权重, 对较长路径赋予低权重。但因该指标的细节部分依赖于大量数据, 因此该指标仍不是广泛使用的度量指标。同样引入新参数的还有 Comin[20]等人提出的考虑节点度和介数的综合指标:

$$C_{DB}(i)=\frac{C_B(i)}{(k(i))^{\gamma}}$$

其中 $C_B(i)$ 为节点 i 的介数值, γ 为最优参数, 使用中 γ 的选择需要着重考虑。

Zhang[21]等人提出了 Kernel 函数法, 将节点影响范围和最短路径考虑在内, 该方法较为准确, 但其计算复杂度仍然较高。

表 1 现有部分研究对比

Table 1 Comparison of existing studies

排序方法	局域性	优势	劣势	时间复杂度
度	局部	直接、简单	单纯考虑一个节点的局部特征	$O(N)$
接近中心性	全局	考虑部分节点的全局影响	不适合于随机网络	$O(N^3)$
介数中心性	全局	考虑了节点的信息传播的中心地位	不适合于大型网络	$O(N^3)$
Eigenvector	全局	考虑节点的相邻节点信息	仅将节点拓朴结构特征进行简单的线性计算	$O(N^2)$
Katz	全局	考虑不同相邻节点的影响	最优参数难以得到	$O(N^2)$
Kernel	全局	考虑全部节点的全局影响	不适合于大型网络	$O(N^3)$
PageRank	全局	考虑网络的全局拓扑结构特征	排序不唯一	$O(MI)$
LeaderRank	全局	排序较 PageRank 稳定, 排序结果唯一	不适合于无向图	$O(MI)$
HITS	全局	时间复杂度较低	结构不稳定	$O(NI)$

(注: N 指网络节点数量; M 指网络边的数量; I 指算法达到收敛的迭代次数)

2.3 随机游走排序方法

随机游走排序方法主要策略是根据网页之间的链接关系来进行网页排序, 此类算法当属 PageRank 最为常见。

PageRank 算法[22]发源于搜索引擎, 其基本思想类似于投票, 当网页 T 有一个链接指向网页 A 时, 网页 A 即根据网页 T 的重要程度获得一定分数, T 越重要, 则 A 获得的分值越高, 即一个页面的得分数由所有链向它的页面的重要性来决定。该分值的计算是一个迭代过程, 最终所有网页根据得分进行排序。

此外, 当网络中存在孤立节点或社团时, PageRank 排序结果并不唯一, 针对这一缺陷, LeaderRank[23]进行了弥补。LeaderRank 排序结果对网络节点重要性排序的效果通常要优于 PageRank 算法, 并且对网络中的噪声容忍性较强。Kleinberg[24]提出的 Hypertext-Induced Topic Search(HITS)算法将网页分成两类, 即较高价值的 Authorities 和将其串联起来的 Hubs, HITS 算法的目标就是寻找 Authorities 值排名较高的网页, HITS 算法是链接分析中非常基础且重要的算法。

2.4 其他节点重要性排序方法

除上述局部、全局和随机游走三大类方法之外, 还有些方法从网络的连通性、节点效率和综合方法的角度入手研究节点重要性度量。

许进[25]教授提出核与核度理论, 用于刻画网络中一组节点的重要性, 核度通过删除一些节点及其相连的边后, 网络中出现的连通分支个数来衡量, 定义了一个网络图 G 的核度为:

$$h(G)=\max\{\omega(G-S)-|S|;S\in C(G)\}$$

其中 $C(G)$ 为图 G 的点割集的集合, $\omega(G-S)$ 为图 G 删去 S 集合产生的连通分支数目。

周漩等人[9]通过定义节点效率和节点重要度评价矩阵, 提出了一种利用重要度评价矩阵来确定复杂网络关键节点的方法, 该方法考虑了节点效率、节点度值和相邻节点的重要度贡献, 用节点度值和效率值来表征其对相邻节点的重要度贡献。

陈静等[26]综合网络局部与全局信息, 根据节点的接近中心性和节点邻域关键度进行排序。文献[26]认为节点在复杂网络中的重要度首先取决于节点网络中的位置, 其次, 节点在网络中的重要度还取决于节点的连通能力, 即经过该节点的最短路径越多, 该节点在网络中的地位越重要, 对阵个网络连通性影响越大。

3 节点局部重要性度量

上一小节阐述的传统方法, 都是基于网络静态拓扑结构进行的。然而真实的网络是处于时刻变化当中的, 节点的重要程度不仅仅取决于网络固有的拓扑结构, 而且还取决于诸如实时流量等动态因素。因此, 本小节首先重定义了拓扑结构的计算方法, 随后加入流量矩阵 F。

选定网络 $G=(V,E)$, 其中 $V=\{v_1,v_2,\cdots,v_n\}$ 为节点集合, $E=\{e_1,e_2,\cdots,e_m\}$ 为边的集合, n 和 m 分别为网络所包含的节点数与边数。网络的邻接矩阵 A 定义为:

$$A=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$

其中 $a_{ij}\in\{0,1\}$, $i,j=1,2,\cdots,n$ 表示节点 i 与节点 j 之间是否有边关联, 如果节点 i, j 之间有边, 则

$a_{ij}=1$; 否则, $a_{ij}=0$。

在第二节中介绍的诸多算法中, 度中心性是最直接的度量方法, 其做法直接简单, 但是效果往往不如人意。考虑全局特征的方法中, 接近中心性和介数中心性在传统上可以很好地度量节点的影响力, 但是其计算复杂度 $O(n^3)$ 较高, 不适合用于大规模网络, Kernel 方法也是如此。

权衡计算复杂度与排序效果, 我们考虑节点的相邻节点与次相邻节点, 这一做法将节点的邻居与邻居的邻居包含在内。另一方面, 一个显然的道理是: 连接节点的边越重要, 则该节点通常也就越重要。因此, 我们重新定义边的权值。

定义网络中边的权重为:

$$w_{ij}=C_L(i)\times C_L(j) \tag{1}$$

其中 $C_L(i)$ 和 $C_L(j)$ 为局部集中度[12], 定义为:

$$C_L(i)=\sum_{u\in\Gamma(i)}\sum_{j\in\Gamma(u)}N(j) \tag{2}$$

$N(j)$ 为与节点 j 相邻的节点数目, $\Gamma(u)$ 为节点 u 的相邻节点集。

我们以图 3 中的 1 号节点举例, 1 号节点有 6 个相邻节点, 分别是节点 2 至 7, 其中节点 6 和 7 还具有节点 1 所不具备的相邻节点 8 和 9, 显然, 按照定义, 节点 1 的 $N(1)=6$。为便于描述, 令中间变量

$$Q(u)=\sum_{j\in\Gamma(u)}N(j) \tag{3}$$

则 $Q(1)=N(2)+N(3)+N(4)+N(5)+N(6)+N(7)=8$, 于是: $C_L(1)=Q(2)+Q(3)+Q(4)+Q(5)+Q(6)+Q(7)=47$。

图 3 中每个节点的相关信息, 包括度数、接近中心性、介数中心性和局部集中度计算结果在表 2 中。

通常, 连接节点的边越重要, 节点越重要。节点权重向量 $W^v=[W_1,W_2,\cdots,W_n]^{\mathrm{T}}$ 可以根据节点关联边的拓扑结构获得:

$$W_i=\sum_{j\in\Gamma(i)}w_{ij} \tag{4}$$

另一方面, 动态地考虑流量对于节点关键性的影响。根据每对节点之间的流量值 $f_{mn}(m,n\in V)$, 可以得到流量矩阵 F 如右侧所示:

$$F=\begin{pmatrix} f_{11} & \cdots & f_{1n} \\ \vdots & \ddots & \vdots \\ f_{n1} & \cdots & f_{nn} \end{pmatrix}$$

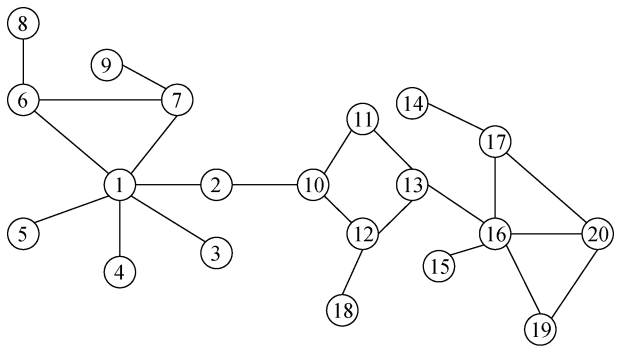


图 3 一个包含 20 个节点和 23 条边的网络

Figure 3 An example network consisted of 20 nodes and 23 edges

对于度量节点的重要性而言, 更需要关注节点总的流量大小而不是单个边的流量值。因此, 我们将计算每一个节点 i 的流量汇集情况 F_i, 即:

$$F_i=\sum_{j=1}^{n}f_{ij} \tag{5}$$

进而, 我们可以得到整个网络的流量向量 $\boldsymbol{F}^v=\left[F_1,F_2,\cdots,F_n\right]^{\mathrm{T}}$ 。

最后, 结合网络的拓扑结构和实况流量, 将节点权重向量 $\boldsymbol{W}$ 与流量向量 $\boldsymbol{F}$ 做 Hadamard 乘积可以得到排序向量 $\boldsymbol{R}=\boldsymbol{F}^v\odot\boldsymbol{W}^v$, 即

$$\boldsymbol{R}=\boldsymbol{F}^v\odot\boldsymbol{W}^v=\begin{bmatrix}F_1\\F_2\\\vdots\\F_n\end{bmatrix}\odot\begin{bmatrix}W_1\\W_2\\\vdots\\W_n\end{bmatrix}=\begin{bmatrix}F_1W_1\\F_2W_2\\\vdots\\F_nW_n\end{bmatrix} \tag{6}$$

该向量 $\boldsymbol{R}$ 是按照节点重要性大小进行排序后的结果, 向量中权值越大的元素对应的节点便越重要。

4 三种中心性方法与节点排序算法

4.1 三种传统中心性方法

本文后续算法是基于三种中心性算法(度中心性(degree centrality)、接近中心性(closeness centrality)和介数中心性(betweenness centrality))的思想形成的, 并且第 5 节实验部分选择与该三种方法进行对比。因此先介绍这三种方法。

定义 1.在无向图中, 度中心性测量网络中一个节点与所有其他节点相联系的程度。对于一个拥有 n 个节点的无向图, 节点 i 的度中心性是 i 与其他 n–1 个节点的关联边总数:

$$C_D(i)=\sum_{j=1}^{n}a_{ij}(i\neq j) \tag{7}$$

其中 $C_D(i)$ 表示节点 i 的度中心度, 若节点 i,j 之间有边, 则 $a_{ij}=1$; 否则 $a_{ij}=0$。

表 2　图 3 中每个节点的相关信息

Table 2　Information about each node in Figure 3

v	度数	C_B	C_C	Q	C_L
1	6	0.5789	0.3065	8	47
2	2	0.5146	0.3276	9	14
3	1	0	0.2375	6	8
4	1	0	0.2375	6	8
5	1	0	0.2375	6	8
6	3	0.1053	0.2500	10	21
7	3	0.1053	0.2500	10	21
8	1	0	0.2021	3	10
9	1	0	0.2021	3	10
10	3	0.5322	0.3393	6	22
11	2	0.2047	0.3167	6	15
12	3	0.3099	0.3276	7	18
13	3	0.4620	0.3065	9	25
14	1	0	0.1845	3	9
15	1	0	0.2159	5	12
16	5	0.4386	0.2714	12	43
17	3	0.1053	0.2236	9	25
18	1	0	0.2500	3	7
19	2	0	0.2194	8	22
20	3	0.0058	0.2235	10	29

度中心性是在网络分析中衡量节点中心性的直接度量指标之一。显然, 一个节点的节点度越大就意味着这个节点的度中心性越高, 该节点在网络中就越重要。节点度中心性取决于每个节点与其它节点的关联性, 同时也受网络规模大小$|V|$的影响。通常情况下, 网络中节点数目越多, 那么度中心性的最大绝对值就可能越大。而在现实度量的情况下, 为了消除网络规模对于度中心性数值大小的影响, Stanley Wasserman 等人根据归一化的思想提出了度中心性的标准化公式:

$$C'_D(i)=\frac{C_D(i)}{n-1} \tag{8}$$

度中心性可以反映网络中节点所处位置的中心程度, 一个节点与越多的节点邻接, 那么这个节点就越处于中心地位, 也就越重要。度中心性是衡量节点重要程度的最直接指标, 而接近中心性反映的是网络中的一个节点在传播信息时对其他节点的依赖程度, 即一个节点到其他节点的距离越短, 那么从该点开始传播信息对其余节点的依赖就越小。接近中心性的定义如下:

定义 2.在一个拥有 n 个节点的无向图中, 节点 i 的接近中心性是该点 i 到其他节点的最短距离之和的倒数:

$$C_C(i)=\frac{1}{\sum_{j=1}^{n}d(i,j)}(i\neq j) \tag{9}$$

其中 $C_C(i)$ 表示节点 i 的接近中心度, $d(i,j)$ 为节点 i 到节点 j 的最短路径长度。

对于一个节点, 它距离其他节点越近, 那么它的接近性中心性越大。同样地, 在不同网络之间进行度量的时候, 为了消除网络规模对于接近中心性绝对值大小的影响, 采用了标准化方法对定义进行了调整:

$$C'_C(i)=\frac{n-1}{\sum_{j=1}^{n}d(i,j)}(i\neq j) \tag{10}$$

另一常用指标——介数中心性, 则以经过某个节点的最短路径数目来刻画节点重要性的指标, 用来描述网络中节点承载最短路径数的能力。节点的介数等于网络中所有最短路径中经过该节点的概率之和, 该指标描述了节点在网络中的影响力与中心性程度。

定义 3.假设节点 i 和节点 j 之间的最短路径数为 δ_{ij} 条, 这两个节点之间经过节点 k 的最短路径条数为 $\delta_{ij}(k)$。比值 $\delta_{ij}(k)/\delta_{ij}$ 能描述节点 k 在节点 i 和节

点 j 之间的重要程度。在此基础上, 将节点 k 的介数定义为:

$$C_B(k)=\sum_{i\in V}\sum_{j\neq i\in V}\frac{\delta_{ij}(k)}{\delta_{ij}} \tag{11}$$

网络中, 介数常用来评价节点的流量承载能力。某个节点的介数 C_B 越大, 说明在信息传播过程中通过该点的数据量就越大, 一些用于数据传输的中枢节点通常使用率会很高。

4.2 本文的节点排序算法

本文提出的节点安全重要性排序算法的具体流程如下所示。首先输入主要参数为图 G 的邻接矩阵 $\boldsymbol{A}$, 图 G 的流量矩阵 $\boldsymbol{F}$。然后依次根据邻接矩阵 $\boldsymbol{A}$ 求出每个节点 i 的局部集中度 $C_L(i)$, 然后根据局部集中度向量求出每一条边的权值得到边权重矩阵 $\boldsymbol{W}$, 随即可得节点权重向量 $\boldsymbol{W}^v$ 与节点流量向量 $\boldsymbol{F}^v$。最后, 根据两向量 $\boldsymbol{W}^v$ 和 $\boldsymbol{F}^v$ 的 Hadamard 乘积, 即可输出排序结果。

算法 1. 节点安全重要性排序算法

1: 输入 $\boldsymbol{A}$、$\boldsymbol{F}$
2: WHILE $i<|V|$ DO
3: $C_L(i)=\sum_{u\in\tau(i)}\sum_{j\in\tau(u)}N(j)$
4: ENDWHILE
5: FOR $i=1$ to $|V|$ DO
6: FOR $j\in Neighbor(i)$ DO
7: $W[i,j]=C_L(i)*C_L(j)$
8: END FOR
9: END FOR
10: FOR $i=1$ to $|V|$ DO
11: FOR $j=1$ to $|V|$ DO
12: $W^v[i]=W^v[i]+W[i,j]$
13: $F^v[i]=F^v[i]+F[i,j]$
14: END FOR
15: END FOR
16: $R=F^v\odot W^v$
17: 输出 R

继续以图 3 为例, 由公式(4)可知节点 1 的权重 $W_1=\sum_{j\in\Gamma(1)}w_{1j}=w_{12}+w_{13}+w_{14}+w_{15}+w_{16}+w_{17}$, 其中 w_{1j} 可根据公式(1)和表 2 计算所得, 其余节点同理可得, 最终形成权重向量 $\boldsymbol{W}^v$。若给图 3 各边赋予权重, 可由(5)式得到整个网络的流量向量 $\boldsymbol{F}^v$, 最终得到排序结果 R。

算法 1 计算局部集中度时, 极端情况下最大算法复杂度为 $O(n+n^2)$, 但实际情况远小于这一数值, 因为真实网络中罕见有每对不同的节点之间都恰有一条边相连的完全图, 甚至更多的是稀疏图, 在稀疏图上最大算法复杂度为 $O(k+kt)$, 其中 k 为选定节点 i 的度数, t 为节点 i 临近节点集的平均度数。后面的二重循环最大算法复杂度为 $O(n^2)$, 但在实际应用中, 对于稀疏图, 可以使用如邻接表等辅助数据结构来加快算法运行速度, 使其实际运行时间小于 $O(n^2)$, 从而获得良好的实际效果。

5 实验分析

5.1 实验数据

为了使算法在真实网络中运行以检测局部重要性算法的性能, 我们使用了 US Top-500 Airport Network[27]数据集。该数据集包含了美国最为繁忙的 500 个商业机场, 每对机场之间的连线代表这两个机场在 2002 年有直飞航线, 其权重值代表该条航线可运载的旅客数目。交通图这一类的数据集可构成一个有向图 $G(V,E)$, 节点代表交通枢纽, 边则代表着两个交通站点之间有路线, 考虑到一般情况下, 每条航线的两个机场之间基本上都会相互通航, 所以该有向图 G 是高度对称的, 因此该数据集版本组成的是无向图。该数据集可以从 the Complex Networks Collaboratory's website 获得。

5.2 实验分析

每一种算法执行后, 可以得到节点按权重大小排序的结果, 显然, 一个节点承担越多的网络流量, 那么这个节点在整个网络中的重要性越高。有效的节点重要性排序方法应能够将越重要的节点排在越靠前的位置上。我们将本文所提出的算法与度中心性、接近中心性和介数中心性进行比较, 针对不同的变量 n, 我们将每种算法得出的有序列表中的前 n 个节点移除, 则与这些节点关联的边也将被删除, 通过观察这一操作后整个网络的流量衰减程度 α 的倒数 $\frac{1}{\alpha}$, 可以得到删除节点对网络的破坏程度。显然, $\frac{1}{\alpha}$ 越大就意味着网络被破坏的程度越严重, 说明被删除的 top-n 个节点越重要, 相对应的排序算法得出的结果就越可靠, 其剩余流量(衰减程度)百分比定义如下:

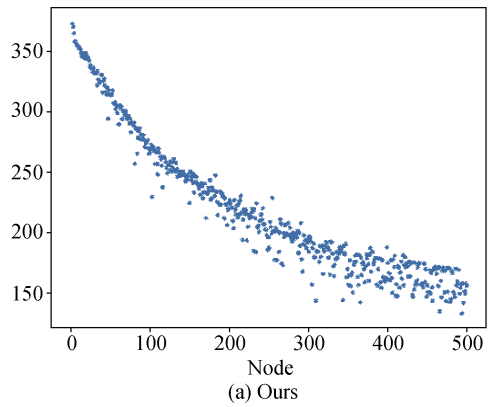
$$\alpha=\frac{f-f_{des}}{f}\times 100\% \tag{12}$$

其中,f 表示网络原有流量,f_{des} 表示删去 top-n 个节点所减少的流量值。

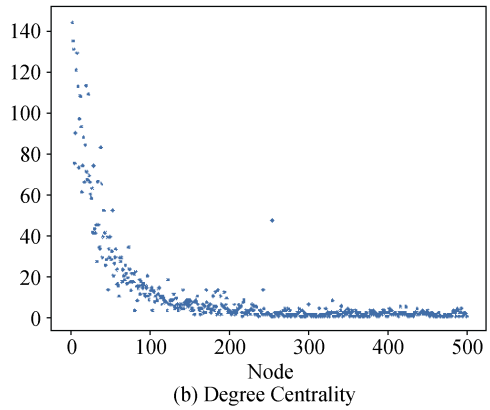
该数据集上四种方法的中心度计算结果如图 4 所示, 可以明显地看出网络中的节点具有长尾分布现象, 具有较大中心度的节点更容易向其他节点传播信息。为防止数据长度过大, 本文算法针对过程中向量进行对数函数 lg 变换。

分别将 top-3、top-5、top-10、top-15 和 top-20 的节点从网络当中删除, 观察网络整体流量的衰减程度, 结果见图 5, 可见删除我们的方法的 top-n 个节点对网络的破坏程度最为严重。

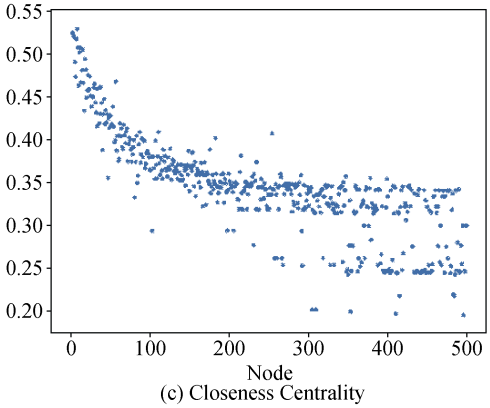
表 3 和图 6 对比了四种算法的实验数据, 可以看出在选取不同的 top-3、top-5、top-10、top-15 和 top-20 时, 本文提出的算法都具有明显优势, 即原网络中不同算法排出来的前 n 个节点, 我们的方法都能选出最为重要的相应节点。



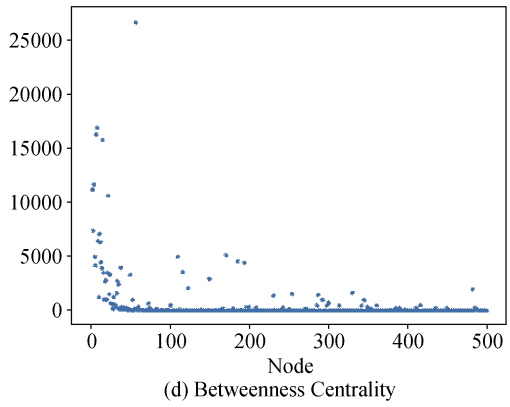
(a) Ours



(b) Degree Centrality



(c) Closeness Centrality



(d) Betweenness Centrality

图 4 四种方法的节点计算结果

Figure 4 The node calculation results of the four methods

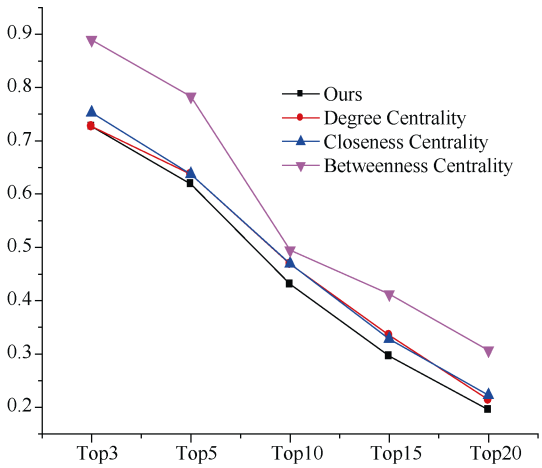


图 5 四种方法的不同 top-n 过程网络流量衰减折线图

Figure 5 Different top-n processes of network attenuation in the four methods

在破坏小规模节点集, 如 top-3 实验中, 通过暴力搜索求解出最大破坏程度的 3 个节点, 该最优解与本文提出的排序方法相吻合, 明显优于接近中心性(closeness centrality)和介数中心性(betweenness centrality)两种方法。

6 结论与展望

在复杂网络数据集 US Top-500 Airport Network 当中, 本文提出的节点安全重要性排序度量方法与度中心性(Degree Centrality)、接近中心性(Closeness Centrality)和介数中心性(Betweenness Centrality)相比, 在破坏网络 top-n 节点之后, 整体网络流量衰减情况表现出本文方法衰减最快; 次之, 度中心性与接近中心性相近; 最后为介数中心性方法, 该方法效果表现不及前述。

表 3 不同 top-n 过程的实验数据对比

Table 3 Experimental data comparison with different top-n process

Method	Top3	Top5	Top10	Top15	Top20
Ours	**1.3753210**	**1.6163288**	**2.3207779**	**3.3787322**	**5.1281525**
Degree Centrality	1.3753210	1.5692502	2.1344535	2.9868667	4.6802894
Closeness Centrality	1.3279313	1.5692502	2.1316189	3.0474333	4.4939983
Betweenness Centrality	1.1231879	1.2766924	2.0208838	2.4257184	3.2604195

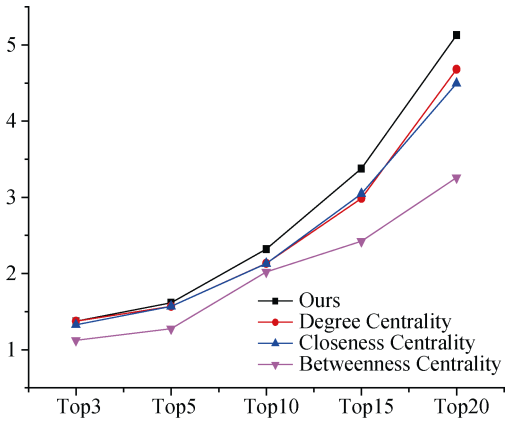


图 6 四种方法的不同 top-n 过程结果折线图

Figure 6 Results of different top-n processes in the four methods

显然，节点越重要则该节点被破坏后对网络整体流量的负面影响越大，出于保护网络整体安全程度的考虑，这一类的节点应当被合适的排序方法找出，并且加以重点保护。同时，关键的节点集合承担了网络中的大部分流量，他们极易成为被攻击的目标，如僵尸网络中被控制的关键计算机集群，交通网络中的关键交通枢纽，甚至是与其他部件关联性极强的漏洞，这些节点被破坏或操纵后，带来的危害远大于其余多数“不那么重要”的节点。

由于本文内容有限，在未来的工作中，我们将尝试在选择重要节点排序时引入随机方法，如随机游走(Random Walk)方法，通过每个结点的游走次数来决定其连接重要程度。另一方面，本文研究假设节点被完全破坏，即被攻击节点与相邻节点的所有通路全部被切断，而在实际情况中，可能存在被攻击节点与部分其他节点仍然可通信的情况，这种情况下，对实时排序结果又会产生不同的影响。

对于网络动态性的研究需要关注。仅从节点重要程度的角度来看，网络的动态性体现在以下两个方面[11]。一方面是从网络的动态发展来看，网络中会不断地涌现出新节点、节点之间会涌现出新的边，同样地，网络中也会不断地失去一些节点，或者失去一些节点之间的关系；网络自身的实时变化就决定了整个网络是一张演变着的动态图。而另一方面，从某个时间切片来看，尽管某一时刻下网络结构是静态的，但节点之间的关系存在多样性，关系种类和节点角色的变化同样影响节点关键程度。因此，对于复杂网络中节点重要性的研究，需要考虑这种发展和变化。

理论之外，在实际使用中，许多方法是在仿真实验平台上针对小规模实验仿真网络进行的模拟计算，当网络规模扩大之后，许多原算法会暴露出其在时空复杂度上的缺陷[11]，因此需要进一步研究近似算法在节点排序中的作用。

此外，实际网络中存在诸多细节因素，因此后续将考虑展开在网络实时攻防对抗条件下的网络节点安全重要性排序的工作，如何在攻防对抗场景下研究有效的节点排序算法也很有意义。

致 谢 本课题得到国家重点研发计划“网络系统安全度量方法与指标体系”项目(No. 2016YFB0800700)资助。

参考文献

[1] M.E.J. Newman, “The structure and function of complex networks,” *SIAM review*, vol. 45, no. 2, pp. 167-256, 2003.

[2] R. Guimera and L.A.N. Amaral, “Modeling the world-wide airport network,” *The European Physical Journal B*, vol. 38, no. 2, pp. 381-385, 2004.

[3] M.E.J. Newman, “The structure of scientific collaboration networks,” *Proceedings of the national academy of sciences*, vol. 98, no. 2, pp. 404-409, 2001.

[4] H. Jeong, B. Tombor, R. Albert, Z.N. Oltval and A.L. Barabási. “The large-scale organization of metabolic networks,” *Nature*, vol. 407, no. 6804, pp. 651-654, 2000.

[5] D.J. Watts and S.H. Strogatz, “Collective dynamics of ‘small-world’ networks,” *Nature*, vol. 393, no. 6684, pp. 440-442, 1998.

[6] L.C. Freeman, “Centrality in social networks conceptual clarification,” *Social networks*, vol. 1, no. 3, pp. 215-239, 1978.

[7] Y.Y. Ahn, S. Han, H. Kwak, S. Moon and H. Jeong, “Analysis of topological characteristics of huge online social networking services,” *Proceedings of the 16th international conference on World Wide Web (WWW’07)*, ACM, pp. 835-844, 2007.

[8] S.H. Strogatz, “Exploring complex networks,” *Nature*, vol. 410, no.

6825, pp. 268-276, 2001.

[9] X. Zhou, F.M. Zhang, K.W. Li, X.B. Hui and H.S. Wu 2012, "Finding vital node by node importance evaluation", *Acta Phys. Sin*, vol. 61, no. 5, pp. 0502011-0502017 (in Chinese), 2012.
(周漩, 张凤鸣, 李克武, 惠晓滨, 吴虎胜, "利用重要度评价矩阵确定复杂网络关键节点", *物理学报*, 2012, 61(5): 0502011-0502017.)

[10] J. Mirkovic and P. Reiher, "A taxonomy of DDoS attack and DDoS defense mechanisms," *ACM SIGCOMM Computer Communication Review*, vol. 34, no. 2, pp. 39-53, 2004.

[11] Y. Liu, Q.X. Wang and J.Y. Luo, "A survey on importance measurement of actors in complex network", *2007th Henan computer federation*, pp. 40-45, 2007.
(刘琰, 王清贤, 罗军勇, "复杂网络中节点重要性度量方法综述", *河南省计算机学会 2007 年学术年会*, 2007, pp. 40-45.)

[12] D. Chen, L. Lü, M.S. Shang, Y.C. Zhang and Z. Tao, "Identifying influential nodes in complex networks," *Physica a: Statistical mechanics and its applications*, vol. 391, no. 4, pp. 1777-1787, 2012.

[13] J.G. Liu, Z.M. Ren, Q. G and B.H. Wang, "Node importance ranking of complex networks", *Acta Phys. Sin*, vol. 62, No. 17, pp.1789011-1789019 (in Chinese), 2013.
(刘建国, 任卓明, 郭强, 等. 复杂网络中节点重要性排序的研究进展[J]. *物理学报*, 2013, 62(17): 1789011-1789019.)

[14] J.W. Wang, L.L. Rong and T.Z. Guo, "A new measure method of network node importance based on local characteristics," *Journal of Dalian University of Technology*, vol. 50, no. 5, pp. 822-826(in Chinese), 2010.
(王建伟, 荣莉莉, 郭天柱, "一种基于局部特征的网络节点重要性度量方法", *大连理工大学学报*, 2010, 50(5): 822-826.)

[15] D. Centola, "The spread of behavior in an online social network experiment," *Science*, vol. 329, no. 5996, pp. 1194-1197, 2010.

[16] L.C. Freeman. "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, no. 1, pp. 35-41, 1977.

[17] L.C. Freeman, D. Roeder, and R.R. Mulholland, "Centrality in social networks: II. Experimental results," *Social networks*, vol. 2, no. 2, pp. 119-141, 1979.

[18] K. Stephenson, M. Zelen, "Rethinking centrality: Methods and examples," *Social networks*, vol. 11, no. 1, pp. 1-37, 1989.

[19] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39-43, 1953.

[20] C.H. Comin and L.F. Costa, "Identifying the starting point of a spreading process in complex networks," *Physical Review E*, vol. 84, no. 5, pp. 056105, 2011.

[21] J. Zhang, X.K. Xu, P. Li, K. Zhang and M. Small, "Node importance for dynamical process on networks: A multiscale characterization," *Chaos: an interdisciplinary journal of nonlinear science*, vol. 21, no.1, pp. 016107, 2011.

[22] P. Berkhin, "A survey on pagerank computing," *Internet Mathematics*, vol. 2, no. 1, pp. 73-120, 2005.

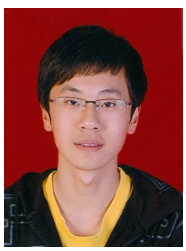
[23] L. Lü, Y.C. Zhang, C.H. Yeung and T. Zhou, "Leaders in social networks, the delicious case," *PloS One*, vol. 6, no. 6, pp. 1-7, 2011.

[24] J.M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, no. 5, pp. 604-632, 1999.

[25] J. Xu, Y.M. Xi and Y.L. Wang, "On system core and coritivity," *Journal of Systems Science and Mathematical Sciences*, vol. 13, no. 2, pp.102-110 (in Chinese), 1993.
(许进, 席酉民, 汪应洛, "系统的核与核度", *系统科学与数学*, 1993, 13(2): 102-110.)

[26] J. Chen and L.F. Sun, "Evaluation of Node Importance in Complex Networks," *Journal of Southwest Jiaotong University*, vol. 44, no. 3, pp. 426-429(in Chinese), 2009.
(陈静, 孙林夫, "复杂网络中节点重要度评估", *西南交通大学学报*, 2009, 44(3): 426-429.)

[26] V. Colizza, R. Pastor-Satorras, A. Vespignani, "Reaction–diffusion processes and metapopulation models in heterogeneous networks," *Nature Physics*, vol. 3, no. 4, pp. 276-282, 2007.



张子超 于 2017 年在兰州大学计算机科学与技术专业获得学士学位。现在北京大学计算机软件与理论专业攻读博士学位。研究领域为社交网络、网络空间安全等。
Email: zhangzch@pku.edu.cn。

郝蔚琳 于 2017 年在北京大学计算机软件与理论专业获得硕士学位。现在北京大学计算机软件与理论专业攻读博士学位。研究领域为生物信息学、网络安全等。
Email: haoweilin@pku.edu.cn。

张伊凡 于 2017 年在哈尔滨工业大学信息安全专业获得学士学位。现在北京大学计算机软件与理论专业攻读博士学位。研究领域为信息安全, 自然语言处理等。
Email: yifanzh@pku.edu.cn。