doi:10.7688/j.issn.1000－1646.2022.05.04

基于边聚类的电力通信网 Sybil 攻击检测算法

党晓婧[1]，张 林[1]，吕启深[1]，彭 浩[2]，张柏松[2]
（1. 中国南方电网有限公司 深圳供电局电力科学研究院，广东 深圳 518000；2. 深圳康托普信息技术有限公司 业务研究中心，广东 深圳 518034）

摘 要：针对电力通信网络频繁遭受 Sybil 攻击的问题，提出了一种基于 *K*-means 边聚类的 Sybil 攻击团体检测算法. 通过优化边聚类和边介数的计算方法，提出了传统 *K*-means 聚类算法的改进方法，计算了通信网络中的聚类系数，根据合法用户的真实数量，建立更加精确的攻击边集合与真实边集合，从而初步检测出所有可疑的攻击边，并使用标签传播算法检测 Sybil 攻击行为所在的恶意团体. 仿真结果表明，与经典的 SybilLimit 算法相比，在所有的攻击路径数量下，该 Sybil 攻击检测算法具有更加优秀的检测性能.

关 键 词：社交网络；Sybil 攻击；*K*-means 算法；攻击检测；边介数；边聚类；欧氏距离；标签传播
中图分类号：TM 73 **文献标志码**：A **文章编号**：1000－1646(2022)05－0502－05

Sybil attack detection algorithm for power communication network based on edge clustering

DANG Xiao-jing[1], ZHANG Lin[1], LÜ Qi-shen[1], PENG Hao[2], ZHANG Bai-song[2]
(1. Shenzhen Electric Power Science Research Institute, China Southern Power Grid, Shenzhen 518000, China; 2. Business Research Center, Shenzhen Comtop Information Technology Co. Ltd., Shenzhen 518034, China)

Abstract: In order to solve the problem that the power communication network had been frequently attacked by Sybil, a Sybil attackgroup detection algorithm based on *K*-means edge clustering was proposed. By optimizing the calculation methods of edge clustering and edge betweenness, an improved method in terms of traditional *K*-means clustering algorithm was suggested, and the clustering coefficient in the communication network was calculated. According to the real number of legitimate users, a more accurate set of attack edges and real edges was established so as to initially detect all suspicious attack edges. In addition, the label propagation algorithm was used to detect the evil group committing Sybil attack. The simulation results show that the as-proposed Sybil attack detection algorithm has better detection performance under all attack paths compared with the classic SybilLimit algorithm.

Key words: social network; Sybil attack; *K*-means algorithm; attack detection; edge betweenness; edge clustering; Euclidean distance; label propagation

为了保障电力系统安全、稳定运行，电力通信网需要制定一系列的安全措施，防止来自电力系统外部与内部的攻击. 电力系统遭受的攻击行为不但包含专门针对电力控制系统的攻击行为，也包含一些适用于互联网与社交网络的通用攻击方法. 在这些通用的攻击方法中，Sybil 攻击是一种广泛应用的攻击算法，Sybil 攻击又称为女巫攻击，其基本思想是利用网络中少数的通信节点控制数量较大的虚假身份，再使用大量的虚假身份对其他正常的通信节点发起攻击. 该攻击最早起源于点对点的通信网络攻击，广泛应用于社交网络和互联网等攻击方法中. 针对 Sybil 攻击的检测与预防，国内外的学者

收稿日期：2019－12－20.
基金项目：国家自然科学基金项目(61033004)；中国南方电网有限公司深圳供电局有限公司科技项目(090000GS62161590).
作者简介：党晓婧(1990－)，女，河南许昌人，工程师，硕士，主要从事电力设备状态监测与评价等方面的研究.

做出了大量的研究,其中,Yu 等[1] 提出了 SybilGuard 算法,检测可疑节点是否为攻击源头;随后,其又提出了分散式 SybilLimit 算法[2],利用随机路径末尾区分合法节点与可疑节点,从而减少了假阴性检测结果的数量. 众多学者在此基础上,做出了较多具有标志性的科研成果,然而,这些检测算法均存在检测精度较低或计算复杂度较高等缺点,不适用于大规模网络中 Sybil 攻击的检测.

针对这一关键问题,本文在详细分析 Sybil 攻击方法的基础上,通过引入边介数和 *K*-means 边聚类算法,提高 Sybil 攻击的检测效率,从而分离边介数较高的攻击边. 在此基础上,利用标签权值检测得到 Sybil 攻击的发起节点,最终完成 Sybil 攻击的检测.

1　边介数计算

一般而言,边介数定义为当前边的最短路径与网络图中所有最短路径的数量比值. 目前,经典算法计算介数的时间较长,难以应用于大规模网络[3-5]. 本文提出了一种边介数的计算算法,网络中有一个节点 a,不妨设起始节点为 s,目标节点为 t,则其介数中心性的定义如下:

1) 对于一个网络 $G(V,E)$,文中统一使用 G 表示网络图,V 和 E 分别表示网络的顶点集和边集. 节点 $a\in V$ 的介数中心性 $C_{\mathrm{B}}(a)$ 的计算公式为

$$C_{\mathrm{B}}(a)=\sum_{s\neq a}\sum_{t\neq a,s}\frac{\sigma_{st}(a)}{\sigma_{st}} \tag{1}$$

式中,σ_{st} 为节点 s 与 t 之间的所有最短路径数目;$\sigma_{st}(a)$ 为 σ_{st} 中包含节点 a 的路径数量. 若节点 s 与 t 之间是不连通的,则 $\dfrac{\sigma_{st}(a)}{\sigma_{st}}=0$.

2) 对于网络 $G(V,E)$,边 l 的边介数中心性 $W_{\mathrm{B}}(l)$ 的计算公式为

$$W_{\mathrm{B}}(l)=\sum_{s\neq t\in V}\frac{\sigma_{st}(l)}{\sigma_{st}} \tag{2}$$

式中,$\sigma_{st}(l)$ 为其中经过边 l 的最短路径数目. 若节点 s 与 t 之间是不连通的,则 $\dfrac{\sigma_{st}(l)}{\sigma_{st}}=0$.

3) 设 ρ 为从节点 s 到节点 v 的路径,$R(u_i)$ 是 u_i 的邻居节点集合,$P[\rho_{s,v}]$ 是从节点 s 把节点 v 作为下一个节点的概率,则顶点 h 的 k 路径边介数中心性 $W_k(h)$ 的计算公式为

$$W_k(h)=\sum_{s\neq h}\sum_{1\leqslant v\leqslant k}\sum_{|\rho_{s,v}|=v}\chi[h\in\rho_{s,v}]P[\rho_{s,v}] \tag{3}$$

$$\chi[i]=\begin{cases}1 & (i\in\rho_{s,v})\\ 0 & (i\notin\rho_{s,v})\end{cases} \tag{4}$$

$$P[\rho_{s,v}]=\prod_{i=1}^{v}\frac{1}{|R(u_{i-1})-\{s,u_1,\cdots,u_{i-2}\}|} \tag{5}$$

在大规模网络中,为了充分利用边介数的性质,本文引入随机游走策略,提出恰当的边介数计算算法,其流程描述为:

1) 输入为网络 $G(V,E)$,迭代轮数 T,随机路径长度 L,边的权值系数 β,输出边介数参数 S;

2) 初始化边介数参数 $S=0$,计算所有节点权值分配值,所有边权值分配值;

3) 若 $i\leqslant T$,步长计数器 $N=0$,选取起始节点;

4) 若步长计数器 $N\leqslant L$,且其邻接边集合元素数量大于标记值,则重复执行这些程序;

5) 将所有边的标记值置为 0;

6) 对于所有的边,设定边介数参数分配值,同时,返回其边介数参数值.

2　边聚类算法

利用边介数算法能够实现真实边与可疑边的初步分离[6]. 为了进一步鉴别可疑边中的真实用户,同时克服传统 *K*-means 算法易受干扰的缺点,基于随机游走边介数的特征[7-8],本文使用优化初始化中心和存储的方法,对传统 *K*-means 聚类算法进行必要的改进,从而更加精确地检测攻击边.

2.1　算法改进

在介绍算法改进策略前,需要引进边聚类系数的概念. 边聚类系数衡量了网络中任意两个节点之间的紧密关系程度[9-10]. 对于一个网络 $G(V,E)$,边 $e\in E$,边的顶点为 q 和 h,则 e 的边聚类系数 $C(e)$ 的计算表达式为

$$C(e)=\frac{1}{2}[C(q)+C(h)] \tag{6}$$

式中,$C(q)$ 与 $C(h)$ 分别为顶点 q 和 h 的聚类系数值.

在传统 *K*-means 算法中,初始中心点对聚类结果的影响较大[11],因此,如何选择更恰当的初始中心点,尽量将初始中心点分布于不同的类簇中是优化 *K*-means 算法的重要策略. 本文对于初始中心点的选取步骤如下:

1) 随机选择网络 $G(V,E)$ 中的一个边,将这个边的数据点设定为中心点;

2) 对于已知的数据点,计算和存储该点与其邻近中心点的最小距离 d_i,把这些距离相加可得 D;

3) 选取某个点的中随机值 $r\in[0,D]$,重复将中随机值 r 累加到最小距离 d_i 中,直至 $r>D$,此时,该点即为下一个中心点;

4）反复执行步骤 2）~3），直至中心点的个数达到聚类数.

选取初始中心点的算法需要频繁存储与查询最小距离数据，这也是传统 *K*-means 算法待优化之处. 为了解决这一问题，本文使用键值的形式，存储边到边索引信息与中心点之间的距离信息.

通常每个中心点需要存储边索引、边介数和边聚类系数等信息[12]. 其中，中心点之间的距离可以用边介数与边聚类系数计算获取，因此，本文使用拉链法存储边到边索引和距离信息. 例如，中心点 0 ~ 1 之间的边为 0，这两点之间的距离为 0.23，则存储为（“0 ~ 1”，0.23）.

若需要增加一个存储元素，则使用键值信息生成散列码，获取数组的索引. 若产生冲突，则令新元素的引用指向与其冲突的元素，使用单链表的形式处理冲突；若需要查询某一个键值信息，则使用该存储结构快速查找到某个中心点的信息，从而提高存储与查询的效率.

2.2 算法步骤

利用优化初始中心点的选取策略与中心点距离的存储方法[12-13]，本文提出了基于 *K*-means 的边聚类算法，其具体执行步骤为：

1）选取 K 个初始中心点；

2）按照 K 个初始中心点，将所有网络节点划分到距离其最近的中心点所属类别中；

3）对于所有类别，计算其数据平均值，生成新的中心点；

4）使用新的中心点，继续对所有的数据点进行划分；

5）反复执行步骤 3）~4），直至划分结果收敛，同时返回最后的聚类结果. 需要说明的是，该聚类算法需要频繁地计算某个网络节点 (x,y) 到中心点 (x_0,y_0) 的距离.

3 团体检测算法

利用边介数和边聚类的计算算法可以检测得到所有的可疑边. 为了进一步精确地检测攻击边，文中提出了一种基于标签权值的团体检测算法. 该算法以网络图 $G(V,E)$ 和可疑边集合为输入，输出只包含实施 Sybil 攻击的节点集合. 基于标签权值的检测算法基本思想为：通过更新种子集节点的标签，令所有节点的标签传播值达到最大.

与传统的标签传播算法相比[14]，本文使用异步的方式更新标签，设迭代轮数为 i，起始节点是 s，其邻近节点主要有 $(s_{j1},\cdots,s_{jm},s_{j(m+1)},\cdots,s_{jk})$，则其标签选择函数 $Z_s(i)$ 表达式为

$$Z_s(i)=f(Z_{s_{j1}}(i),\cdots,Z_{s_{jm}}(i),Z_{s_{j(m+1)}}(i),\cdots,Z_{s_{jk}}(i)) \tag{7}$$

由式(7)可知，第 i 轮的迭代标签由第 $i-1$ 轮、邻近节点 $(s_{j1},s_{j2},\cdots,s_{jm})$ 第 i 轮和邻近节点 $(s_{j(m+1)},s_{j(m+2)},\cdots,s_{jk})$ 第 $i-1$ 轮的标签状态决定，而函数 f 的功能是选择影响力最大的标签. 在将节点变化的扩散过程中，本文需要频繁地更新网络中的顶点标签. 起始节点为 s，其邻近节点标签为 m 的标签传播值为 $mp(m)$，则其标签更新表达式为

$$mp(m)=\frac{\sum C(m_i)}{N(m)}+\left(1-\mathrm{e}^{-\sum \deg(m_i)}\right) \tag{8}$$

式中：$C(m_i)$ 为邻近节点标签为 m 的边聚类系数；$N(m)$ 为标签为 m 的节点个数；$\sum \deg(m_i)$ 为标签 m 的顶点度数之和. 基于标签更新的表达式，s 的标签选择函数 $M(s)$ 为

$$M(s)=\arg\max mp(m)=\arg\max\left[\frac{\sum C(m_i)}{N(m)}+\left(1-\mathrm{e}^{-\sum \deg(m_i)}\right)\right] \tag{9}$$

团体检测算法的具体步骤为：

1）利用边聚类结果，构造种子集；

2）根据节点度的大小，对所有节点进行从大到小的排序；

3）按照步骤 2）的顺序结果，使用式(8)更新所有节点的标签；

4）测试所有节点的标签传播值是否最大，若是，则算法终结，返回结果；否则，反复执行步骤 3）~4）.

4 仿真验证与分析

为了验证攻击检测算法的有效性，本文利用不同规模的数据集，分别对攻击检测算法与 SybilLimit 算法进行仿真. 此外，文中还详细地对比分析了这两种算法的仿真结果.

4.1 仿真数据与环境

本文选用来自 Amazon 的通信网络数据集，Amazon 数据集拥有 335 874 个通信节点和 924 589 条边. 对该通信网络的数据集进行一定的扩展和操作，即可得到本文的实验数据集. 其具体的过程描述为：从真实数据集中，随机选取被攻击节点，与 Sybil 节点共同构成攻击边，然后利用 PA（preferential attachment）模型构造实验所需的拓扑结构.

此外，在仿真实验中，Sybil 攻击节点的数量

分别被设置为 10 000、6 000 和 1 000,攻击路径数量分别设定为 20 ~ 90. 使用的软件开发工具为 MyEclipse 9. 0,编程语言是 Java,硬件配置为 Intel Core i7-4790 处理器.

4. 2　评价指标

为了准确地评价攻击检测算法和 SybilLimit 算法的效果,本文使用假负率和模块化度量等指标进行评价. 设 n_{tp} 是检测得到的真实攻击节点数量,n_{fn} 是被检测为合法节点的攻击节点数量,则假负率 R_{fn} 的计算表达式为

$$R_{fn}=\frac{n_{fn}}{n_{fn}+n_{tp}} \tag{10}$$

此外,令 x 代表已划分团体,X 为总团体的数量,e_c 为团体 x 中边数量,q_c 为团体 x 到其他团体的边数,A 为所有边的数量,则模块化度量 Q 的计算表达式为

$$Q=\sum_{x\in X}\left[\frac{e_c}{A}-\left(\frac{2e_c+q_c}{2A}\right)^2\right] \tag{11}$$

4. 3　仿真结果与分析

在上述仿真数据集与软硬件环境下,本文对攻击检测算法和 SybilLimit 算法进行仿真. 两种算法在不同节点下的假负率仿真结果如图 1 ~ 3 所示.

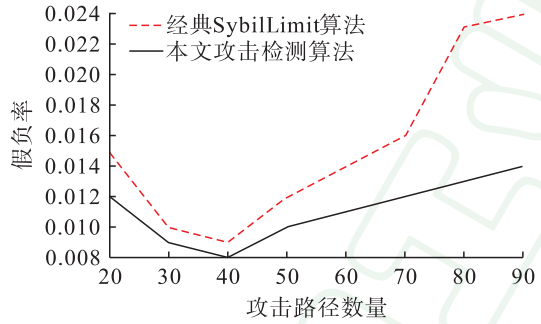

图 1　10 000 个 Sybil 攻击节点时假负率仿真图

Fig. 1　Simulation of false negative rate with 10 000 Sybil attack nodes

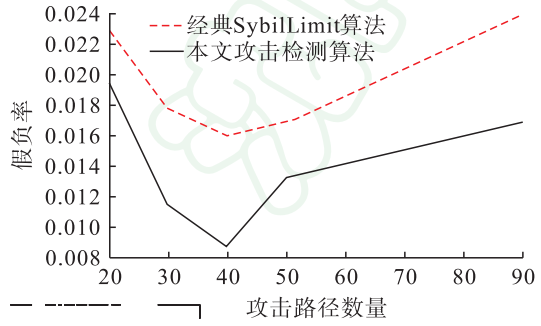

图 2　6 000 个 Sybil 攻击节点时假负率仿真图

Fig. 2　Simulation of false negative rate with 6 000 Sybil attack nodes

由图 1 ~ 3 可知,在算法与模型相同的情况下,随着攻击路径数量增加,假负率 R_{fn} 先降低,再

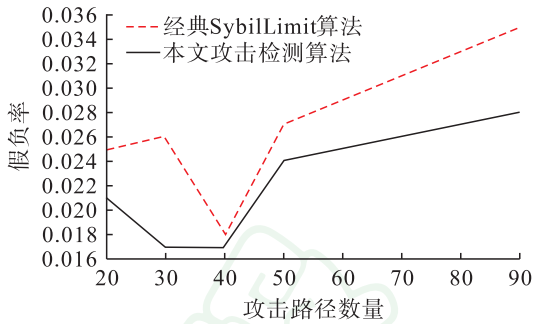

图 3　1 000 个 Sybil 攻击节点时假负率仿真图

Fig. 3　Simulation of false negative rate with 1 000 Sybil attack nodes

逐步提高. 其主要原因是在攻击路径增加的时候,Sybil 攻击的群体特征逐渐显现,而当攻击路径数量小于 40 的时候,由于群体特征不够明显,所以算法的检测难度较高,随着攻击路径数量的增加,算法的检测难度降低,此时假负率 R_{fn} 也逐渐减少;当攻击路径数量超过 40 之后,通信网络将遭受更多攻击,其攻击种类也在大量增加,这直接导致算法检测攻击的难度增加,同时假负率 R_{fn} 也逐渐提高. 在路径数量与模型相同的情况下,本文攻击检测算法的假负率 R_{fn} 显著低于经典 SybilLimit 算法,其主要原因是本文攻击检测算法的设计保留了经典 SybilLimit 算法的优点,同时避免了经典算法存在的缺点;由图 1 ~ 3 的比较可知,所提攻击检测算法和经典 SybilLimit 算法随着攻击节点的增加,其检测效果也更加优秀. 这表明,本文的攻击检测算法适用于大规模网络的攻击检测,而且其假负率指标水平低于经典 SybilLimit 算法.

在 10 000 个 Sybil 攻击节点和 PA 模式下,本文还统计了 SybilLimit 算法和攻击检测算法的模块化度量指标,其指标随路径数量变化的统计结果如图 4 所示.

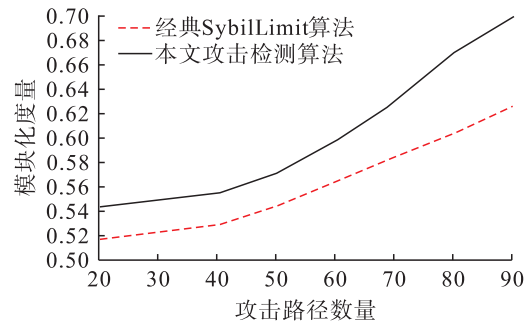

图 4　模块化度量仿真结果对比

Fig. 4　Comparison of simulation results for modularity measurement

一般而言,模块化度量主要衡量网络结构强度,该指标越大,表明算法的团体检测结果更加准确. 根据图 4 可知,本文攻击检测算法和经典 Sybil-

Limit 算法的模块化度量指标随着攻击路径的增加而提高,而在各种相同的外部条件和路径数量下,本文攻击检测算法的模块化度量指标要大于经典 SybilLimit 算法,这些仿真结果进一步表明,本文提出的算法具有更加优秀的团体检测结果.

综合假负率和模块化度量的对比结果可以看出,文中提出的攻击检测算法的表现优于经典SybilLimit算法,这是因为文中提出的攻击检测算法使用了更加合理的边聚类算法和标签传播算法,增大了真实节点和 Sybil 攻击节点之间的差距,降低了真实节点被误判为攻击节点的概率,从而实现了更加精确的团体攻击检测.

5 结 论

基于边介数和边聚类计算算法,本文提出了适用于电力通信网的 Sybil 攻击检测算法. 仿真结果表明,该算法的表现优于经典 SybilLimit 算法. 但是,这种攻击检测算法还可能存在一定的缺陷,这是因为文中所有仿真均使用了非常成熟的静态数据,尽管这些攻击数据来源于实际网络状态的统计,然而现实网络环境非常复杂,攻击技术和手段更新周期变化非常快,所以该算法是否可以动态检测现实网络的 Sybil 攻击,依然是未知的,需要进一步的研究与改进,下一步将致力于解决该问题.

参考文献(References):

[1] Yu H F, Kaminsky M, Gibbons P B, et al. SybilGuard:defending against Sybil attacks via social networks [J]. IEEE Transactions on Networking, 2008, 16(3):576 – 589.

[2] Yu H F, Gibbons P B, Kaminsky M, et al. SybilLimit: a near-optimal social network defense against sybil attacks [J]. IEEE/ACM Transactions on Networking, 2010, 18(3):885 – 898.

[3] Vinayagam S S, Parthasarathy V. A secure restricted identity-based proxy re-encryptionbased routing scheme for Sybil attack detection in peer-to-peer networks [J]. Journal of Computational and Theoretical Nanoscience, 2018, 15(1):210 – 221.

[4] Ying Z, Ziwen S. Double level Sybil attack detection scheme for heterogeneous industrial wireless sensor networks [J]. Information & Control, 2018, 47(1):41 – 47.

[5] Pecori R. A trust and reputation method to mitigate a Sybil attack in kademlia [J]. Computer Networks, 2016, 94(7):205 – 218.

[6] Jhaveri H, Jhaveri H, Sanghavi D, et al. Sybil attack and its proposed solution [J]. International Journal of Computer Applications, 2014, 105(3):17 – 19.

[7] Yu B, Huang M, Huang Y, et al. Adaptive link fingerprint authentication scheme against Sybil attack in ZigBee network [J]. Journal of Electronics & Information Technology, 2016, 38(10):2627 – 2632.

[8] Sharma T, Singh L. Analysis of non simultaneous Sybil attack on DSR [J]. International Journal of Computer Applications, 2015, 109(9):8 – 10.

[9] Anamika P, Mayank S. Detectionand prevention of Sybil attack in MANET using MAC address [J]. International Journal of Computer Applications, 2015, 122(21):20 – 23.

[10] 费贤举,李虹,田国忠. 基于特征加权理论的数据聚类算法 [J]. 沈阳工业大学学报,2018,40(1):77 – 81.
(FEI Xian-ju, LI Hong, TIAN Guo-zhong. Data clustering algorithm based on feature weighting theory [J]. Journal of Shenyang University of Technology, 2018, 40(1):77 – 81.)

[11] 杨慧婷,杨文忠,殷亚博,等. 基于深度信念网络的 *K*-means 聚类算法研究 [J]. 现代电子技术,2019,42(8):145 – 150.
(YANG Hui-ting, YANG Wen-zhong, YIN Ya-bo, et al. Research on *K*-means clustering algorithm based on deep belief network [J]. Modern Electronics Technique, 2019, 42(8):145 – 150.)

[12] 赵凯,侯玉强. 基于自组织映射神经网络 *K*-means 聚类算法的风电场多机等值建模 [J]. 浙江电力,2019,38(8):30 – 36.
(ZHAO Kai, HOU Yu-qiang. Multi-machine equivalent modeling of wind farms using SOM-based *K*-means clustering [J]. Zhejiang Electric Power, 2019, 38(8):30 – 36.)

[13] 袁兆祥,余春生. 基于 DBSCAN 聚类的电力工程数据完整性分析 [J]. 沈阳工业大学学报,2019,41(3):246 – 250.
(YUAN Zhao-xiang, YU Chun-sheng. Integrity analysis of power engineering data based on DBSCAN clustering [J]. Journal of Shenyang University of Technology, 2019, 41(3):246 – 250.)

[14] 王永程,孟艳红. 针对有向社交网络的 Sybil 检测方法 [J]. 西安电子科技大学学报,2016,43(2):199 – 204.
(WANG Yong-cheng, MENG Yan-hong. Sybil detection method for directed social networks [J]. Journal of Xidian University, 2016, 43(2):199 – 204.)

(责任编辑:景 勇 英文审校:尹淑英)