doi :10.3969/j.issn.1671-1122.2016.06.006

一种改进的图分割算法在用户行为异常检测中的应用

杨连群[1]，温晋英[1]，刘树发[2]，王峰[3]

（1. 天津市滨海新区公安局，天津 300450；2. 天津市公安局，天津 300000；3. 91655 部队，北京 100036）

摘　要：基于模拟随机流的马尔科夫分类算法（Markov Cluster Algorithm，MCL）是一种快速且可扩展的无监督图分割算法，在用户行为异常检测中具有广泛的应用，但时间复杂度为 $O(N^3)$，不利于处理海量数据。为提高分割质量，同时减少计算时间，文章提出了一种改进的 MCL 模型。采用调整互信息（Adjusted Mutual Information，AMI）指标，对不同时间的图分割结果的相似度进行比较，判断是否有异常发生。实验表明，相较多层图分割算法（METIS），文章所提出的改进的 MCL 模型具有以下优点：1）无需事先规定聚类的数目；2）不易被数据中的拓扑噪声所影响；3）适合处理长尾分布的数据；4）在计算时间一定的情况下，能获得质量较高的分割结果。

关键词：图分割；马尔科夫分类算法；异常检测；多层图分割算法

中图分类号：TP309　　**文献标识码：**A　　**文章编号：**1671-1122（2016）06-0035-06

中文引用格式：杨连群，温晋英，刘树发，等．一种改进的图分割算法在用户行为异常检测中的应用 [J]. 信息网络安全，2016（6）：35-40.

英文引用格式：YANG Lianqun, WEN Jinying, LIU Shufa, et al. An Improved Graph Partitioning Algorithm for User Behavior Abnormal Detection [J].Netinfo Security, 2016 (6) : 35-40.

An Improved Graph Partitioning Algorithm for User Behavior Abnormal Detection

YANG Lianqun[1], WEN Jinying[1], LIU Shufa[2], WANG Feng[3]

(1. *Binhai New Area Public Security Bureau, Tianjin* 300450, *China*; 2. *Tianjin Public Security Bureau, Tianjin* 300000, *China*; 3. *Army* 91655, *Beijing* 100036, *China*)

Abstract: The MCL algorithm is short for Markov Cluster Algorithm, a fast and scalable unsupervised partitioning algorithm for graphs. MCL has been widely applied to anomaly detection. However, it requires $O(N^3)$ time, Which is no good for a wide range of data processing. In order to improve the quality of clustering and save time consumption, an improved MCL algorithm is proposed. With AMI (Adjusted Mutual Information) index, the similarities of clusters of different periods are compared to determine whether the abnormal behavior occurs. Compared with multilevel k-way partitioning scheme (METIS) for graphs, experiments show that the proposed MCL algorithm has following strengths: 1) Number of clusters not specified ahead of time. 2) Robust against noise in graph data. 3) Suitable for clusters with long tail distribution. 4) Produce better clustering results in case of certain time consumption.

Key words: graph partitioning; Markov Cluster Algorithm; abnormal detection; METIS

收稿日期：2016-05-15

基金项目：国家高技术研究发展计划（国家 863 计划）[2013AA01A214]

作者简介：杨连群（1978—），男，天津，高级工程师，硕士，主要研究方向为计算机算法；温晋英（1980—），男，天津，工程师，主要研究方向为计算机算法；刘树发（1978—），男，天津，工程师，硕士，主要研究方向为软件工程；王峰（1977—），男，山东，博士，主要研究方向为计算机系统结构。

通信作者：曾凤 916211986@qq.com

0 引言

近年来，由于计算机网络的广泛使用和网络之间信息传输量的急剧增加，入侵行为的案例不断增长，其攻击手段也日趋复杂，网络信息安全问题日益严重。异常检测作为一种积极主动的安全防御技术，已经成为信息安全领域的热门研究课题。它通过对入侵行为过程和特征的研究，提供了对内部攻击、外部攻击和误操作的实时保护，结合联动机制，在网络系统受到危害之前拦截并响应入侵行为。

在异常检测[1]中，图分析[2-5]广泛应用在账户交易异常检测、不同事件关联等多种场景下。与机器学习类算法比较，图分析方法符合人的思维方式，分析过程能实现直观可视。例如，从一张分析图可快速定位爆发次数最多的病毒、违规使用同台机器的用户、读写过同一个USB设备的机器。

但是，随着分析图的尺寸增大，分析过程会变慢，因此，图分割算法被提出。它利用实际经验中图的社区性特征，把图分割成若干个强连通的区域，对每一个区域进行分析和可视化操作。

近年来，人们在图分割算法上取得了很多成果，如谱分割算法[6-8]、多层图分割算法(METIS)[9-11]等，但在实际运用中仍然有很多不足之处。

1) 谱分割算法由于需要计算矩阵的特征值和特征向量，导致计算时间太长；2) METIS算法虽然以计算速度快著称，但它是将图的权重进行均等划分，不适合长尾分布的数据，且每层之间依赖性较高，较易产生噪声点；3) 谱分割算法和METIS都需借助先验知识定义递归终止条件，即不具有智能识别图类别数目的能力。因此找到一种计算时间较快、分割质量高且无需事先规定聚类数目的图分割算法是很有必要的。

基于模拟随机流的马尔科夫分类算法(Markov Cluster Algorithm，MCL)[12-16]是一种比较新的快速图形聚类算法，较以往的图分割算法，MCL算法在异常检测中的应用具有以下优势。

1) 它是基于随机流的一种方法，通过图中节点之间转移概率矩阵的简单几何运算和反复修改，模拟随机流更易理解；2) 不易被数据中的拓扑噪声所影响；3) 无需预先了解有哪些潜在的簇结构。但MCL也有一些不足之处，例如，时间复杂度为$O(N^3)$，不适合图直径较大的图等。

本文提出一种基于改进的MCL的异常检测算法。首先，对图进行预处理，去掉一些对分割结果影响不大的节点以缩短图直径，从而加快得到高质量的分割结果的速度；再将去除的点归于相应的簇；最后，使用调整互信息(Adjusted Mutual Information，AMI)指标衡量不同时间图分割结果的相似度，判断是否有异常行为。实验表明，相较METIS算法，本文提出的方法具有以下优点：1) 无需事先规定聚类的数目，具有智能识别图的内部结构的能力；2) 不易被数据中的拓扑噪声所影响；3) 适合处理长尾分布的数据；4) 在计算时间一定的情况下，能获得质量较高的分割结果。

1 改进的MCL算法

1.1 马尔科夫过程和随机游走

设随机过程$\{X(t),t\in T\}$，其状态空间为S。若对参数集T中任意n个数值$t_1<t_2<\cdots<t_n,(n\geqslant 3,t_i\in T)$，有

$$P\{X(t_n)\leqslant i_n|X(t_1)\leqslant i_1,\cdots,X(t_{n-1})\leqslant i_{n-1}\}$$
$$=P\{X(t_n)\leqslant i_n|X(t_{n-1})\leqslant i_{n-1}\}$$

则称过程$\{X(t),t\in T\}$具有马尔科夫性，并称此过程为马尔科夫过程。通俗地说，就是在已经知道过程“现在”的条件下，其“将来”不依赖于“过去”。

随机游走基于这样一个假设：当划分图时，会发现同一类的内部节点连边比较多，而不同类之间连边相对较少。因此，如果以图中某一点为起点，随机游走到下一个节点时，则游走到同一类中节点的概率大于游走到不同类中节点的概率，即该节点很可能在某一固定类中游走，而非在不同类之间来回游走。

1.2 转移概率矩阵

一个图通常表示为$G=(V,E,W)$，其中V是节点的集合，E是边的集合，W是边权重。记$\boldsymbol{A}$是图G的邻接矩阵，$\boldsymbol{M}$为图G的转移概率矩阵，$\boldsymbol{M}_{(i,j)}$代表节点v_i到节点v_j的转移概率，因此矩阵$\boldsymbol{M}$每列之和为1。定义转移概率矩阵$\boldsymbol{M}$与邻接矩阵$\boldsymbol{A}$之间的关系如下：

$$M_{(i,j)}=\frac{A_{(i,j)}}{\sum_{k=1}^{n}A_{(k,j)}} \quad (1)$$

MCL算法是通过反复修改转移概率矩阵来模拟图中节点的转移过程，该过程则是通过扩展(Expansion)操作和膨胀(Inflation)操作来实现的。

1.3 扩展操作和膨胀操作

扩展操作通过扩展参数来对转移概率矩阵 M 进行幂乘操作。当扩展参数为 e 时，该操作如下：

$$M_{exp}=Expand(M,e)=M^e \quad (2)$$

它将两个节点之间的相似度转化成两个节点与共同邻接点的相似度的乘积之和，直至它不再改变为止。扩展参数 e 越大，即经历的步数越多，则流出所在类而进入其他类的概率就越大。扩展操作实则为求马尔科夫链的极限分布的过程，目的是将图中不同的区域连接起来。

膨胀操作先通过膨胀参数对矩阵 M 中的每一列进行幂乘操作，再对每一列进行归一化操作。当膨胀参数为 r 时，该操作如下：

$$M_{inf}=Inflate(M,r)=\frac{M(i,j)^r}{\sum_{k=1}^{n}M(k,j)^r} \quad (3)$$

膨胀操作使得矩阵 M 中概率大的更大，概率小的更小，这就保证联系频繁的节点更加紧密，联系稀疏的节点更加疏远。同时，为保证矩阵 M 每列元素之和为 1，需要对矩阵进行归一化处理，即需要将每个元素除以整列元素之和。

1.4 改进的MCL 算法描述

基于模拟随机流的 MCL 算法是由 DONGEN 在文献 [12] 中提出的，一种新的图聚类算法，该算法无需预先了解有哪些潜在的类结构。转移概率矩阵 M 中第 j 列可以看成是从第 j 个节点流出到其他节点的流。每个节点在各个可能的方向都会有遍历的机会，且通常选择不同的路径。该算法的关键思想就是“随机漫游者抵达稠密的类后，在抵达大部分节点之前不会轻易离开该类”。MCL 算法针对转移概率矩阵 M，交替使用扩展和膨胀两个操作反复迭代，不断修改 M 的值，从而使得 M 达到稳定状态实现聚类。MCL 算法实现步骤如下：

输入：无向图 G（有权图或无权图均可），扩展参数 e 以及膨胀参数 r。

1）计算图 G 的邻接矩阵 A。

2）对每个节点添加自环，即 $A:=A+I$，I 为对角线元素为 1 的对角矩阵。相当于每个节点都有一步到本节点的路径，目的是使随机游走的步数既有奇数也有偶数。

3）利用公式（1），计算转移概率矩阵 M。

4）利用公式（2），对 M 进行扩展操作。取扩展参数 $e\in N>1$，模拟在当前的转移矩阵随机游走 e 步。

5）利用公式（3），对 M 进行膨胀操作，膨胀参数 $r\in R^+$。

6）将每列中接近零值的元素移除，目的是为节省存储空间与计算时间。

7）重复步骤 5），步骤 6），直至收敛。判断本次矩阵 M 与上次矩阵 M 是否有变化，若无变化，则表示收敛。

8）根据最后得到的矩阵 M 进行划分。矩阵 M 中的点分为两类：主动吸引节点和被动吸引节点。在相同的行，主动吸引节点至少有一个值为正的被吸引节点，且每一行具有相同的值。如图 1 所示，在第一行中，$M(1,1)$,$M(1,6)$,$M(1,7)$,$M(1,10)$ 元素的值均为 1.0，说明节点 1 是主动吸引节点，6、7、10 是被节点 1 所吸引的节点，因此将 1、6、7、10 归为一类。以此类推，最后得到的聚类结果为 {1、6、7、10}，{2、3、5}，{4、8、9、11、12}。

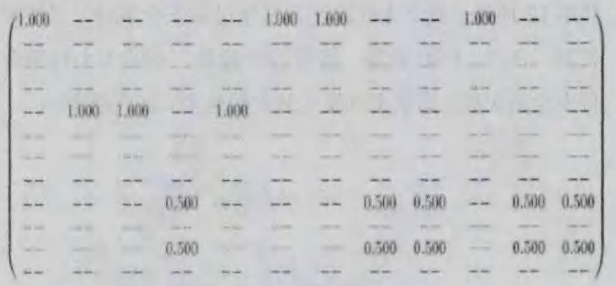


图1 矩阵M聚类结果

虽然文献 [12] 中并没有证明该算法的收敛性，但证明了 MCL 算法很容易在经过有限次的扩展和膨胀之后结束，且大量实验也说明，MCL 算法一般经历 10~100 扩展和膨胀次便可达到稳定状态。因此，随着图的尺寸变大，MCL 算法可扩展性比较好。

MCL 算法也有一些不足之处：1）时间复杂度为 $O(N^3)$；2）不适合直径太大的图。因为直径太大，相同类之间的“跨度”更大，需要更多的扩展操作和较少的膨胀操作，否则相同的类很可能会被分开。但随着迭代次数的增加，MCL 算法也会变得更敏感，会降低分割质量。因此，适当减少图的直径以及降低时间复杂度，在实际应用中具有重要意义。

2 基于改进的 MCL 算法的异常检测模型

在很多实际情况下，图的度分布符合幂律分布定理（Power-law）[16,17]，即绝大多数事件的规模很小，而只有少数事件的规模相当大。利用这一定理，首先对数据进行预

处理，去掉一些对聚类结果影响不大的节点，如度比较小的点，从而大大减少处理的数量，降低时间复杂度，同时也缩短图的直径，加快得到高质量的分割结果；之后计算度比较小的点与每个类的相似度，将这些点归于相应的类。

假设企业内各种事件实体之间的交互关系在较长的一段时间内稳定不变。参照图计算术语，把企业固定时间周期（如一天）内发生的各种操作转化为一张无向图，图中一个点代表一个操作的源或者目的，一条边代表源节点对目的节点的操作。异常检测模型基于这样的假设：在正常情况下，当前周期和前几个周期的图分割结果应保持基本一致；当企业内部发生异常事件时，当前周期和前几个周期的图分割结果将出现较大的差异。

图 2 所示某用户企业中 5 类安全事件：用户登录机器、用户使用机器、用户连接 USB 设备、机器病毒报警和机器声明 IP 地址。图 2 中的边代表发生过某安全事件；点代表机器、用户、USB 设备、病毒、IP 地址。点的大小代表事件发生的次数，即某事件发生的次数越多，该点越大。

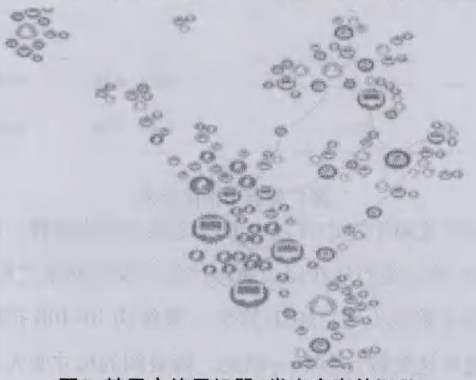


图2 某用户使用机器5类安全事件汇总

衡量两个时间周期图分割结果的相似度可以使用评价聚类结果的各种指标，本文中使用 AMI[18] 指标。AMI 计算公式如下：

$$AMI_{(X,Y)}=\frac{I_{(X,Y)}-E\{I_{(X,Y)}\}}{\max\{H_{(X)},H_{(Y)}\}-E\{I_{(X,Y)}\}} \quad (4)$$

其中，$E\{I_{(X,Y)}\}$ 表示划分结果 X,Y 的互信息的期望，可见 AMI 也是在 0 到 1 之间的值，划分 X,Y 越相似，AMI 越趋向于 1。

基于改进的 MCL 算法的异常检测模型步骤如下：

输入：时间周期总数 k，需要比较的前周期数 c，AMI 异常阈值 t。

1）从 1 到 k 的时间周期的事件构成图 $G_1,G_2,\cdots,G_k$。

2）用改进的 MCL 算法对 $G_1,G_2,\cdots,G_k$ 进行划分，结果标注为 $P_1,P_2,\cdots,P_k$。

3）从第 c+1 个划分起，计算它和前面 c 个划分的两两 AMI 指标。c 个 AMI 指标总和如果低于 AMI 异常阈值 t，则第 c+1 个划分对应的周期是异常的。

4）定位此异常周期中的异常点：从此时间周期对应的划分和从前 c 个划分中把点和点连接的边依次删除，每删除一个点，重新计算 AMI 指标总和。所有提升 AMI 总值的点 X 被认为是异常点。

输出：点 X。

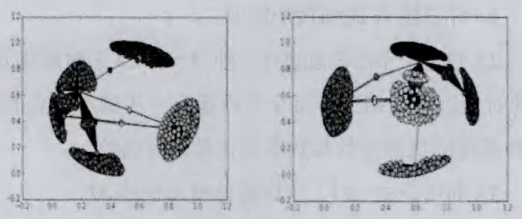
3 实验及分析

本文实验环境为：Intel Core i7-4790K CPU@ 4.00GHz；主存 32GB；系统平台为 Ubuntu14.04 LTS 英文版操作系统；软件实施平台为 Python 2.7.9(Anaconda 2.2.0)。每个实验结果均由运行 10 次，取均值得来。

3.1 MCL与METIS的比较

为测试不同的膨胀参数 r 对聚类效果的影响，现利用 MCL 算法对小数据集进行分割。MCL 算法接受边列表的输入形式，即每条边以“n_1n_2 Weight”的列表形式建立输入文档，其中 n_1 和 n_2 代表节点的名字，无需对原始节点重新编号。MCL 算法提供多个参数供选择，提供多种条件下的聚类。本实验中，选择在其他参数缺省的条件下，测试不同的膨胀参数 r 对聚类效果的影响。

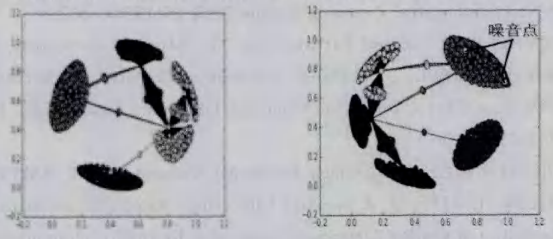
给定图节点总数 803 个，边总数 859 条。为防止聚类的数目过大或者过小，一般情况下膨胀参数 r 取值范围为 1.2~4。本文分别取 r=1.2 和 r=4 进行计算，分别得到 3 个聚类和 5 个聚类，如图 3 所示，计算时间分别为 0.04 秒和 0.03 秒。该测试说明 MCL 算法无需事先给定聚类数目，具备自动识别图内部结构的能力。



a）r_1=1.2时得到3个聚类　　b）r_2=4时得到5个聚类

图3 用MCL算法聚类效果图

为比较 MCL 算法与多层图分割算法（METIS），现利用 METIS 算法对相同的图进行测试，分别形成 3 个聚类和 5 个聚类。METIS 输入文档第一行为图的节点总数和边的总数，第二行及以后以邻接列表形式输入。在使用之前需将原始节点重新编号，即每个邻接列表以 "n_1 n_2 n_3 n_4 n_5" 建立输入文档，其中 $n_1,n_2,\cdots,n_5$ 代表节点编号。METIS 算法提供多个参数供选择，提供多种条件下的聚类。本实验中，选择在其他参数缺省的条件下，测试聚类数目为 3、5 的两种情况，计算时间均为 0.005 秒，结果如图 4 所示。



a）分割成3个聚类　　b）分割成5个聚类

图4 利用METIS算法的聚类效果图

比较图 3 和图 4，可知虽然都是得到了 3 个聚类和 5 个聚类，但是 MCL 和 METIS 这两种算法所得的结果截然不同，这主要是由于这两种算法原理不同。MCL 算法通过对图中节点之间的概率转移矩阵的简单几何运算来模拟随机流，这种做法使得划分的类不易出现噪声点，且每个类之间紧密度更高。METIS 算法采用的是多层分割结构，经粗化阶段和还原阶段后，每层之间的依赖性较高，较易得到噪声点，如图 4b）所示。

此外，METIS 聚类成权重均分的类，不适合长尾分布的数据。从直观上看，当分割成 3 个聚类时，MCL 得到的结果显然比 METIS 好，因为 MCL 相同类的内部连接边数更多，联系更加紧密；当分割成 5 个聚类时，METIS 出现噪声点。虽然 MCL 计算时间慢于 METIS，但是可通过调整参数提高计算速度，且随着节点数目的增加，MCL 计算时间增加也不快。

3.2 改进的MCL算法对大数据进行分割

Stanford 大型网络数据集[19]是关于图形的一个比较全面的数据集，数据收集时间也比较新，所以本文采用该数据集。表 1 列出了 Stanford 数据集中 3 个子数据集的大致情况。数据集 A 代表 com_Amazon 数据集，样本数最少，由爬虫 Amazon 网站获得，表示用户与商品之间的购买情况。如果两位用户买过相同的商品，则这两位用户是连接关系。数据集 B 代表 com_DBLP 数据集，来自计算机类英文文献的集成数据库 DBLP。如果两个作者一起发表至少一篇论文，则这两个作者是连接关系。数据集 C 代表 com_YouTube 数据集，来自 YouTube。如果两位用户是朋友关系，则这两位用户为连接关系。表 2 中 redA、redB 以及 redC 分别表示对数据集 A、B、C 去除度为 1 的点以及对应的边之后图的信息情况。对比表 1 和表 2 可知，数据集越大，Power-law 定理表现更明显，即本文的方法更有效。

表1 实验数据集信息

数据集	A	B	C
点总数	334,863	317,080	1,134,890
边总数	925,872	1,049,866	2,987,624

表2 去除度为1的点实验数据集信息

数据集	red A	red B	red C
点总数	309,154	273,899	532,351
边总数	900,163	1,006,685	2, 385,085

MCL、METIS 和改进的 MCL 三种图分割算法在 redA、redB、redC3 个数据集上的运行时间对比结果如图 5 所示、分割质量对比结果如图 6 所示。从图 5 可知，虽然 METIS 运行时间快于改进的 MCL，但改进的 MCL 对百万数量级的数据计算时间仍低于 100 秒，且随着图的点数和边数的增加，运行时间增加并不快。从图 6 可知，改进的 MCL 切割的边数远远低于 METIS，即分割质量优于 METIS。综合图 5、图 6 可以看出，改进的 MCL 算法要优于 METIS。

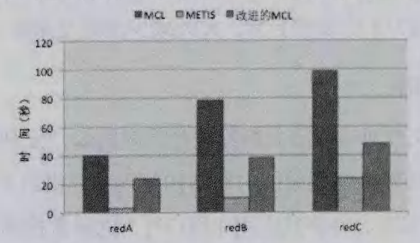


图5 三种分割算法运行时间对比结果

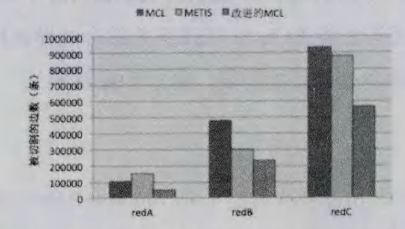


图6 三种图分割算法分割质量对比结果

3.3 异常检测实验

本实验数据源为某机关两个月内的网络数据，包括登录、USB 使用、上网、机器之间网络通信、使用 USB 设备、机器分配 IP 地址。数据量平均一天 4,300,000 条。对上网行为的目的节点，也就是访问的网站，根据注册信息做了简单的归类，即把网站分为 1 天前、1 周前、1 月前、1 年前注册这 4 个节点。这样避免了每个不同的网站就是一个节点，图中各种网站节点极度不平衡的情况出现。

1）按节点类别检测数量如下：

（1）多台机器违规使用同一个 USB 盘 31 次。

（2）局域网网络拓扑变化 19 次。

（3）用户违规访问其他部分文件服务器 7 次。

（4）用户登录他人机器 3 次。

2）主要误报来源分析如下：

（1）某用户更换部门，导致对（业务）服务器访问模式发生变化。

（2）路由配置错误导致各服务器连接断线，网络拓扑发生变化。

（3）多台机器上新安装的一个内部通信软件缺省使用 P2P 通信模式。

3）用户反馈认为改进的 MCL 算法的优点在于：

（1）异常结果容易通过可视化理解。

（2）子图划分结果很大程度与部门划分重合，而且子图往往小到可以在屏幕上全部显示。

（3）运维人员可以单个节点为单位处理异常，也可以一类节点为单位处理异常，大大减少运维复杂度。

4 结束语

本文针对当前图分割算法计算时间长、分割质量不高以及无法智能识别图的内部结构等现状，提出了一种改进的 MCL 算法，并将其应用于用户行为的异常检测。实验结果表明，该方法较 METIS 能获得质量较高的分割结果。下一步可考虑构建将 MCL 和 METIS 相结合的异常检测模型，以适应数据急速增长的需要。（责编 程斌）

参考文献：

[1]CHARU C. AGGARWAL. Outlier Analysis[M].Berlin Heidelberg : Springer,2013.

[2]CHANDOLA V, BANERJEE A, KUMER V. Anomaly Detection for Discrete Sequences: A Survey [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 24(11):823-839.

[3]NOBLE C C, COOK D J. Graph-based Anomaly Detection[C]// ACM, KDD '03.9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. August 24-27,2003. Washington,DC,USA. New York:ACM, 2003:631-636.

[4]TONG Hanghang, LIN Chingyung . Non-Negative Residual Matrix Factorization with Application to Graph Anomaly Detection[C]// SIAM, Eleventh SIAM International Conference on Data Mining, SDM 2011, April 28-30, 2011, Mesa, Arizona, USA. Philadelphia, PA : SIAM, 2011:143-153.

[5]LEE D D. Algorithms for Non-negative Matrix Factorization[J]. Advances in Neural Information Processing Systems, 2015, 13(6):556--562.

[6]YAO S Z. Spectral Partitioning: The More Eigenvectors, The Better[C]// IEEE, 32nd Design Automation Conference : proceedings 1995, June 12-16, 1995, San Francisco, USA. New Jersey:IEEE, 1995, 90(3):195-200.

[7]FAN R K C. Spectral Graph Theory[M].Washington, DC : AMS,1997.

[8]LIN H, ZHU Q. A Spectral Clustering-Based Dataset Structure Analysis and Outlier Detection Progress[J]. Journal of Computational Information Systems, 2012, 8(1):115-124.

[9]KARYPIS G, KUMAR V. Parallel Multilevel k-way Partitioning Scheme for Irregular Graphs [J].Siam Review, 1999, 41(2):278-300.

[10]HENDRICKSON B, LELAND R. A Multilevel Algorithm for Partitioning Graphs[C]// Supercomputing. Proceedings of the IEEE/ACM SC95 Conference. December 3-8,1995,San Diego,California, USA. New York:Association for Computing Machinery, 1995:28.

[11]KARYPIS G, KUMAR V. A Fast And High Quality Multilevel Scheme For Partitioning Irregular Graphs[J]. SIAM Journal on Scientific Computing, 2006, 20(1):359--392.

[12]DONGEN S M V. Graph Clustering by Flow Simulation [D]. Netherlands: Utrecht University, 2001.

[13]BROHEE S, HELDEN J V. Evaluation of Clustering Algorithms for Protein-protein Interaction Networks[J]. BMC Bioinformatics, 2006, 7(1602):2791-2797.

[14] 尚进，谢军，蒋东毅，等．现代网络安全架构异常行为分析模型研究 [J]. 信息网络安全，2015，(9)：15-19.

[15] 牛秦州，陈艳．基于 MCL 与 KNN 的混合聚类算法 [J]. 桂林理工大学学报，2015，35 (1)：181-186.

[16]MICHAEL MITZENMACHER. A Brief History of Generative Models for Power Law and Lognormal Distributions[J].Internet Mathematics. 2004,1(2):226-251.

[17]GOLDSTEIN M L, MORRIS S A, YEN G G. Problems with Fitting to the Power-Law Distribution[J]. Physics of Condensed Matter, 2004, 41(2):255-258.

[18]VINH N X, EPPS J, BAILEY J. Information Theoretic Measures for Clusterings Comparison[J]. Journal of Machine Learning Research, 2010, 11(1):2837-2854.

[19]JURE L, ANDREJ K. Large Network Dataset Collection [EB/OL]. http://snap.stanford.edu/data, 2014-2-26.