# 基于概率图的银行电信诈骗检测方法

刘 枭 王晓国

(同济大学电子与信息工程学院计算机科学与技术系 上海 201800)

**摘 要** 近几年,经由电信网络实施的诈骗频发,给银行用户带来了巨大的经济损失。现有的银行欺诈检测方法通常先提取账户交易的RFM(Recency,Frequency,Monetary Value)特征,然后采用有监督的方法训练分类器来识别诈骗交易。但是,这类方法没有考虑交易网络的结构特征。电信诈骗具有明显的集团特性,在交易网络中会呈现出特定的结构特征,使用交易网络的结构特征有助于识别电信诈骗。针对电信诈骗的集团特性,设计相应的马尔可夫网络用于识别电信诈骗中的欺诈账户。给出了该马尔可夫网络的线性迭代优化式,并证明了其理论收敛条件。最后在模拟数据和真实数据上测试了所提方法的性能,并将其与CIA和SybilRank进行比较。实验结果表明,所提方法具有更低的假阳性和更好的抗噪性。在真实数据上,将基于账户交易特征的方法与所提方法结合,可以取得比单独使用两种方法更好的识别性能。

**关键词** 欺诈检测,半监督学习,数据挖掘,电信诈骗,马尔可夫网络

**中图法分类号** TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.07.020

## Probabilistic Graphical Model Based Approach for Bank Telecommunication Fraud Detection

LIU Xiao WANG Xiao-guo

(Department of Computer Science and Technology,College of Electronics and Information Engineering,
Tongji University,Shanghai 201800,China)

**Abstract** Over the past few years,telecommunication fraud has caused enormous economic losses for bank users. Existing detection methods firstly extract statistical features,such as RFM (Recency,Frequency,Monetary Value) of user transactions,and then use supervised learning algorithms to detect fraud transactions or fraud accounts through training classifiers. However,the RFM features don't make use of the network structure of the transaction network. This paper designed a pairwise markov random field to capture the characteristics of the network structure in telecommunication fraud. Then,it exploited a linear loopy belief propagation algorithm to estimate the posterior probability distribution and predict the label of an account. Finally,it compared the proposed method with CIA and SybilRank on both synthetic dataset and real-world dataset. The results show that the proposed method outperforms other methods and can improve the F1-score of the RFM features based method.

**Keywords** Fraud detection,Semi-supervised learning,Data mining,Telecommunication fraud,Markov random field

## 1 引言

近年来,随着我国互联网金融的快速发展,电信诈骗也持续高发。相对于传统诈骗,电信诈骗更具智能性和隐蔽性,且具备有明显的集团特征。

从银行提供的数据来看,可将涉及欺诈的交易分为诈骗交易和洗钱交易。诈骗交易是指欺诈者通过各种手段从正常账户骗取金钱的交易;而洗钱交易则是指欺诈者将骗取的金钱进行分散和转移的交易。图1给出了简化后的欺诈交易结构(其中,圆圈代表本行账户,正方形代表非本行账户,白色代表正常账户,灰色代表欺诈账户),图中$a$到$e$是诈骗交易,而$e$到$i$、$g$到$f$等则是洗钱交易。如果能准确识别出由欺诈者控制的欺诈账户,那么银行便能制定相应的策略来应对诈骗交易和洗钱交易。

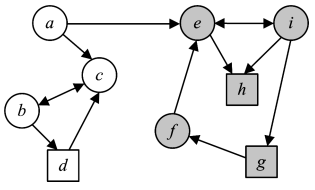



图1 电信诈骗的欺诈交易结构

Fig. 1 Fraud transaction structure of telecommunication fraud

在银行交易网络中,正常账户通常只与正常账户进行交易,欺诈账户只与欺诈账户进行交易。欺诈账户和正常账户之间只有少量的因为诈骗交易而产生的连接。假设图1是整

到稿日期:2018-03-01 返修日期:2018-05-08

刘 枭(1989-),男,博士生,主要研究方向为数据挖掘、欺诈检测,E-mail:nanaya100@gmail.com;王晓国(1966-),男,博士,教授,主要研究方向为数据挖掘、企业信息化,E-mail:xiaoguowang@tongji.edu.cn(通信作者)。

个交易网络，并且已知该网络中存在欺诈账户。如果确认 $b$ 是正常账户，那么可以推测 $c$ 也是正常账户，且 $a$ 和 $d$ 很有可能也是正常账户。而 $e$，$f$，$g$，$h$ 和 $i$ 则可能是欺诈账户。类似地，如果已知 $i$ 是欺诈账户，那么可以推测 $e$ 也是欺诈账户，且 $g$，$h$ 和 $f$ 很有可能是欺诈账户。而 $a$，$b$，$c$ 和 $d$ 则可能是正常账户。

形式化地，可以将上述由已知部分账户的标签数据来推测其余账户的标签的问题定义为网络中的标签传递问题。

**定义 1**（银行网络中的标签传递问题）　定义银行交易网络为有向图 $G(V,E,l)$，其中节点 $v\in V$ 表示实际数据中的账户，边 $(u,v)\in E$ 表示账户 $u$ 到账户 $v$ 的一笔转账交易，$l(v)\in\{1,-1\}$ 表示账户 $v$ 的真实标签，1 表示账户 $v$ 为欺诈账户，$-1$ 表示账户 $v$ 为正常账户。假设 $S\subset V$ 表示一部分已知标签的账户节点集合，$y(v)\in\{1,0,-1\}$ 和 $\hat{y}(v)\in\{1,-1\}$ 分别表示节点 $v$ 的现有的标签信息和预测标签，其中 0 表示节点标签未知，银行网络中的标签传递可以定义为：

$$\min_{\hat{y}}\sum_{v\in V}\frac{|l(v)-\hat{y}(v)|}{|V|} \tag{1}$$

其中，集合 $S$ 中的帐户的已知标签可能存在噪声，即 $\sum_{v\in S}|l(v)-y(v)|\geqslant 0$。

针对上述问题的解决方法主要有基于随机游走的方法和基于概率图的方法两类。基于随机游走的方法是一种通用的解决方法，可以解决不同应用中的欺诈检测问题。其基本思想是选择网络中已知标签的正常节点作为种子节点，从种子节点出发生成随机游走序列。由于正常节点和欺诈节点之间的联系通常比较少，这些随机游走能访问到的节点可以被认为是正常节点，而不能访问到的节点则被认为是欺诈节点。

基于概率图的方法则是针对不同的应用，选择或设计合适的图模型，并设计其图结构以及相应的概率分布来描述问题的性质。

本文将从以下 3 个方面展开研究：

1）针对电信诈骗数据的特征，设计相应的马尔可夫网络图模型，并研究该模型在大型网络上的优化问题；

2）基于模拟数据，研究不同的噪声比例和参数设置对所提方法的识别性能的影响，并将其成效与现有的基于随机游走的方法进行比较；

3）基于真实数据，将基于账户交易特征的方法与本文提出的方法相结合，以研究其对提升欺诈的识别性能的成效。

## 2　相关工作

Sybilguard[1] 最早将随机游走应用于欺诈检测中，并给出了随机路径的概念。随机路径是一种特殊的随机游走，在一条随机路径中任意节点的先继节点决定了后继节点。Sybilguard 在预处理阶段根据每个节点邻居的全排列生成上述路径表。然后以已知的可信节点作为起点，生成 $m$ 条随机路径，这些路径上的点被称为认证点。对于需要检测的点，同样生成 $m$ 条随机路径，如果随机路径经过认证点则称该路径被认证；如果 $m$ 条路径中多数为认证路径，则认为该节点是正常节点，否则为欺诈节点。Sybillimit[2] 在 Sybilguard 的基础上通过引入相交和平衡两个条件来限制生成的随机路径，从而降低假阴性。Sybilinfer[3] 将随机游走的思想与贝叶斯推理相结合，在实验中表现出了比 Sybillimit 和 Sybilguard 更好的识别性能。但是，Sybilinfer 的算法复杂度较高，理论上为 $O(n^2\log n)$，很难应用到实际应用中。

GateKeeper[4] 提出了一种票分配系统，在 Sybilguard 的基础上结合广度优先搜索和随机游走算法来保证搜索的有效性。但是该算法与上文提到的算法不同，其假设网络的生成是随机扩展的，而该假设在真实网络中无法成立，在实际应用中会出现非常高的假阳性和假阴性[5]。Sybildefender[5] 引入社区发现到欺诈节点的检测中。

不同于之前的方法简单地把整个网络分为可信区域和欺诈区域，Sybilshield[6] 假设网络可以分为几个大小不同的社区，各个社区间存在少量连接的边。因为正常节点和欺诈节点间的连接非常少，连接各个社区间的边中正常账户与正常账户间的边要远多于正常账户和欺诈账户间的边。Sybilshield 在流程上基本与 Sybilguard 一样，但是在节点第一次被判断为欺诈节点时，Sybilshield 就进行了第二步验证。在第二步验证中，Sybilshield 提出了代理节点的概念。代理节点是从可信区域以外的社区中选择的节点，如果在第一步中被判为欺诈节点的随机路径经过这些代理节点，那么就认为该节点是正常节点。实验中，Sybilshield 可以大幅降低 Sybilguard 的假阳性。

与上述方法不同，Sybilrank[7] 针对每个节点，生成一个可疑分值而不是直接进行二分类。Sybilrank 的思想与半监督的垃圾网页检测方法 Trustrank[8] 相似。Sybilrank 与 TrustRank 主要有两个不同点：1）Sybilrank 仅迭代 $O(\log n)$ 次，而 TrustRank 迭代到收敛；2）Sybilrank 最终根据节点的度数对可信值进行调整。对第 1 个不同点进行改进：因为网络中的可信区域和欺诈区域分离，且可信区域的子网络是 fast mixing 的假设，所以可以使经由攻击边传递到欺诈区域的可信值尽可能少。因为 PageRank 偏向于给度数较高的节点较高的分值，所以对于第 2 个不同点，可以减少高度数节点的假阴性率和低度数节点的假阳性率。

Sybilrank 和 Sybilshield 均假设了多社区的网络结构，并提出首先在网络中找出所有的社区，然后在每个社区中选择一部分节点进行人工确认后作为种子节点。

CIA（Criminal account Inference Algorithm）[9] 与上述从正常节点出发的算法不同，其从已知的欺诈节点出发来找出所有的欺诈节点。CIA 与 Sybilrank 借鉴 TrustRank 的思想类似，主要借鉴了 BadRank[10] 的思想。

所有基于随机游走的方法都假设网络是 fast mixing 或者正常节点组成的可信区域是 fast mixing。在一个网络中的马尔可夫的混合时间被定义为：

$$T(\varepsilon)=\max_{i}\min\{t\,|\,|\pi-\pi^{(0)}P^t|_1<\varepsilon\} \tag{2}$$

其中，$\pi$ 表示最终的平稳概率分布，$\pi^{(0)}$ 表示初始的概率分布，$P$ 是网络 $G$ 的概率转移矩阵，$|\cdot|$ 表示概率分布间的总变差（Total Variation Distance）。

与文献[1,5,11]相同，当网络满足 fast mixing 的假设时，认为 $T(\varepsilon)=O(\log n)$。文献[12]中给出了混合时间 $T(\varepsilon)$ 的上限和下限：

$$\frac{\mu}{2(1-\mu)}\log(\frac{1}{2\varepsilon})\leqslant T(\varepsilon)\leqslant\frac{\log(\mu)+\log(\frac{1}{\varepsilon})}{1-\mu} \tag{3}$$

其中，$\mu$ 是网络的概率转移矩阵 $P$ 的第二大特征值。文献[13]给出了 $\mu$ 和网络的电导(conductance)$\Phi(G)$的关系：

$$1-2\Phi(G)\leqslant\mu\leqslant1-\frac{\Phi(G)^2}{2} \tag{4}$$

网络的电导衡量了网络的稀疏程度。根据式(4)，网络的连通性越强，电导值就越大，网络转移矩阵的 $\mu$ 就越小，从而由式(3)可以得出网络的混合时间就越小。因此，fast mixing 的假设相当于假设了网络的强连通性或者网络中可信区域的强连通性。而由文献[1]可发现：该假设在很多现实的网络中都不成立。本文的交易网络中节点数为 10000 的可信区域的 $\mu$ 值平均为 0.9972，这同样不符合 fast mixing 的假设。

应用基于随机游走的方法的另一个难点在于如何选择可信种子。现实世界的网络通常呈多社区模型，如果选择的可信种子没有覆盖所有的正常社区，则算法的假阳性将会非常高。Sybilrank 中使用 Louvain 算法[14]将整个网络分割为多个社区，再从每个社区中选择一部分账户进行人工认证，并将其中的正常节点作为可信种子。本文对由每个月的交易数据形成的网络使用 Louvain 算法后，得到的社区数在 28～32 万之间。即使每个社区仅选取一个账户进行人工确认，也需要人工确认 30 万左右的账户。在实际应用中，特别是项目初期标记数据积累不足时，几乎无法实现该工作。从欺诈种子出发的 CIA 算法虽然不存在上述问题，但是 CIA 只使用欺诈样本而无法利用正常样本；当欺诈样本中存在噪声时，CIA 无法利用正常样本来减少噪声的影响。

与基于随机游走的方法针对通用的欺诈问题不同，基于概率图的方法是针对不同的应用选用合适的图模型，并设计图结构以及相应的概率分布来描述问题的性质。因此，基于概率图模型的方法没有统一的假设，而是针对不同的应用提出与数据相符的假设，并将其反映到概率图结构的设计中。如 Netprobe[15]中发现在 eBay 的交易网络中欺诈交易呈现如下结构：欺诈者通常会控制欺诈账户和中继账户两类账户；中继账户通过正常账户进行交易来提升自己的可信度，再通过与欺诈账户进行交易将可信度传递给欺诈账户。Netprobe 基于上述发现采用马尔可夫网络进行建模，并设计了相应的概率转移函数。Shebuti 等[16]发现欺诈用户通常给欺诈 APP 较高的评分，给正常 APP 较低的评分。而正常用户刚好相反。Shebuti 等通过分析 APP 的评价文本找出可能的虚假评价，再构建用户-评论的概率图模型将欺诈分值在网络中传递。该方法可以有效地发现没有评论的虚假评价和欺诈账户。

## 3 马尔可夫网络的构建

本节介绍根据电信诈骗特点构建的马尔可夫网络。

根据前文对银行交易网络的定义，将马尔可夫网络定义为 $\overline{G}(\overline{V},\overline{E})$，其中 $\overline{V}=V$，$\overline{E}=\{(u,v)\mid(u,v)\in E\vee(v,u)\in E\}$。节点 $v$ 对应的随机变量 $x_v\in\{1,-1\}$。根据实际的数据情况，我们可以做出以下基本假设：

1)欺诈账户只会给欺诈账户转账，不会给正常账户转账；

2)除了诈骗交易，正常账户不会给欺诈账户转账。

使用 $\Phi(x_v)$表示节点 $v$ 的先验概率分布，即在已知标签的情况下，账户 $v$ 是欺诈账户的概率分布。定义 $N(v)$，$P(v)$和 $S(v)$分别表示账户 $v$ 的双向边邻居节点集合、单向边先继节点集合和单向边后继节点集合。$N(v)$表示既向账户 $v$ 转过账，也接受过账户 $v$ 转账的账户的集合；$P(v)$表示向账户 $v$ 转过账，但没有接受过账户 $v$ 转账的账户的集合；$N(v)$表示接受过账户 $v$ 转账，但没有向账户 $v$ 转过账的账户的集合。因为正常账户和欺诈账户之间不会有交易，所以账户 $v$ 的标签趋于与双向边邻居集合 $N(v)$中的已知标签的账户一样。由此定义：

$$\Pr(x_v=1)=\frac{1}{1+\exp(-\sum\limits_{u\in N(v)}(w(v,u)+w(u,v))x_u)} \tag{5}$$

如果已知账户 $v$ 的单向边先继邻居 $u$ 是欺诈账户，那么账户 $v$ 也趋向于是欺诈账户，因为欺诈账户只会给欺诈账户转账。但是如果已知账户 $u$ 是正常账户，则不会对账户 $v$ 的标签产生影响，因为账户 $v$ 可能是由于受到诈骗而向账户 $u$ 转账。由此定义：

$$\Pr(x_v=1)=\frac{1}{1+\exp(-\frac{1}{2}\sum\limits_{u\in P(v)}w(u,v)(1+x_u)x_u)} \tag{6}$$

如果已知账户 $v$ 的单向边后继邻居 $u$ 是正常账户，那么账户 $v$ 也趋向于是正常账户，因为欺诈账户不会给正常账户转账。但是如果已知账户 $u$ 是欺诈账户，则不会对账户 $v$ 的标签产生影响，因为账户 $v$ 可能是由于受到诈骗而向账户 $u$ 转账。由此定义：

$$\Pr(x_v=1)=\frac{1}{1+\exp(-\frac{1}{2}\sum\limits_{u\in S(v)}w(v,u)(1-x_u)x_u)} \tag{7}$$

根据上述定义，节点 $v$ 的先验概率分布 $\Phi(x_v)$可以设置为：

$$\Phi(x_v)=\begin{cases}\dfrac{1}{1+\exp(-I_N(v)-I_P(v)-I_S(v)-\alpha x_v)}, & \text{if } x_v=1\\ 1-\dfrac{1}{1+\exp(-I_N(v)-I_P(v)-I_S(v)-\alpha x_v)}, & \text{if } x_v=-1\end{cases} \tag{8}$$

其中，$I_N(v)=\sum\limits_{u\in N(v)}(w(v,u)+w(u,v))x_u$ 表示所有双向边邻居的影响的总和，$I_P(v)=\frac{1}{2}\sum\limits_{u\in P(v)}w(u,v)(1+x_u)x_u$ 表示所有单向边先继邻居的影响的总和，$I_S(v)=\frac{1}{2}\sum\limits_{u\in S(v)}w(v,u)(1-x_u)x_u$ 表示所有单向边后继邻居的影响的总和。$\alpha x_v$ 表示本身的标签信息，其中 $\alpha$ 是常数，本文中设置 $\alpha$ 为 10。

与节点先验概率类似，可以定义网络 $G$ 中双向边的势函数为：

$$\Psi(x_v,x_u)=\begin{cases}w, & \text{if } x_vx_u=1\\ 1-w, & \text{if } x_vx_u=-1\end{cases} \tag{9}$$

定义单向边的势函数为：

$$\Psi(x_v,x_u)=\begin{cases}w, & \text{if } x_v=-1\vee x_u=1\\ 1-w, & \text{otherwise}\end{cases} \tag{10}$$

其中，$w\in[0,1]$，本文中设置 $w>0.5$。上述的边势函数定义中，$w$ 可以理解为在网络 $G$ 中双向边连接的两个节点的标签相同的概率。

根据先验概率分布 $\Phi(x_v)$ 和边势函数 $\Psi(x_v,x_u)$ 的定义，可将马尔可夫网络的联合概率分布定义为：

$$\Pr(X=x)=\frac{1}{Z}\prod_{v\in\overline{V}}\Phi(x_v)\prod_{(u,v)\in\overline{E}}\Psi(x_u,x_v) \tag{11}$$

其中，$Z=\sum_{X}\prod_{v\in\overline{V}}\Phi(x_v)\prod_{(u,v)\in\overline{E}}\Psi(x_u,x_v)$。

## 4　循环置信传播

在概率图中进行统计推断(Inference) 通常是很困难的。在一般的概率图模型中，随着图中的顶点与边的数目的增加，统计推断的计算复杂度会以指数的形式增加。但我们可以使用循环置信传播算法(Loopy Belief Propagation)[17]来求解马尔可夫网络中各个节点的后验概率。循环置信传播算法通过在马尔可夫网络中的相邻节点间迭代传播消息(message)来使节点的后验逐步趋向稳定。具体地，在算法第 $t$ 次迭代时从节点 $v$ 向节点 $u$ 传播的消息 $m_{vu}^{(t)}(x_u)$ 可表示为：

$$m_{vu}^{(t)}(x_u)=\sum_{x_v}\Phi(x_v)\Psi(x_v,x_u)\prod_{k\in Neig(v)\backslash u}m_{kv}^{(t-1)}(x_v) \tag{12}$$

其中，$Neig(v)\backslash u$ 表示在网络 $\overline{G}$ 中去除了节点 $u$ 后的节点 $v$ 的所有邻居节点的集合。式(12)表示在循环置信传播算法第 $t$ 轮迭代中，节点 $v$ 接受除了节点 $u$ 以外所有邻居节点的第 $t-1$ 轮消息，并根据先验概率分布 $\Phi(x_v)$ 和节点 $u$ 与节点 $v$ 之间的边势函数 $\Psi(x_v,x_u)$ 计算新的由节点 $v$ 向节点 $u$ 传播的消息。当新消息与旧消息之间的变化非常小(例如变化的 L1 范数小于 $10^{-7}$)时，可以认为循环置信传播算法收敛。循环置信传播算法虽然在理论上无法保证收敛，但是在许多实际应用中通常可以很好地收敛。

循环置信传播算法收敛后，节点 $u$ 的后验概率 $\Pr(x_u)$ 可以通过式(13)计算：

$$\Pr^{(t)}(x_u)\propto\Phi(x_u)\prod_{k\in Neig(u)}m_{ku}^{(t)}(x_u) \tag{13}$$

为方便表示，令 $\theta_u=\Phi(x_u=1)$，$p_u^{(t)}=\Pr^{(t)}(x_u=1)$，$m_{vu}^{(t)}=m_{vu}^{(t)}(x_u=1)$，$\Phi(x_u=-1)=1-\theta_u$，$\Pr^{(t)}(x_u=-1)=1-p_u^{(t)}$，$m_{vu}^{(t)}(x_u=-1)=1-m_{vu}^{(t)}$。因为在每一轮计算消息后对消息进行归一化处理不会影响最终结果，所以假设 $m_{vu}^{(t)}(x_u=1)+m_{vu}^{(t)}(x_u=-1)=1$。由此，式(13)可以表示为：

$$p_u^{(t)}=\frac{\theta_u\prod_{k\in Neig(u)}m_{ku}^{(t)}}{\theta_u\prod_{k\in Neig(u)}m_{ku}^{(t)}+(1-\theta_u)\prod_{k\in Neig(u)}1-m_{ku}^{(t)}} \tag{14}$$

### 4.1　线性化

循环置信传播算法有两个缺点：1)没有收敛保证，在带环的网络中可能出现震荡不收敛的情况[17]；2)算法迭代时需要在边上保存消息数据，并且在更新消息时需要去除传出边节点的消息，这会在空间和时间上产生额外开销。

Wang 等[18]发现在计算第 $t$ 轮节点 $v$ 向节点 $u$ 传播的消息 $m_{vu}^{(t)}(x_u)$ 时，不去除节点 $u$ 向节点 $v$ 传播的消息，而直接使用节点 $v$ 所有邻居节点的第 $t-1$ 轮消息，马尔可夫网络中的后验概率分布最终仍然能收敛。因此本文采用相同的思想将式(12)修改为：

$$m_{vu}^{(t)}(x_u)=\sum_{x_v}\Phi(x_v)\Psi(x_v,x_u)\prod_{k\in Neig(v)}m_{kv}^{(t-1)}(x_v) \tag{15}$$

将式(13)代入式(15)，得到：

$$m_{vu}^{(t)}(x_u)\propto\sum_{x_v}\Psi(x_v,x_u)\Pr^{(t-1)}(x_v) \tag{16}$$

因为 $m_{vu}^{(t)}(x_u=1)+m_{vu}^{(t)}(x_u=-1)=1$，根据式(16)很容易验证：当且仅当 $\Pr^{(t-1)}(x_v)=0$ 且 $v\in P(u)$ 或者 $\Pr^{(t-1)}(x_v)=1$ 且 $v\in S(u)$ 时，$m_{vu}^{(t)}(x_u=1)=m_{vu}^{(t)}(x_u=-1)=0.5$。上述结果可以理解为正常节点不会对其后继节点传递有效信息，而欺诈节点不会对其先继节点传递有效信息，这与电信诈骗问题的特征相符。将式(16)进一步简化为：

$$m_{vu}^{(t)}=\begin{cases}0.5, & \text{if } p_v^{(t-1)}<0.5\wedge v\in P(u)\\ 0.5, & \text{if } p_v^{(t-1)}>0.5\wedge v\in S(u)\\ wp_v^{(t-1)}+(1-w)(1-p_v^{(t-1)}), & \text{otherwise}\end{cases} \tag{17}$$

置信传播算法根据式(14)和式(17)迭代更新消息，而不需要在每条边上维护消息数据，从而降低了空间复杂度和时间复杂度。在式(14)中对后验概率的更新是非线性的，Wang 等[18]和 Gatterbauer 等[19]提出将式(14)线性化，以保证算法的收敛性。下面推导出式(14)的线性迭代式。

**定义 2**(残差变量和残差向量)[19]　定义残差变量 $\hat{x}=x-0.5$，残差向量 $\hat{x}=[x_0-0.5,x_1-0.5,\cdots]$。

为方便描述，定义列向量 $p^{(t)}=[p_0^{(t)};p_1^{(t)};\ldots;p_{|V|-1}^{(t)}]$ 表示账户节点在 $t$ 轮迭代后的后验概率向量。按照定义 2，用 $\hat{p}^{(t)}$ 表示 $p^{(t)}$ 的残差向量。相似地，定义 $\theta^{(t)}$ 和 $\hat{\theta}^{(t)}$ 表示节点的先验概率的列向量和残差向量。定义 $A_b\in R^{|V|\times|V|}$ 和 $A_u\in R^{|V|\times|V|}$ 分别表示银行交易网络 $G$ 的双向边邻接矩阵和单向边邻接矩阵。$A_b$ 的第 $u$ 行表示 $u$ 的双向边邻居，而 $A_u$ 的第 $u$ 行表示 $u$ 的单向边后继邻居。令 $\hat{P}^{(t)}=[\hat{p}^{(t)},\hat{p}^{(t)},\cdots,\hat{p}^{(t)}]$ 表示将列向量 $\hat{p}^{(t)}$ 重复叠放 $|V|$ 次后得到的矩阵，$\hat{P}^{(t)}\in R^{|V|\times|V|}$。

**定理 1**　式(14)的近似线性迭代式为[18-20]：

$$\begin{aligned}&\hat{p}^{(t)}=\hat{\theta}+2\,\hat{w}M^{(t-1)}\hat{p}^{(t-1)}\\&M^{(t-1)}=(A_b+I(A_u\circ\hat{P}^{(t-1)^{\mathrm{T}}})+I(-A_u^{\mathrm{T}}\circ\hat{P}^{(t-1)^{\mathrm{T}}}))\end{aligned} \tag{18}$$

其中，$I(A)$ 是一个指示函数，它将矩阵中的非负数设置为 0，其余数设置为 1。$\circ$ 算符表示两个矩阵间对应元素相乘。

证明：参考文献[18-20]。

算法 1 给出了整个算法的伪代码。算法首先根据式(8)初始化先验概率分布的残差；再根据迭代式(18)更新后验概率的残差，当后验概率的残差的相对误差小于阈值或者达到设置的算法最大迭代次数时，算法停止；最后输出原始的后验概率。

**算法 1**　LinearLBS

输入：交易网络 G(V,E,w)，马尔科夫网络 $\overline{G}(\overline{V},\overline{E})$，w，ε，T，已知标签数据

1. 根据已知标签和式(8)初始化 $\hat{\theta}$；
2. 初始化 $\hat{p}^{(0)}$ 为 $\hat{\theta}$；
3. t=1；
4. while $\|\hat{p}^{(t)}-\hat{p}^{(t-1)}\|/\|\hat{p}^{(t)}\|>\varepsilon$ and t<T do
5. 　使用式(18)更新 $\hat{p}^{(t)}$；
6. 　t=t+1；
7. end while
8. return $\hat{p}^{(t)}+0.5$.

### 4.2 收敛性

本节从理论上来分析算法 1 的收敛条件。

**定义 3**(谱半径(Spectral radius)) 用 $\lambda_1,\lambda_2,\cdots,\lambda_n$ 表示矩阵 $A\in C^{n\times n}$ 的特征值。定义矩阵 $A$ 的谱半径为:

$$\rho(A)=\max\{|\lambda_1|,|\lambda_2|,\cdots,|\lambda_n|\} \tag{19}$$

**定理 2**(线性迭代式收敛的充分必要条件[21]) 给定线性迭代式 $y^{(t)}\leftarrow c+My^{(t-1)}$ 和任意初始值 $y^{(t)}$,其收敛的充分必要条件是矩阵 $M$ 的谱半径小于 1,即 $\rho(M)<1$。

**引理 1**(算法 1 收敛的充分必要条件) 式(18)收敛的充分必要条件是:

$$\overset{\wedge}{w}<\frac{1}{2\rho(M^{(t')})} \tag{20}$$

其中,$t'=\arg\max_t \rho(M^{(t)})$。

证明:由定理 2 可知,当且仅当 $\rho(2\overset{\wedge}{w}M^{(t)})<1$ 对所有 $t$ 都满足时,式(18)收敛。由于 $\rho(2\overset{\wedge}{w}M^{(t)})=2\overset{\wedge}{w}\rho(M^{(t)})$,因此式(18)收敛的充分必要条件是当 $\rho(M^{(t)})$ 取最大值 $\rho(M^{(t')})$ 时,满足 $2\overset{\wedge}{w}\rho(M^{(t')})<1$。而 $\rho(M^{(t')})>0$,两边除以 $2\rho(M^{(t')})$ 即可得证。

引理 1 给出了算法收敛的充分必要条件。Wang 等[18]在引理 1 的基础上进一步证明了一个较易计算的收敛的充分条件。给定矩阵 $A$,对于其任意的特征值和特征向量对$(\lambda,x)$,有 $Ax=\lambda x$,而 $\|\lambda x\|=\|\lambda\|\|x\|=\|Ax\|\leqslant\|A\|\|x\|$,所以 $\|\lambda\|\leqslant\|A\|$。这对于任意特征值和任意范数运算都成立,取 L1 范数可以得到 $\rho(A)\leqslant\|A\|_1$。

**引理 2**(算法 1 收敛的充分条件[18]) 式(18)收敛的充分条件为:

$$\overset{\wedge}{w}<\frac{1}{2\max_{u\in V}\deg(u)} \tag{21}$$

证明:$I(A_u\circ\overset{\wedge}{P}^{(t-1)^{\mathrm{T}}})$ 和 $A_u$ 都是 01 矩阵,在任意一轮迭代中,对于 $A_u$ 的任意元素,有 $I(A_u\circ\overset{\wedge}{P}^{(t-1)^{\mathrm{T}}})\leqslant A_u\circ I(\overset{\wedge}{P}^{(t-1)^{\mathrm{T}}})\leqslant A_u$;同理有:$I(-A_u^{\mathrm{T}}\circ\overset{\wedge}{P}^{(t-1)^{\mathrm{T}}})\leqslant A_u^{\mathrm{T}}$。所以在任意一轮迭代中有 $\|M^{(t)}\|_1\leqslant\|A_b+A_u+A_u^{\mathrm{T}}\|_1$。所以,$\overset{\wedge}{w}<\frac{1}{\|A_b+A_u+A_u^{\mathrm{T}}\|_1}\rightarrow\overset{\wedge}{w}<\frac{1}{2\rho(M^{(t')})}$。不难证明 $\|A_b+A_u+A_u^{\mathrm{T}}\|_1=\max_{u\in V}\deg(u)$。

根据引理 2,当设置边势函数中的 $w$ 为小于 0.5 加上网络中的最大度数的两倍的倒数时,算法 1 从理论上可以保证收敛。但在实际应用中,通常只要设置 $w>0.5$ 就可以获得比较稳定的结果。

## 5 实验分析

### 5.1 模拟数据分析

本节研究本文提出的马尔可夫网络模型(简称 MRF)在不同比例的噪声样本和不同算法参数下的性能,并将其与 Sybilrank 和 CIA 进行比较。为了方便调整各种参数,本节采用模拟数据进行实验。网络通过 Power-law Cluster[22]算法进行模拟,以保证模拟网络中节点的度数符合幂定律(Power-law)。模拟数据的基本信息如表 1 所列,模拟网络的节点度数分布如图 2 所示。

表 1 模拟数据的基本信息

Table 1 Basic information of synthetic data

| 属性 | 数值 |
|---|---|
| 正常样本数 | 10000 |
| 欺诈样本数 | 100 |
| 正常社区数 | 500 |
| 正常社区平均节点数 | 20 |
| 欺诈社区数 | 10 |
| 欺诈社区平均节点数 | 10 |
| 节点平均度数 | 2.6 |
| 正常节点间的边数 | 12856 |
| 欺诈节点间的边数 | 307 |
| 正常节点与欺诈节点间的边数 | 20 |

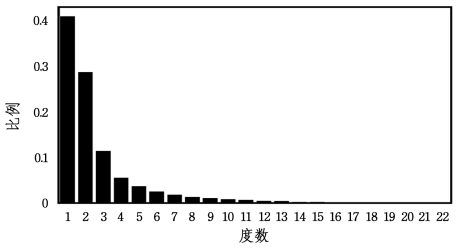


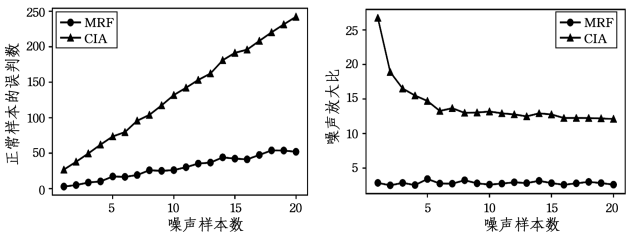

图 2 模拟数据的节点度数分布

Fig. 2 Node degree distribution of synthetic data

#### 5.1.1 噪声标记的影响

实际中得到的标记数据通常存在噪声,本节分析噪声数据对算法性能的影响。

图 3 显示了噪声样本数对算法的正常样本误判数和噪声放大比的影响,噪声放大比=正常样本的误判数/噪声样本数。实验中,将欺诈样本的标记数和正常样本的标记数分别固定为 20 和 2000。从图 3(a)可以看出,随着噪声样本数的增加,CIA 和 MRF 对正常样本的误判数都线性增加,但是 MRF 的增速明显低于 CIA。这主要由于 MRF 可以利用标记正确的正常样本来降低噪声对样本的影响,而 CIA 无法利用标记正确的正常样本。如图 3(b)所示,CIA 的噪声放大比在 12 左右,而 MRF 的噪声放大比稳定在 3 以下,这说明 MRF 有效利用了标记正确的正常样本来降低噪声样本的影响。



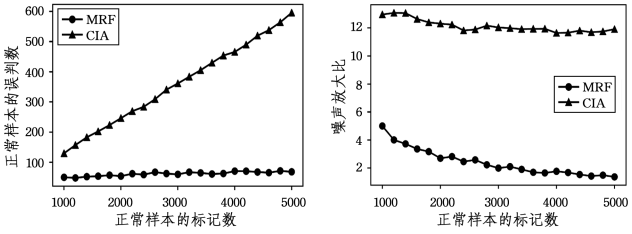

(a)噪声样本数对正常样本误判数的影响 (b)噪声样本数对噪声放大比的影响

图 3 噪声样本数对算法性能的影响

Fig. 3 Effect of number of noise labels on performance of algorithms

图 4 显示了当噪声比例为 1% 时,正常样本的标记数对算法的正常样本误判数和噪声放大比的影响。实验中,将欺诈样本的标记数固定为 20。从图 4(a)可以看出,随着正常样本的标记数的增加,噪声样本数量呈线性增加,导致 CIA 对正常样本的误判数也呈线性增加。而 MRF 对正常样本的误判数则稳定在 60 左右,这说明 MRF 主要受噪声样本的比例影响,噪声样本的绝对数量对其影响较小。图 4(b)中显示了 CIA 的噪声放大比稳定在 12 左右,而 MRF 的噪声放大比随

着正常样本的标记数的增多而逐步降低。在实际应用中,随着标记样本的积累,MRF 的性能将逐步提升。

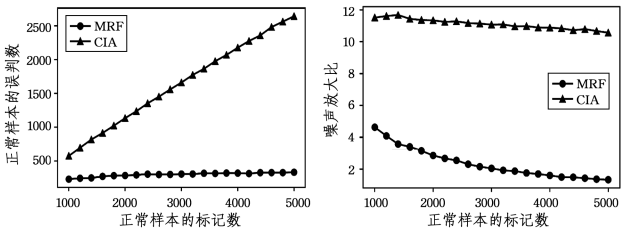



(a)正常样本的标记数对正常样本误判数的影响　　(b)正常样本的标记数对噪声放大比的影响

图 4　当噪声比例固定为 1%时正常样本的标记数对算法性能的影响

Fig. 4　Effect of number of normal labels on performance of algorithms with ratio of noise is fixed to 1%

图 5 显示了当噪声比例为 5%时,正常样本的标记数对算法的正常样本误判数和噪声放大比的影响。从图 5(a)可以看出,CIA 由于噪声样本的绝对数量增多,对正常样本的误判数大幅增加。而 MRF 由于噪声样本的比例增加,对正常样本的误判数也大幅增加。对比图 4(b)和图 5(b)可以发现,噪声样本的绝对数量和噪声样本的比例对 CIA 和 MRF 的噪声放大比均没有影响。

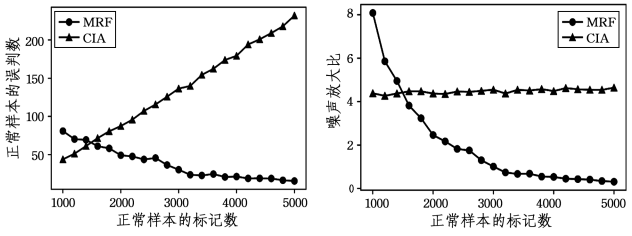



(a)正常样本的标记数对正常样本的误判数的影响　　(b)正常样本的标记数对噪声放大比的影响

图 5　当噪声比例固定为 5%时正常样本的标记数对算法性能的影响

Fig. 5　Effect of number of normal labels on performance of algorithms with ratio of noise is fixed to 5%

为了测试网络结构对算法性能的影响,我们调整了网络节点的度数,使平均度数增加到 6.7,并保持其他设置不变。图 6 显示了在该结构下,当噪声比例为 1%时,正常样本的标记数对算法的正常样本误判数和噪声放大比的影响。




(a)正常样本的标记数对正常样本的误判数的影响　　(b)正常样本的标记数对噪声放大比的影响

图 6　当网络节点的度数为 6.7 且噪声比例固定为 1%时正常样本的标记数对算法性能的影响

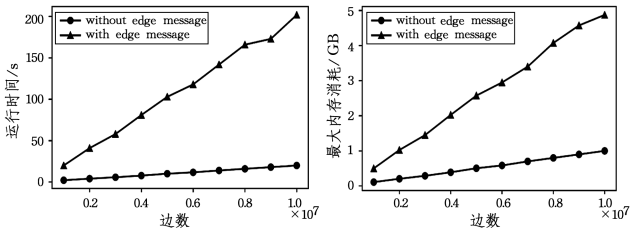
Fig. 6　Effect of number of normal labels on performance of algorithms when average degree in network is 6.7 and percentage of noise is 1%

对比图 4(a)和图 6(a)可以发现,图 6(a)中 CIA 对正常样本的误判出现了下降,特别是正常样本的标记数较少时,误判数甚至比 MRF 更少。但 CIA 对正常样本的误判数仍然随着噪声样本的增多而线性增加。MRF 对正常样本的误判数随着噪声样本的增多呈现下降的趋势,当正常样本的标记数为 5000 时,图 6(a)中的误判数明显少于图 4(a)中的误判数。对比图 6(b)和图 4(b)可以发现,CIA 的噪声放大比从 12 左右下降到了 4 左右。MRF 在正常样本的标记数较少时,噪声放大比有所上升,但是当正常样本的标记数超过 2500 时,噪声放大比下降到 1 以下。出现上述现象的主要原因是:CIA 的主要思想是将风险值从种子节点传播出去,网络节点的平均度数越高,一个种子节点的风险值会传播给更多的节点,分值会相对分散,误判数相对就会减少;相反,MRF 的主要思想是利用节点标签间的趋同性,网络节点的平均度数越高,一个种子节点将影响越多的节点。当正常样本的标记的绝对数较少时,噪声样本与正常样本的交集较少,相对地,噪声样本会影响更多的正常节点。而当正常样本的标记的绝对数较多时,噪声样本和正常样本的交集增加,噪声样本的影响变小。

总体而言,CIA 的性能随着噪声样本的绝对数量的增加而线性下降,噪声放大比随着网络的平均节点度数的增加而下降。MRF 的性能随着噪声比的增加而下降。在噪声比固定时,噪声放大比随着正常样本的标记数量的增加而下降。随着标记数据的积累,MRF 的性能将逐步提升。当正常样本的标记数据充足时,噪声放大比随着网络的平均度数的增加而下降。

### 5.1.2　参数设置和算法复杂度

图 7 给出了算法运行时间和最大消耗内存随着网络大小的变化。从图 7(a)中可以发现,省去循环置信传播在边上保存消息后,算法的时间复杂度的常数项大幅下降。整体而言,算法与网络中的边数呈线性相关。从图 7(b)中可以发现,空间复杂度的常数项的值也因为不用维护每条边上的消息而降低。




(a)时间复杂度　　(b)空间复杂度

图 7　算法复杂度

Fig. 7　Complexity of algorithms

图 8 给出了不同的 $w$ 值对算法性能的影响。可以发现本文算法对 $w$ 的取值不敏感,实际应用中设置 $w>0.5$ 即可。

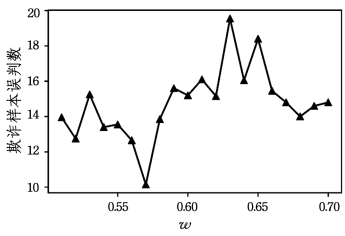



图 8　$w$ 对算法性能的影响

Fig. 8　Effect of $w$ on algorithm performance

### 5.2 银行数据上的实验

本文使用银行提供的 2016 年 1 月 1 日至 2017 年 7 月 1 日的电子银行的交易历史数据进行实验。去除如企业交易、内网交易等特殊交易后，数据基本情况如表 2 所列，其确认的欺诈账户仅包括本行账户。

表 2 银行提供的数据的基本情况

Table 2 Basic information of data provided by bank

| 交易总数 | 账号总数 | 本行账号数 | 确认的欺诈账户数 |
|---|---|---|---|
| 17441949 | 3046079 | 686807 | 195 |

由于银行提供的欺诈账户仅包含本行账户，本文仅使用本行账户来评估不同算法的性能。实验数据集选取 195 个被确认为欺诈的本行账户和 10000 个至少有 50 条交易记录的已确认的正常本行账户(来源于银行账户白名单和人工筛选数据)。实验中各算法的配置如下。

1)LightGBM：使用基于账户交易特征的分类算法，算法采用文献[23]中提出的特征。如果一个账户的 25%以上的交易被判断为欺诈交易，则认为该账户是欺诈账户。树数量设置为 200，树叶子节点数设置为 31，行采样系数设置为0.7，列采样系数设置为 0.7，shrinkage 设置为 0.05。使用 20%的数据进行训练，80%的数据进行测试。

2)MRF：以 20%的数据作为已知的标记数据。在连续 30 天的交易数据形成的网络上运行 MRF。

3)SybilRank：以 20%的正常样本数据作为已知的标记数据。在连续 30 天的交易数据形成的网络上运行 SybilRank。

4)CIA：以 20%的欺诈样本数据作为已知的标记数据。在连续 30 天的交易数据形成的网络上运行 CIA。

5)LightGBM＋MRF：以 LightGBM 识别出的欺诈账户以及 LightGBM 的训练数据作为已知的标记数据。在连续 30 天的交易数据形成的网络上运行 MRF。

6)LightGBM＋CIA：以 LightGBM 识别出的欺诈账户以及 LightGBM 的训练数据作为已知的标记数据。在连续 30 天的交易数据形成的网络上运行 CIA。

实验使用 Precision，Recall 和 F1-score 来评估算法的性能。实验重复运行 20 次，并取平均值作为最终结果。

表 3 列出了不同算法的实验结果。SybilRank 能检测到绝大部分账户，但是误报非常多。SybilRank 的结果验证了正常账户和欺诈账户基本分离的假设。

表 3 真实数据上的实验结果

Table 3 Experimental results on real-world dataset

| 算法 | *Precision* | *Recall* | *F1-score* |
|---|---|---|---|
| LightGBM | 0.748 | 0.374 | 0.499 |
| MRF | 0.979 | 0.479 | 0.643 |
| SybilRank | 0.0559 | 0.974 | 0.106 |
| CIA | 0.937 | 0.478 | 0.633 |
| LightGBM＋MRF | 0.685 | 0.679 | **0.682** |
| LightGBM＋CIA | 0.315 | 0.679 | 0.430 |

CIA 和 MRF 的召回率比模拟数据低，这是因为相对于模拟数据，实际数据中欺诈账户间的连通性更低，形成了更多更小的欺诈社区。使用 LightGBM 的结果和训练集的标记数据作为 MRF 的输入标签，提高了算法的整体的召回率，而精准率只出现了小幅度的下降。相对地，以 LightGBM 的结果和标记数据作为 CIA 的输入标签，虽然得到了同样的召回率提升，但是由于 CIA 的噪声放大比过大，导致精准率出现大幅下降。

**结束语** 在研究电信诈骗数据的基础上，本文设计了符合电信诈骗特征的马尔可夫网络，给出了相比传统循环置信传播更为高效的后验概率优化算法，使所提方法在大型网络上的优化更为高效，并对其收敛条件进行了证明。在模拟数据上的实验表明，在不同噪声比例的数据中，本文所提方法相比基于随机游走的方法具有更好的抗噪性。基于真实数据的实验表明，将基于账户交易特征的方法的结果作为所提方法的输入得到的结果，比单独使用两种方法得到的结果更好。

今后可以进一步研究将所提方法应用到更复杂的网络中，例如在网络中加入 IP 地址和 MAC 地址信息。

## 参 考 文 献


[1] YU H F, KAMINSKY M, GIBBONS P B, et al. Sybilguard: Defending against sybil attacks via social networks[J]. IEEE/ACM Transactions on Networking(TON), 2008, 16(3): 576-589.

[2] YU H, GIBBONS P B, KAMINSKY M, et al. Sybillimit: A nearoptimal social network defense against sybil attacks[J]. IEEE/ACM Transactions on Networking, 2010, 18(3): 885-898.

[3] DANEZIS G, MITTAL P. Sybilinfer: Detecting sybil nodes using social networks[C]// Proceedings of the Network and Distributed System Security Symposium, NDSS. San Diego, California, USA, 2009.

[4] NGUYEN T, JINYANG L, LAKSHMINARAYANAN S, et al. Optimal sybil-resilient node admission control[C]// Proceedings of IEEE INFOCOM. Shanghai, China, 2011.

[5] WEI W, FENGYUANF X, CHIU C T, et al. Sybildefender: A defense mechanism for sybil attacks in large social networks[J]. IEEE Transactions on Parallel and Distributed Systems, 2013, 24(12): 2492-2502.

[6] LU S, SHUCHENG Y, WENJING L, et al. Sybilshield: An agent-aided social network-based sybil defense among multiple communities[C]// Proceedings IEEE INFOCOM. Turin, Italy, 2013.

[7] CAO Q, SIRIVIANOS M, YANG X, et al. Aiding the detection of fake accounts in large scale social online services[C]// Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation(NSDI'12). 2012: 15.

[8] ZOLTA G, HECTOR G M, JAN P. Combating web spam with trustrank[C]// Proceedings of the 30th International Conference on Very Large Databases. 2004.

[9] YANG C, HARKREADER R, ZHANG J, et al. Analyzing spammers' social networks for fun and profit: A case study of cyber criminal ecosystem on twitter[C]// Proceedings of the 21st International Conference on World Wide Web. New York, NY, USA: ACM, 2012: 71-80.


(下转第 134 页)

[11] WANG J Y, FENG L X, ZHENG X F, et al. Research Status and Development Trends of Access Control Model [J]. Journal of Central South University (Science and Technology), 2015, 46(6): 2090-2097. (in Chinese)
王静宇,冯黎晓,郑雪峰. 一种面向云计算环境的属性访问控制模型[J]. 中南大学学报(自然科学版),2015,46(6):2090-2097.

[12] LI F H, WANG W, MA J F, et al. Action-based Access Control Model and Administration of Actions [J]. Acta Electronica Sinica, 2008, 36(10): 1881-1890. (in Chinese)
李凤华,王巍,马建峰,等. 基于行为的访问控制模型及其行为管理[J]. 电子学报,2008,36(10):1881-1890.

[13] SU M, LI F H, SHI G Z. Action-based Multilevel Access Control Model [J]. Journal of Computer Research and Document, 2014, 51(7): 1604-1613. (in Chinese)
苏铓,李凤华,史国振. 基于行为的多级访问控制模型[J]. 计算机研究与发展,2014,51(7):1604-1613.

[14] LANG B. Access Control Oriented Quantified Trust Degree Representation Model for Distributed Systems [J]. Journal on Communications, 2010, 31(12): 45-54. (in Chinese)
郎波. 面向分布式系统访问控制的信任度量化模型[J]. 通信学报,2010,31(12):45-54.

[15] FU X, XU S, ZHOU D M. Research on Trust-based Access Control Model in Cloud Computing Environment [J]. Computer Technology and Development, 2015, 25(9): 139-143. (in Chinese)
付雄,徐松,周代明. 云计算环境下基于信任的访问控制模型研究[J]. 计算机技术与发展,2015,25(9):139-143.

[16] DU S, JOSHI J B D. Supporting Authorization Query and Inter-domain Role Mapping in Presence of Hybrid Role Hierarchy[C]// Proceedings of the 11th ACM Symposium on Access Control Models and Technologies. New York: ACM, 2006: 228-236.

[17] YANG L, TANG Z, LI R F, et al. Roles Query Algorithm in Cloud Computing Environment Based on User Require [J]. Journal on Communications, 2011, 32(7): 169-175. (in Chinese)
杨柳,唐卓,李仁发,等. 云计算环境中基于用户访问需求的角色查找算法[J]. 通信学报,2011,32(7):169-175.

[18] ZHANG Y, JOSHI J B D. Uaq: A Framework for User Authorization Query Processing in RBAC Extended with Hybrid Hierarchy and Constraints[C]// Proceedings of the 13th ACM Symposium on Access Control Models and Technologies. New York: ACM, 2008: 83-92.

[19] LU J, JOSHI J B D, JIN L, et al. Towards Complexity Analysis of User Authorization Query Problem in RBAC [J]. Computers & Security, 2015, 48(C): 116-130.

(上接第 128 页)

[10] KOLDA T G, PROCOPIO M J. Generalized badrank with graduated trust[R]. Sandia National Laboratories, 2009.

[11] LESNIEWSKILAAS C, KAASHOEKM F. Whanau: A sybil-proof distributed hash table[C]// Proceedings of the 7th USENIX Conference on Networked Systems Design and Implementation. Berkeley, CA, USA: USENIX Association, 2010: 8.

[12] MOHAISEN A, YUN A, KIM Y. Measuring the mixing time of social graphs[C]// Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement. New York, NY, USA: ACM, 2010: 383-389.

[13] BEHRENDS E. Introduction to markov chains: with special emphasis on rapid mixing[M]// Advanced Lectures in Mathematics, 2000.

[14] BLONDEL V D, GUILLAUME J L, LAMBIOTTE R, et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, 2008(10): 155-168.

[15] PANDIT S, CHAU D H, WANG S, et al. Netprobe: A fast and scalable system for fraud detection in online auction networks [C]// Proceedings of the 16th International Conference on World Wide Web. New York, NY, USA: ACM, 2007: 201-210.

[16] SHEBUTI R, LEMAN A. Collective Opinion Spam Detection: Bridging Review Networks and Metadata[C]// Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15). ACM, New York, NY, USA, 2015: 985-994.

[17] PEARL J. Probabilistic reasoning in intelligent systems: Networks of plausible inference[M]. San Francisco: Morgan Kaufmann Publishers Inc., 1988.

[18] JIA J, WANG B, ZHANG L, et al. AttriInfer: Inferring User Attributes in Online Social Networks Using Markov Random Fields[C]// International Conference on World Wide Web. 2017: 1561-1569.

[19] GATTERBAUER W, GUNNEMANN S, KOUTRA D, et al. Linearized and single-pass belief propagation[J]. Proceedings of the Vidb Endowment, 2014, 8(5): 581-592.

[20] WANG B, GONG N Z, FU H. Gang: Detecting fraudulent users in online social networks via guilt-by-association on directed graphs[C]// 2017 IEEE International Conference on Data Mining (ICDM). New Orleans, LA, USA, 2017: 465-474.

[21] SAAD Y. Iterative methods for sparse linear systems (2nd ed) [M]. Reading, MA: Society for Industrial and Applied Mathematics, 2003.

[22] HOLME P, KIM B J. Growing scale-free networks with tunable clustering[J]. Physical Review E, 2002, 65(22): 026107.

[23] BAHNSEN A C, AOUADA D, STOJANOVIC A. Feature engineering strategies for credit card fraud detection[J]. Expert Systems with Applications An International Journal, 2016, 51(C): 134-142.