基于 k-度匿名的社会网络隐私保护方法

龚卫华[1],兰雪锋[1],裴小兵[2],杨良怀[1]
(1. 浙江工业大学计算机科学与技术学院,浙江杭州 310023; 2. 华中科技大学软件学院,湖北武汉 430074)

摘 要: 针对当前社会网络的匿名化隐私保护方法存在信息损失量巨大、网络关系结构被改变严重等问题,提出一种保持网络结构稳定的 k-度匿名隐私保护模型 SimilarGraph,运用动态规划方法对社会网络按照节点度序列进行最优簇划分,然后采用移动边操作方式重构网络图以实现图的 k-度匿名化. 区别于传统的数值扰乱或图修改如随机增加、删除节点或边等方法,该模型的优势在于既不增加网络边数和节点数,也不破坏网络原有连通性和关系结构. 实验结果表明,SimilarGraph 匿名化方法不仅能有效提高网络抵御度属性攻击的能力,并且还能保持网络结构稳定,同时具有较理想的信息损失代价.

关键词: 社会网络; 隐私保护; k-度匿名; 信息损失
中图分类号: TP309.2 **文献标识码:** A **文章编号:** 0372-2112 (2016)06-1437-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2016.06.026

Privacy Preservation Method Based on k-Degree Anonymity in Social Networks

GONG Wei-hua[1], LAN Xue-feng[1], PEI Xiao-bing[2], YANG Liang-huai[1]
(1. *School of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, Zhejiang* 310023, *China*;
2. *School of Software Engineering, Huazhong University of Science and Technology, Wuhan, Hubei* 430074, *China*)

Abstract: To preserve the privacy of social networks, most existing methods are applied to satisfy different anonymity models, but some serious problems are involved such as often incurring large information losses and great structural modifications of original social network after being anonymized. Therefore, an improved privacy protection model called SimilarGraph is proposed, which is based on k-degree anonymous graph derived from k-anonymity to keep the network structure stable. Where the main idea of this model is firstly to partition network nodes into optimal number of clusters according to degree sequences based on dynamic programming, and then to reconstruct the network by means of moving edges to achieve k-degree anonymity with internal relations of nodes considered. To differentiate from traditional data disturbing or graph modifying method used by adding and deleting nodes or edges randomly, the superiority of our proposed scheme lies in which neither increases the number of nodes and edges in network, nor breaks the connectivity and relational structures of original network. Experimental results show that our SimilarGraph model can not only effectively improve the defense capability against malicious attacks based on node degrees, but also maintain stability of network structure. In addition, the cost of information losses due to anonymity is minimized ideally.

Key words: social network; privacy preservation; k-degree anonymity; information loss

1 引言

近年来,社会网络的流行已深刻地改变了人们的日常生活和交流方式,国内外著名社交网站如 Facebook、QQ、人人网等注册用户数量不断攀升,以 Facebook 为例,用户总数在 2013 年已突破 10 亿,其中包含 1500 亿条朋友链接,这些社会网络数据蕴含巨大的商业价值和应用前景,例如可促进广告、游戏、零售等业务迅速增长. 然而,人们在使用基于社会网络的应用同时面临着严重的隐私信息泄露和恶意攻击问题. 因此,研

收稿日期:2015-01-25;修回日期:2015-10-23;责任编辑:梅志强
基金项目:浙江省自然科学基金(No. LY13F020026, No. Y1080102, No. LY14F020017, No. LY14C130005);国家自然科学基金(No. 61571400, No. 61070042);中国博士后科学基金(No. 2015M581957);浙江省博士后科研项目择优资助(No. BSH1502019)

究社会网络的隐私保护技术显得尤为重要.

社会网络属于复杂网络的研究范畴,关注的是社会个体及个体间的互动和联系,同样具有“小世界”现象和幂律分布特征[1~3],但这使得社会网络所包含的2类重要隐私信息(节点属性数据和关系数据)极易遭受节点度攻击、链接攻击等结构化攻击. 目前针对社会网络的隐私保护问题已取得一些研究成果,如从节点属性数据角度出发的隐私保护类似于数据发布研究中的隐私保护方法[4~6],侧重保护标识或敏感属性如姓名、电话、地址等,常采用已比较成熟的数据泛化[7~10]、扰动[5,11]或添加噪声节点[12]等方法. 而针对关系数据的隐私保护则是亟待人们深入探索的研究热点,通常被建模为图数据并采用数值扰乱法或图修改法如随机增加、删除节点或边[13],以及修改边权重值[14]来实现隐私保护. 总体上看,现有的社会网络隐私保护方法大多基于如何实现各种匿名化模型如节点 k-匿名、子图 k-匿名等[15],但他们都面临由于匿名化而带来巨大的信息损失问题,甚至还会严重破坏社会网络关系结构,显著降低了网络数据的效用.

本文针对社会网络中关系数据这类隐私对象提出一种改进的基于图的 k-度匿名模型 SimilarGraph,该模型首先运用动态规划思想进行基于节点度的最优簇划分,然后,通过移动边方式重构网络图实现图的 k-度匿名化. 该方法不仅能克服传统匿名化算法所存在严重的信息损失缺点,还有效保持了社会网络原有连通性和内在关系结构稳定,并提高了抵御度属性攻击的能力.

2 相关工作

目前,现有针对网络关系数据的隐私保护研究大多数都采用匿名化模型来防止隐私信息泄露和恶意攻击,其主要途径有基于聚类方法和图修改方法.

基于聚类的匿名化方法是先对节点、边或两者同时聚类成簇,然后通过泛化方式来达到匿名化效果. 文献[16]提出将网络中相似节点聚合为一簇,每个簇所包含的节点数$\geqslant k$,这样使得攻击命中率降为 $1/k$. Campan 等[8]采用贪心策略对网络中属性相似的节点进行聚类并使用边泛化方法实现 k 匿名的网络,该方法考虑了匿名化过程中的信息损失问题. 文献[17]对加权无向网络采用节点聚类和边聚类相结合的泛化方式实现 k-匿名模型,但缺点是严重改变了网络结构,同时还降低了匿名化后的网络数据效用.

近年来,采用图修改方法实现网络匿名化已成为国内外研究者关注的热点,Liu 等[18]提出图的 k-度匿名概念,即要求图中任一顶点都至少有 $k-1$ 个顶点与其度数相同,并运用贪心策略采用增加边的方式来实现匿名图,以抵御节点度属性攻击,该方法虽然考虑了图修改的最小代价问题,但破坏了网络连通性使得网络内在关系结构发生重大变化. Yuan[19]和 Zhou[20]都针对具有节点属性标签的社会网络提出了 k-度 $-l$ 多样化匿名模型,该模型在 k 度匿名的基础上要求相同度数的 k 个节点必须有 l 种不同标签,并通过增删边和添加噪声节点的方法实现属性匿名,但他们都没有考虑匿名化所造成的信息损失影响. Zheleva 等[21]将关系边区分为敏感边和非敏感边并提出通过删除敏感边的方式实现图的匿名化,以防止链接再识别攻击,其不足之处在于数据匿名化的效用由删除边的数量多少决定,缺乏对信息损失量的考虑,严重破坏原有网络的连通性. 此外,Zou 等[22]运用图同构理论提出 k-同构匿名模型防御结构化攻击,要求网络任一子图至少有 $k-1$ 个与其同构的子图,其缺点是同构图的匹配和重构造代价较大,特别是图转化时需要复制边的操作破坏了原有网络的结构特性.

综上所述,基于聚类的匿名模型由于泛化后存在严重的信息损失问题,导致网络结构发生巨大变化,数据效用急剧降低. 而针对图数据修改或转化的匿名化方法大多都采用添加、删除节点或边以及子图同构等扰动方式实现 k-度匿名,但这种图随机修改策略忽略了社会网络内在结构特性,仍无法克服较大的信息损失问题. 为此,本文提出的隐私保护模型 SimilarGraph 与传统的数值扰乱或图修改方法不同之处在于采用移边方式替代随机增、删节点或边等操作,并能在网络节点数和边数都保持不变条件下以最小的信息损失代价移动关系边实现网络的 k-度匿名化,因而既不损害社会网络原有连通性和关系结构,还有效提高了抵御度属性攻击的能力.

3 相关定义

为了便于研究,本文将社会网络建模为无权无向图 $G=(V,E)$,其中 V 表示为社会网络中的节点集,E 表示节点间的关系边集,且 $E\subseteq V\times V$. 一般情况下,图中节点及其关系极易受到节点度攻击、链接攻击等结构化攻击,因此,实现图中节点及关系边的匿名化是一种重要的隐私保护方法,下面先给出一些基本定义.

定义 1 图的 k-度匿名模型:给定图 $G'=(V',E')$,$\forall v_i'\in V'$,$\exists m(m\geqslant k-1)$ 个节点的度数与 $D_g(v_i')$ 相等,则称该图 G' 满足 k-度匿名模型,其中 $D_g(v_i')$ 表示节点 v_i' 的度数.

图的 k-度匿名借鉴了传统数据表中的 k-匿名思想[11],使得图中节点间关系及其度分布趋于同构,这将有效降低结构化攻击的概率,至少小于等于 $1/k$. 从另一角度看,社会网络可看成由若干子图构成,每个子图

都满足 k-度匿名模型,这样得出网络的 k-度匿名概念.

定义 2　网络的 k-度匿名化:将社会网络图 $G=(V,E)$ 中的节点按照度数划分成若干簇 $C^L=\{c_1^{'},c_2^{'},\cdots,c_m^{'}\}$,其中任意簇都是一个满足 k-度匿名的子图,即 $c_m^{'}=\{v_{m_1},\cdots,v_{m_j}\mid m_j\geqslant k\}$ 且 $D_g(v_{m_1})=\cdots=D_g(v_{m_j})$.

由定义 2 可知,社会网络被划分成满足 k-度匿名的各簇实际上可称为匿名簇,同一簇内的节点都具有相同的度属性,而不同的匿名簇间满足不同的 k-度匿名要求.对于相同簇中的节点由于具有同构特征而不易受攻击,并且如果簇越大、簇数量越多,其遭受攻击的难度也越大.因此,当社会网络被划分成满足定义 2 的 m 个簇时,受到恶意攻击的概率将进一步下降到 $1/(m\cdot k)$.

为了便于社会网络按照节点度特征划分成各匿名簇,下面给出基于递减度的序列结构.

定义 3　递减度的节点序列 $S^q(\langle v_1\cdots v_i\rangle)$:如果网络图 G 的节点集 $V=\{v_1,\cdots,v_i\}$ 中所有节点按照递减度的偏序关系排列,即满足 $D_g(v_1)\geqslant\cdots\geqslant D_g(v_i)$,则该递减度节点序列表示为 $S^q(\langle v_1\cdots v_i\rangle)$.

根据定义 3,如果节点序列 $S^q(\langle v_1\cdots v_i\rangle)$ 中所有节点的度数都相等,并且序列的节点数 $|S^q|\geqslant k$,则该序列 S^q 可看作一个符合 k-匿名要求的簇序列.

当社会网络节点被划分到不同的簇序列 $c_i^{'}$ 时,为了满足簇的匿名度要求,节点需要通过删除或增加边来改变原先节点度数,而节点度变化会随之影响网络原有结构,因此,本文给出簇的信息损失量定义.

定义 4　簇的信息损失量 $I(c_m^{'})$:当节点序列 $S^q(\langle v_1\cdots v_i\rangle)$ 被划分成满足某个 k-度匿名的簇序列 $c_m^{'}$ 时,簇内每个节点匿名化前后的度变化之和称为簇序列的信息损失量 I,即为:

$$I(c_m^{'})=\sum_{v_i\in c_m^{'}}|D_g^{'}(c_m^{'})-D_g(v_i)| \tag{1}$$

其中,$D_g(v_i)$ 表示节点 v_i 匿名化前的度数,而 $D_g^{'}(c_m^{'})$ 表示簇序列 $c_m^{'}$ 的匿名化度数,等于簇内所有节点度的平均值即

$$\lfloor\frac{1}{|c_m^{'}|}\sum_{v_i\in c_m^{'}}D_g(v_i)\rfloor$$

这里 $|c_m^{'}|$ 表示簇 $c_m^{'}$ 的节点数.

定义 4 中,簇的信息损失量衡量了单个匿名簇内节点度变化对网络原有结构造成的影响程度.在此基础上,可进一步通过累加所有匿名簇的信息损失量获得整个社会网络匿名化的信息损失代价,即原始网络 G 与匿名网络 G' 间的节点度变化量为:

$$\begin{aligned}I(G'/G)&=\sum_{c_m^{'}\in C^{L}}I(c_m^{'})\\&=\sum_{c_m^{'}\in C^{L}}\sum_{v_{m_i}\in c_m^{'}}|D_g^{'}(c_m^{'})-D_g(v_{m_i})|\end{aligned} \tag{2}$$

定义 5　信息损失率(R):满足 k-度匿名的社会网络 G' 的信息损失量与其原始网络 G 中总度数的比值称为信息损失率:

$$R=\frac{I(G'/G)}{2\times|E|} \tag{3}$$

式(3)中,$I(G'/G)$ 表示整个社会网络匿名化的信息损失量,由式(2)计算;而对于原始网络 G 的节点总度数,由图的握手定理可得:当网络 G 的边数为 $|E|$ 时,其总度数和为 $2|E|$.

4　图的 k-度匿名隐私保护方法

针对建模成图结构的社会网络,本文提出基于移边操作的 k-度匿名隐私保护方法,基本思路是将整个匿名化过程分为两个步骤:(1)基于度的最优簇划分;(2)移边操作重构网络图实现 k-度匿名化.

4.1　基于度的最优簇划分

最优簇划分是以信息损失量最小化代价为目标对网络节点进行簇划分,并确定簇内每个节点满足 k-度匿名的度数.为了实现该目标,本文先将社会网络 $G=(V,E)$ 中节点集 V 按照定义 3 排序成递减度序列形式:

$$\begin{aligned}S^q(\langle v_1v_2\cdots v_n\rangle)=\{&\langle v_1v_2\cdots v_n\rangle\mid\forall i,j=1,\cdots,n,\\&\text{当 }i<j\text{ 时 }D_g(v_i)\geqslant D_g(v_j)\}\end{aligned}$$

然后基于节点度划分成 m 个匿名簇,并使其满足定义 2 中的 k-度匿名要求,这样匿名簇的度序列转变为如下结构:

$$\begin{aligned}&S^{q'}(\langle v_{11}v_{12}\cdots v_{1t_1},v_{21}v_{22}\cdots v_{2t_2},v_{m1}v_{m2}\cdots v_{mt_m}\rangle)\\&=\{c_1^{'},c_2^{'},\cdots,c_m^{'}\mid\forall i=1\cdots m,c_i^{'}\\&=\langle v_{i1}v_{i2}\cdots v_{it_i}\rangle,\\&t_i\geqslant k\text{ 且 }D_g(v_{i1})=D_g(v_{i2})=\cdots=D_g(v_{it_i})\}\end{aligned}$$

可以看出,对整个社会网络节点的簇划分等价于递减度序列的簇划分,并且要求信息损失量最少.我们采用动态规划方法对递减度序列结构 S^q 进行簇划分,动态规划特别适合具有重叠子过程的多阶段决策问题,要求出一个过程的最优解必须求出其子过程的最优解,这样逐步递推直到求出整个过程的最优解.因此,本文提出最优簇划分的代价函数如式(4)所示.

$$\begin{aligned}&S_L^Q(\langle v_1\cdots v_n\rangle)\\&=\text{Min}\Big\{\min_{k\leqslant t\leqslant n-k}\{(I(\langle v_1\cdots v_t\rangle)+S_L^Q(\langle v_{t+1}\cdots v_n\rangle)\},\\&\quad I(\langle v_1\cdots v_n\rangle)\Big\}\end{aligned} \tag{4}$$

约束为

$$\begin{cases}S_L^Q(\langle v_i\cdots v_n\rangle)=\sum_{m=i}^{n}|D_g(v_m)-D_g^{'}(\langle v_i\cdots v_n\rangle)|,\\\quad\text{当节点序号 }n-2k+1<i\leqslant n-k+1\end{cases} \tag{5}$$

式(4)中,$S_L^Q(\langle v_{t+1}\cdots v_n\rangle)$ 表示子序列的最优簇划

分代价即最小信息损失量,而 $I(\langle v_1 \cdots v_t \rangle)$ 则表示簇序列的信息损失量,可根据式(1)计算. 式(5)中,$S_L^Q(\langle v_i \cdots v_n \rangle)$ 为终端条件,表示从 v_n 开始向前划分的子簇序列的最小代价. $D_g'(\langle v_i \cdots v_n \rangle)$ 表示该子簇序列的平均度,而 $D_g(v_m)$ 表示匿名前的节点度数.

综合式(4)和式(5),当序列 $S^q(\langle v_1 \cdots v_n \rangle)$ 中 $n < 2k$ 时,无法再划分成满足 k-度匿名要求的子簇,因而整个序列将自成一簇;当序列 $S^q(\langle v_1 \cdots v_n \rangle)$ 中 $n \geqslant 2k$ 时,整个序列的子簇划分候选方案共有 $n-2k+1$ 种,而其中每种候选子簇 $\langle v_{t+1} \cdots v_n \rangle$ 的划分又是一个递归调用,其最小划分代价 $S_L^Q(\langle v_{t+1} \cdots v_n \rangle)$ 则需由终端状态开始逆向递推计算,直至获得所有候选方案的最优划分结果. 然后,再将该序列的最小划分代价与其单独成簇时的信息损失量比较. 基于该过程,最优簇划分算法的具体实现见算法 1.

算法 1 最优簇划分算法

输入:网络图 G 中的节点递减度序列 $S^q(v_1, v_2, \cdots, v_n)$,匿名 k 度值.
输出:最优匿名簇 $S^{q'}$ 的划分序列号 $t_1, \cdots, t_m$.

```
1.  if n < 2k then
2.     return 簇序列 S^q'(v1, v2, ..., vn);
3.  else  //对于 n≥2k 情况
4.  for i = n - k + 1 to k do
5.        if  i > n - 2k + 1 then  //当 n - 2k + 1 < i≤n - k + 1 时
6.           for m = i to n do
7.           S_L^Q[i] + = |D_g(v_m) - D_g'(⟨v_i ... v_n⟩)|;
8.           endfor
9.        elseif i > k then //当 k < i≤n - 2k + 1 时
10.          由式(1)计算 I(⟨v_i ... v_n⟩);
11.       endif
12.    endfor
13.    for t = k to n - k do
14.       由式(1)计算 I(⟨v_1 ... v_t⟩);
15.    S_L^Q[t]←Min{I(⟨v_1 ... v_t⟩) + Min_IL(⟨v_{t+1} ... v_n⟩)};  //
       递归调用函数 Min_IL 获得子簇最小划分代价
16.    endfor
17.    Min{S_L^Q[t], I(⟨v_1 ... v_n⟩)};//由式(4)选取最优簇划分,如
       果小于则 t_1 = t,否则 t_1 = 1
18.    return 最优簇序列 S^q' 的划分序号[t_1, ..., t_i];
19. endif
```

算法 1 中,步骤 4 ~ 12 计算子序列 $S^q(\langle v_i \cdots v_n \rangle)$ 终端状态下的最小划分代价以及其单独成簇时的信息损失量,步骤 13 ~ 17 则从整个序列 $S^q(\langle v_1 \cdots v_n \rangle)$ 的 $n-2k+1$ 种候选划分方案中选取最小划分代价,其中步骤 15 通过递归函数 Min_IL 实现最优的子簇划分目标,即 $\text{Min}\{(I(\langle v_i \cdots v_t \rangle) + S_L^Q(\langle v_{t+1} \cdots v_n \rangle), I(\langle v_i \cdots v_n \rangle)\}$.

4.2 网络图重构算法

经过最优簇序列划分后,网络图中每个节点将获得实现 k-度匿名化所属簇的平均度数. 本文采用移边方式实现匿名化操作,即将高于簇平均度的节点上的边移动到低于簇平均度的节点上. 实际上,移边操作可等价于先删除边再增加边这两步原子操作,成功的移边操作应使其两端节点都同时满足度匿名的变化方向.

假设任意节点 v_i 的现有度数 D_g 与其所属匿名簇 c' 的平均度数 D_g' 之间的关系函数 $\gamma(v_i)$ 如式(6):

$$\gamma(v_i) = \begin{cases} D_g(v_i) < D_g'(c'), & \text{节点 } v_i \text{ 需要增加边} \\ D_g(v_i) > D_g'(c'), & \text{节点 } v_i \text{ 需要删除边} \\ D_g(v_i) = D_g'(c'), & \text{节点 } v_i \text{ 满足匿名化} \end{cases} \tag{6}$$

对于网络中的任意边来说,其两端节点 v_i 和 v_j 的函数 γ 状态共同决定了该边是否符合增删操作要求,如图 1 所示 6 种状态,除了图 1(f)中边上两端节点都已满足匿名化要求外,剩余 5 种情况图 1(a) ~ (e)都需要通过增删边来改变节点度数. 不难得知,由于图 1(b)、(c)和(e)都至少有一端存在度关系“ < ”,因而不满足移边操作中需先删除边的前提条件,而只有图 1(a)和(d)满足该前提条件,且节点度符合匿名变化方向.

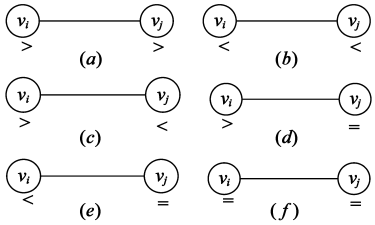


图1 任意边的两端节点度匿名关系

为了保持图结构的连通性,移边操作中的删除边与增加边间存在必要的关联条件是这两条边的端点在图中体现互为连通邻居. 具体地,针对图 1(a)和图 1(d)的移边方法分别对应图 2(a)和图 2(b),图中移边的先后步骤等于①删除边 + ②增加边(虚线表示). 图 2(a)中新增边的两节点 v_p 和 v_q 分别是被删边上节点 v_i 和 v_j 的连通邻居,并且都有增加节点度要求. 而图 2(b)中为了维持被删边上的节点 v_j 度不变的要求,新增边的一端必须从 v_j 出发,而另一端则是 v_i 中需增加节点度的连通邻居.

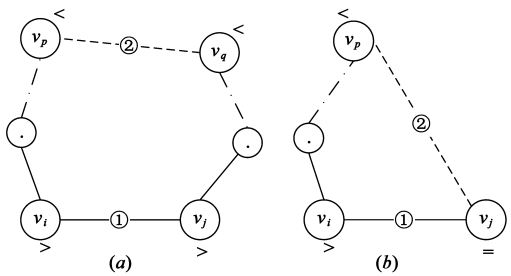


图2 两种移边匿名化操作方法

为了实现基于移边的网络图匿名化,本文给出满足 k-度匿名的重构网络图算法 2,算法中假设已知原始图中各节点 v_i 的度数 $D_g(v_i)$.

算法 2　重构网络图算法

输入:原始网络图 $G=(V,E)$ 和划分匿名簇 C^L 的平均度 $\{D_g'(c_1'),D_g'(c_2'),\cdots,D_g'(c_m')\}$

输出:重构后的 k-度匿名网络图 G'

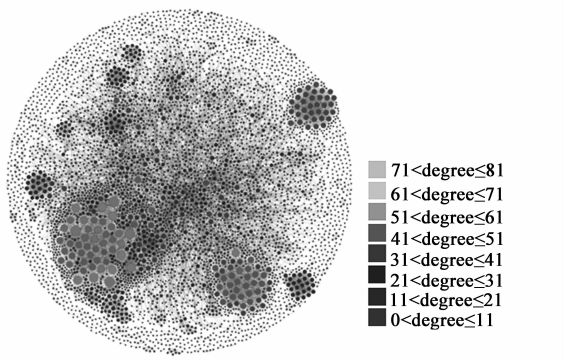
```
for each edge(v_i,v_j) ∈ E do
if D_g(v_i) > D_g'(v_i) && D_g(v_j) > D_g'(v_j) then
    for v_p ∈ N (v_i 的连通分量) do
        if D_g(v_p) < D_g'(v_p) then
            for v_q ∈ N (v_j 的连通分量) do
            if D_g(v_q) < D_g'(v_q) && ∄ edge(v_p,v_q) then
                {删除 edge(v_i,v_j)后两端节点度 -1;
                增加 edge(v_p,v_q)后两端节点度 +1;}
            endif
        endfor
    endif
endfor
else if D_g(v_i) > D_g'(v_i) && D_g(v_j) = = D_g'(v_j) then
    for v_p ∈ N(v_i 的连通分量) do
            if D_g(v_p) < D_g'(v_p) && ∄ edge(v_p,v_j) then
                {删除 edge(v_i,v_j)后节点 D_g(v_i) -1;
                增加 edge(v_p,v_j)后节点 D_g(v_p) +1;}
            endif
    endfor
    endif
endfor
return 重构后的网络 G'
```

5　仿真实验及结果分析

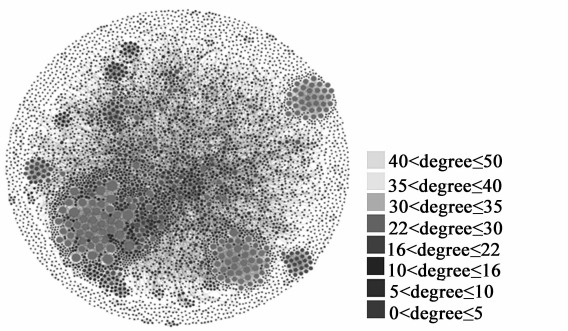
本文采用 CA-GrQc 数据集构建社会网络进行实验与分析,该数据集包括 5242 个节点,14496 条无向边,度分布服从幂律分布. 为了便于实验比较和说明,我们将第 4 节所提出的社会网络基于图的 k-度匿名隐私保护方法称为 SimilarGraph 模型,算法代码用 Python 编程实现,实验环境为 Intel(R) Core™ i5 CPU 2.3GHz,4GB 内存,操作系统为 Windows7. 实验方法是先由算法 1 对原始网络数据集进行最优的 k-度匿名簇划分,再用算法 2 进行移边操作来重构匿名化的网络图,然后采用 Gephi 工具对其可视化并对比网络匿名化前后节点度变化及分布特征.

图 3(a)展示了原始社会网络的节点度分布图,节点度数越多则呈现越大,图中共标注了 8 种度区间的节点分布情况. 图 3(b)则显示当 $k=50$ 时匿名化网络的分布图,其度特征明显下降,节点共被划分成 21 个簇,与图 3(a)对比后发现,原始社会网络中节点度大于 70 的显著节点只有 4 个,对其成功攻击的概率有 1/4,而在匿名后的图 3(b)中,至少有 50 个以上节点与其相似,这样攻击概率便降至 1/50 以下.

图 4 显示了不同匿名 k 值下社会网络度的幂律分布规律,图中 $k=0$ 时表示原始社会网络的度服从幂律分布,其度数介于 10 到 80 之间的节点分布不均匀且同构节点数偏少,度数大的节点最容易遭受攻击,而实现不同 k-度匿名化后的网络度分布虽然也满足幂律特征,但其结构趋于均匀,最大节点度数随着匿名 k 值增大而逐渐减少,节点聚集特性也越明显,特别是当 k 值越大时匿名网络中节点度大于 10 以上的同构节点数越多,这样大大增加了针对网络度属性攻击的难度.



(a) 原始社会网络的节点度分布



(b) k=50时匿名化的社会网络节点度分布

图3　社会网络匿名化前后对比图

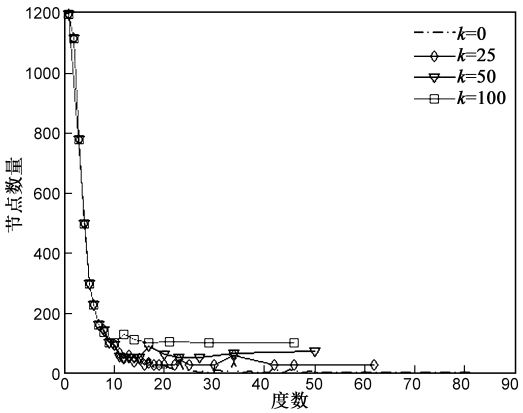


图4　不同k-度匿名值下的网络度幂律分布

下面,将本文提出的模型 SimilarGraph 与经典的 k-

度匿名方法 SuperGraph[18] 和最近 Yuan 等[19] 提出的模型 KDLD 进行各项实验指标对比,三者区别在于 SimilarGraph 采用移边方法而 SuperGraph 则采用随机增加边方式实现网络匿名化,对于 KDLD 则是通过增加噪声节点来实现 k-度匿名化. 图 5 比较了三种方法在实现不同 k-度匿名化网络过程中发生边移动、增加或因噪声节点而增加边的变化数量,当匿名 k 值增大时,SimilarGraph 实现匿名化所需移动的边数增长较小且比较平稳,而 SuperGraph 所需改变的边数从 222 增加到 2675 条,KDLD 也与其较一致,增长幅度都很显著. 总体上看,SimilarGraph 的边变化数远小于 SuperGraph 和 KDLD.

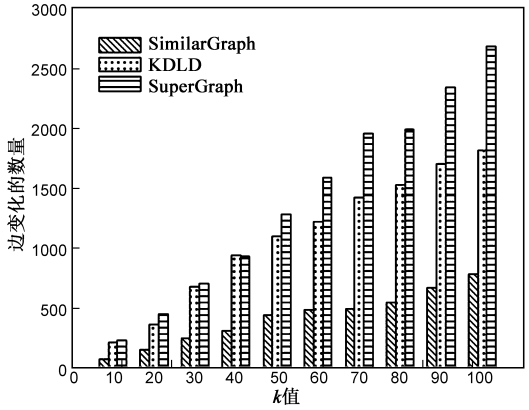


图5 k-度匿名化网络中的边变化数

图 6 进一步统计了三种方法实现匿名化后带来的信息损失率结果,该指标由式(3)计算. 图 6 中 SimilarGraph 在实现不同 k 值匿名化网络时由移边操作所引起的信息损失率非常小,而 SuperGraph 和 KDLD 两者都增加了大量边而造成较大的信息损失率且增长趋势较明显,由此可见,SimilarGraph 方法具有最理想的移边代价.

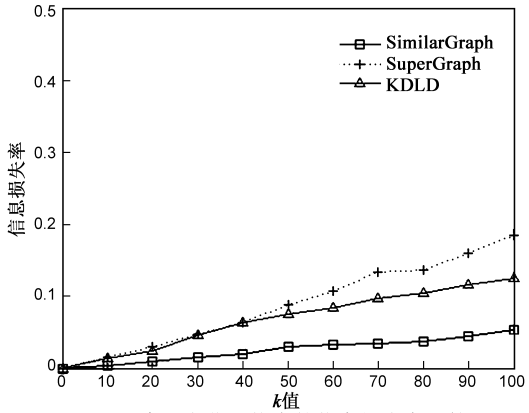


图6 k-度匿名化网络中的信息损失率比较

另外,为了对比网络匿名化前后的结构特性变化,图 7、图 8 和图 9 分别给出了三种方法在不同 k-度匿名化网络中的聚类系数(CC)、节点平均度和平均路径长度(APL)等指标结果,图中用虚线表示了原始网络的相关指标值,它不随匿名 k 值而变化. 由图 7 可知,KDLD 方法当 k 在 50 ~ 70 区间时由于增加了一些噪声节点以及需增加、删除相关边,导致其 CC 指标出现较明显的先升后降趋势,整体网络结构变化较大,表现不稳定,而 SuperGraph 方法随 k 值增大而所增边数越多造成 CC 指标逐渐下降. 总体上看,本文的 SimilarGraph 方法在不同 k 值下一直最接近于原始网络的聚类系数值,对匿名化后的网络结构影响最小.

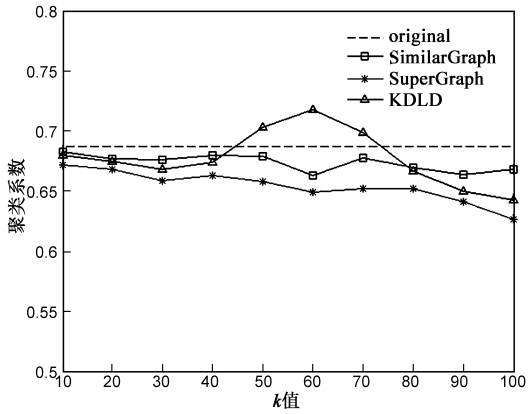


图7 不同k-度匿名化网络中的聚类系数

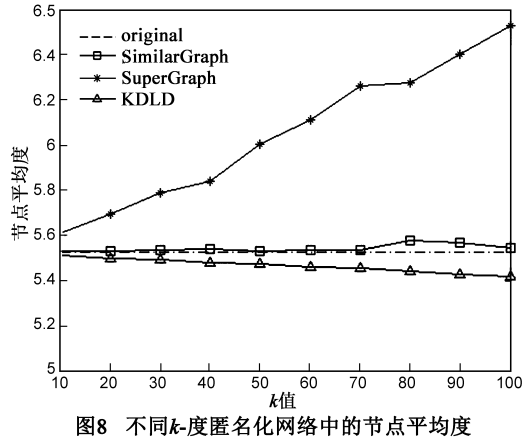


图8 不同k-度匿名化网络中的节点平均度

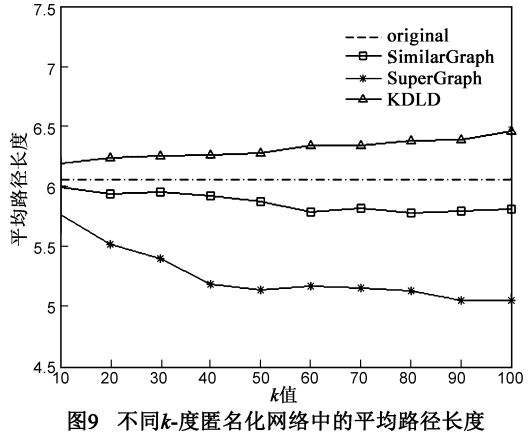


图9 不同k-度匿名化网络中的平均路径长度

图 8 中当匿名 k 值增大时,SimilarGraph 产生的匿

名化网络中节点平均度数与原始网络基本相同,而 KDLD 方法使得不同 k 值匿名化的网络节点平均度逐渐下降,对网络结构影响较小,SuperGraph 则使匿名后的节点平均度增幅较大,表明该匿名方法比较严重地破坏了原始网络结构.

图9 比较了网络匿名化前后的平均路径长度(APL)指标,三者之中本文的 SimilarGraph 表现最好,该方法使得匿名化的网络 APL 在不同 k 值下都保持较小幅的下降且比较平稳,而 KDLD 在匿名化后由于增加了一些噪声节点导致 APL 指标有小幅度上升,SuperGraph 则采用随机增加边方式引起匿名化网络的 APL 指标有较大的下降. 由此表明,SimilarGraph 能保持比较稳定的网络内在关系结构.

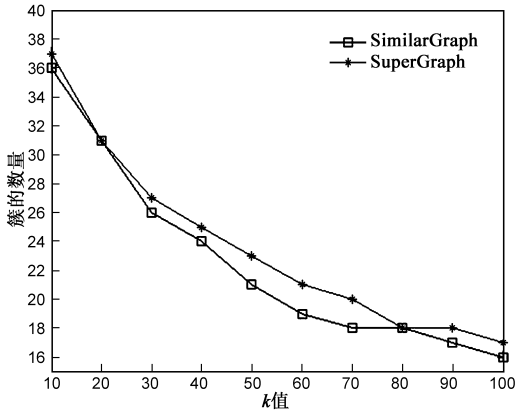


图10　不同k-度匿名系数下的划分簇数量

最后,由于本文实验所选取的数据集 CA-GrQc 中节点无属性标签,因此,KDLD 模型无法在相同条件下与 SimilarGraph 和 SuperGraph 比较抗恶意攻击能力,图 10 和图 11 分别对比了 SimilarGraph 和 SuperGraph 两种方法在不同 k-度匿名值下的网络划分簇数量和遭受度攻击的平均概率. 从图 10 统计的匿名簇数量对比来看,当匿名 k 值增大时,SimilarGraph 和 SuperGraph 两者在实现匿名化网络时所划分的簇数量都是逐渐减少且大致接近. 另一方面,图 11 中的平均攻击概率等于对所有簇节点攻击的概率平均值,概率值越小表示匿名化网络抵御节点度攻击的能力越强,由图 11 结果可知,两种方法都使得匿名化网络遭受度攻击的概率大大减小,而 SimilarGraph 抵御恶意攻击的能力总体上优于 SuperGraph.

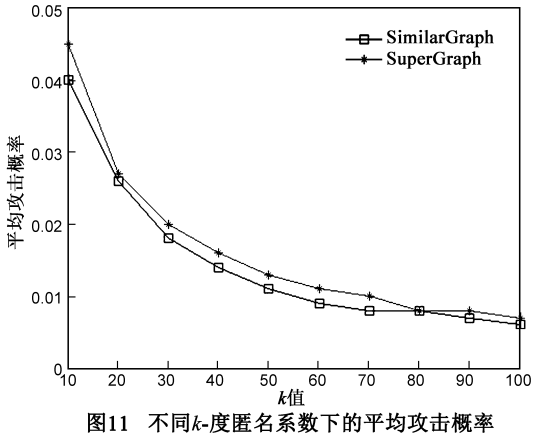


图11　不同k-度匿名系数下的平均攻击概率

6　总结

现有社会网络的隐私保护方法普遍存在比较严重的信息损失,以及匿名化后网络结构特征发生巨大改变的问题. 针对这些不足,本文提出一种保护社会网络关系数据的 k-度匿名模型 SimilarGraph,该模型先从网络节点度序列出发运用动态规划方法进行最优簇划分,然后,采用移动边方式对网络进行扰动,并进一步重构网络实现基于图的 k-度匿名化的隐私保护. 最后,采用 CA-GrQc 数据集构建社会网络进行实验与分析,各项实验结果表明 SimilarGraph 方法能在网络节点数和边数都保持不变条件下以最小的信息损失代价移动关系边实现网络的 k-度匿名化,克服了传统匿名化算法存在严重的信息损失缺点,而且还有效保持了社会网络结构和内在联系的稳定,同时提高了网络抵御度属性攻击的能力. 限于篇幅,我们下一步研究工作是改进本文所提出的匿名化模型实现并行化以求改变全局优化过程计算复杂的局面,并考虑在更大的实际网络数据集上进行实验验证其有效性.

参考文献

[1] Boccaletti S, Latora V, Moreno Y, et al. Complex networks: structure and dynamics[J]. Physics Reports, 2006, 424(4): 175 – 308.

[2] Wang X F, Chen G R. Complex networks: small-world, scale-free and beyond[J]. IEEE Circuits and Systems Magazine, 2003, 3(1): 6 – 20.

[3] Faloutsos M, Faloutsos P, Faloutos C. On power-law relationships of the internet topology[A]. ACM SIGCOMM99[C]. Cambridge, Massachusetts: ACM, 1999. 251 – 262.

[4] 童云海,陶有东,唐世渭,等. 隐私保护数据发布中身份保持的匿名方法[J]. 软件学报,2010,21(4):771 – 781.
Tong Yun-hai, Tao You-dong, Tang Shi-wei, et al. Identity-reserved anonymity in privacy preserving data publishing[J]. Journal of Software, 2010, 21(4): 771 – 781. (in Chinese)

[5] 黄茂峰,倪巍伟,王佳俊,等. 一种面向聚类的对数螺线数据扰动方法[J]. 计算机学报,2012,35(11):2275 – 2282.
Huang Mao-feng, Ni Wei-wei, Wang Jia-jun, et al. A logarithmic spiral based data perturbation method for clustering[J]. Chinese Journal of Computers. 2012, 35(11): 2275 – 2282. (in Chinese)

[6] 张啸剑,孟小峰. 面向数据发布和分析的差分隐私保护[J]. 计算机学报,2014,37(4):927 – 949.

Zhang Xiao-jian, Meng Xiao-feng. Differential privacy in data publication and analysis[J]. Chinese Journal of Computers, 2014, 37(4): 927 – 949. (in Chinese)

[7] Campan A, Truta T M, Cooper N. P-sensitive K-anonymity with generalization constraints[J]. Transactions on Data Privacy, 2010, 3(2): 65 – 89.

[8] Campan A, Truta TM. A clustering approach for data and structural anonymity in social networks[A]. 2nd ACM SIGKDD International Workshop on Privacy, Security, and Trust in KDD(PinKDD 08)[C]. Las Vegas, NV: ACM, 2008. 33 – 54.

[9] 王智慧,许俭,汪卫,等. 一种基于聚类的数据匿名方法[J]. 软件学报,2010,21(4):680 – 693.
Wang Zhi-hui, Xu Jian, Wang Wei, et al. Clustering-Based approach for data anonymization[J]. Journal of Software, 2010, 21(4): 680 – 693. (in Chinese)

[10] 张健沛,谢静,杨静,等. 基于敏感属性值语义桶分组的 *t*-closeness 隐私模型[J]. 计算机研究与发展,2014,51(1):126 – 137.
Zhang Jianpei, Xie Jing, Yang Jing, et al. A *t*-closeness privacy model based on sensitive attribute values semantics bucketization[J]. Journal of Computer Research and Development, 2014, 51(1): 126 – 137. (in Chinese)

[11] 付艳艳,张敏,冯登国,等. 基于节点分割的社交网络属性隐私保护[J]. 软件学报,2014,25(4):768 – 780.
Fu Yan-yan, Zhang Min, Feng Deng-guo, et al. Attribute privacy preservation in social networks based on node anatomy[J]. Journal of Software, 2014, 25(4): 768 – 780. (in Chinese)

[12] Wu W T, Xiao Y H, Wang W, et al. *k*-symmetry model for identity anonymization in social networks[A]. 13th International Conference on Extending Database Technology(EDBT'10)[C]. Lausanne, Switzerland: ACM, 2010. 111 – 122.

[13] Ying X, Wu X. On link privacy in randomizing social networks[J]. Knowledge and Information Systems, 2011, 28(3): 645 – 663.

[14] 刘华玲,郑建国,孙辞海. 基于贪心扰动的社交网络隐私保护研究[J]. 电子学报,2013,41(8):1586 – 1591.
Liu Hua-ling, Zheng Jian-guo, Sun Ci-hai. Privacy preserving in social networks based on greedy perturbation[J]. Acta Electronica Sinica, 2013, 41(8): 1586 – 1591. (in Chinese)

[15] 刘向宇,王斌,杨晓春. 社会网络数据发布隐私保护技术综述[J]. 软件学报,2014,25(3):576 – 590.
Liu Xiang-yu, Wang Bin, Yang Xiao-chun. Survey on privacy preserving techniques for publishing social network data[J]. Journal of Software, 2014, 25(3): 576 – 590. (in Chinese)

[16] Hay M, Miklau G, Jensen D, et al. Resisting structural re-identification in anonymized social networks[J]. The VLDB Journal, 2010, 19(6): 797 – 823.

[17] Skarkala M E, Maragoudakis M, Gritzalis S, et al. Privacy preservation by k-anonymization of weighted social networks[A]. Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining(ASONAM)[C]. Istanbul, Turkey: IEEE Computer Society, 2012. 423 – 428.

[18] Liu K, Terzi E. Towards identity anonymization on graphs[A]. 2008 ACM SIGMOD International Conference on Management of Data(SIGMOD'08)[C]. New York: ACM, 2008. 93 – 106.

[19] Yuan M X, Chen L, Yu P S, et al. Protecting sensitive labels in social network data anonymization[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(3): 633 – 647.

[20] Zhou B, Pei J. The K-anonymity and L-diversity approaches for privacy preservation in social networks against neighborhood attacks[J]. Knowledge and Information Systems, 2011, 28(1): 47 – 77.

[21] Zheleva E, Getoor L. Preserving the privacy of sensitive relationships in graph data[A]. 1st ACM SIGKDD Workshop on Privacy, Security, and Trust in KDD (PinKDD'07)[C]. San Jose, CA: ACM, 2007. 153 – 171.

[22] Zou L, Chen L, Özsu M T. K-automorphism: a general framework for privacy preserving network publication[J]. Proceedings of the VLDB Endowment, 2009, 2(1): 946 – 957.

作者简介

龚卫华 男,1977 年生于湖北武汉,博士,现为浙江工业大学计算机学院副教授. 主要研究方向:数据挖掘、社会网络、大数据计算等.
E-mail: whgong@sohu.com

兰雪锋 男,1990 年生于浙江丽水,浙江工业大学硕士生. 主要研究方向:社会网络、隐私保护.

裴小兵 男,1971 年生于湖北,博士,现为华中科技大学软件学院副教授. 主要研究方向:机器学习、数据挖掘、软件工程、电信网络管理.
E-mail: xiaobingp@hust.edu.cn

杨良怀 男,1967 年生于浙江新昌,博士,现为浙江工业大学计算机学院教授,主要研究方向:数据库系统、数据挖掘、大数据计算等.
E-mail: yanglh@zjut.edu.cn