基于向量相似的权重社会网络隐私保护

兰丽辉[1,2],鞠时光[1]
(1.江苏大学计算机科学与通信工程学院,江苏镇江 212013;2.沈阳大学信息工程学院,辽宁沈阳 110000)

摘 要: 针对权重社会网络发布,提出采用基于向量相似的随机扰动方法实现多个发布场景下网络结构和边权重的隐私保护.该方法以边空间理论为基础,采用基于节点聚类的分割方法构建权重社会网络的向量集模型;以加权欧氏距离作为向量相似的度量标准,根据选定阈值构建发布候选集;从候选集随机选取向量实现权重社会网络的发布;可抵御多种节点识别攻击,迫使攻击者在一个向量发生概率相同的庞大结果集中进行重识别,增加了识别的不确定性.实验结果表明,该方法在确保社会个体隐私安全同时可保护社会网络分析所需的某些结构特征,提高发布数据效用.

关键词: 社会网络; 边权重; 隐私保护; 向量集模型; 加权欧氏距离
中图分类号: TP309.2 **文献标识码:** A **文章编号:** 0372-2112 (2015)08-1568-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2015.08.015

Privacy Preserving Based on Vector Similarity for Weighted Social Networks

LAN Li-hui[1,2], JU Shi-guang[1]
(1. *School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang, Jiangsu* 212013, *China*;
2. *School of Information Engineering, Shenyang University, Shenyang, Liaoning* 110000, *China*)

Abstract: Aiming at the publication of weighted social networks, a random perturbation method based on vector similarity is proposed. It can protect network structures and edge weights in multiple release scenarios. It constructs vector set models by segmentation based on vertex cluster using edge space theory. It adopts weighted Euclidean distance as similarity metrics to construct the released candidate sets according to the threshold. It randomly selects vectors from candidate sets to construct the published weighted social networks. The proposed method can resist multiple vertex recognition attacks, force attackers to re-identify in a large result set that the existential probabilities of the vectors are same, and increase the uncertainty of recognition. The experimental results demonstrate that it can preserve individuals' privacy security, meanwhile it can protect some structure characteristics for networks analysis and improve data utility.

Key words: social networks; edge weight; privacy preserving; vector set model; weighted Euclidean distance

1 引言

社会网络是社会个体因互动而形成的网状关系结构,是多种社会现象的表示模型.由于科学研究、数据共享等需要,社会网络要进行发布,为确保社会个体的敏感信息不泄露,在数据发布前需进行隐私保护处理.已提出的隐私保护方法可分为基于聚类(泛化)和基于网络结构修改两类.结构修改方法相对聚类方法而言可保持社会网络的原有规模,数据缺损相对较小,可获得相对较高的数据效用.针对无权社会网络,基于 K-匿名模型,Zhou 等[1]提出采用 K-邻域匿名模型阻止攻击者以目标节点的邻域作为背景知识进行个体识别攻击;Liu 等[2]提出采用 K-度匿名模型阻止攻击者以目标节点的度作为背景知识进行个体识别攻击;Zou 等[3]提出采用 K-自同构方法抵御多种基于结构查询的个体识别攻击.Hay 等[4]通过等概率的删除 M 条边和插入 M 条的随机化构建方法实现隐私保护;Ying 等[5]采用和文献[4]相似的方法,通过随机增加边、删除边和交换边实现匿名保护.

针对权重社会网络,文献[6]提出采用高斯随机乘法扰动策略,通过添加噪声进行干扰实现边权重保护;文献[7]应用线性不等式系统捕获社会网络的特征参

收稿日期:2014-08-07;修回日期:2015-01-08;责任编辑:马兰英
基金项目:国家自然科学基金(No.61003288,61111130184);国家教育部博士点基金 (No.20093227110005);江苏省普通高校研究生科研创新计划项目(No.CX10B_006X)

数,提出采用线性规划技术实现边权重保护.文献[8]和文献[9]基于 K-匿名思想,提出采用 K-权重匿名实现节点和边权重保护.

本文在已有研究成果基础上,面向权重社会网络提出了一种基于向量相似的随机扰动隐私保护方法.

2　社会网络模型

2.1　发布场景

对社会网络发布进行隐私保护,先要明确发布场景的三个要素:攻击者的背景知识、发布数据用途和需要保护的隐私信息.

本文选取权重社会网络的节点(包括与节点相连边的权重)作为隐私信息,发布数据用于进行网络结构特征分析(重点关注平均路径长度、聚类系数),拟定攻击者分别拥有三种关于节点的背景知识(度、子图、边权重)情况下展开隐私保护研究.

2.2　图的边空间

文献[10]提出采用线性空间理论研究图,得到图的边空间定义如下:

定义 1(图的 m 维边空间).已知简单图 G,取数域 $P=\{0,1\}$,对 G 中的边从 1 到 m 进行编号,得到 $E(G)=\{e_1,e_2,\cdots,e_m\}$.将 G 的全体边子集构成的集合记作:$\psi(G)$,$\psi(G)$中每个元素是分量为 0 或 1 的 m 维向量,当边 $e_i(0<i\leqslant m)$在边子集中时,对应 m 维向量的第 i 分量为 1,否则为 0.将 $\psi(G)$在数域 P 上构成的 m 维线性空间称为 G 的边空间.

本文研究无向权重社会网络 $\boldsymbol{G}^S=(V,E,W)$,$V(\boldsymbol{G}^S)$是社会个体对应的节点集,$E(\boldsymbol{G}^S)$是表示个体间关系的边集,$W(\boldsymbol{G}^S)$是表示个体间连接强度的权重集.基于边空间理论,我们提出采用向量集作为 $\boldsymbol{G}^S$的发布模型.

2.3　向量集模型

2.3.1　社会网络分割

已知 $\boldsymbol{G}^S$,$|V(\boldsymbol{G}^S)|=N$,$|E(\boldsymbol{G}^S)|=M$,采用基于节点聚类的分割方法构建向量集模型,具体如下:

(1)基于共同邻居的数量进行节点聚类.节点相似性指标定义见式(1),$\Gamma(u)$和$\Gamma(v)$分别为节点 u 和 v 的邻居节点的集合.根据式(1)的计算结果,将 $V(\boldsymbol{G}^S)$划分为 d 个聚类集合:$\bigcup V_i(0<i\leqslant d)$(具体算法见 4.1 节).其中,集合 $V_1,V_2,\cdots,V_{d-1}$中节点的数量为 $t=\lfloor N/d\rfloor$ 个,集合 V_d中的节点数量大于等于 t.

$$S_{u,v}=|\Gamma(u)\bigcap\Gamma(v)|,u,v\in V(\boldsymbol{G}^S)\qquad(1)$$

(2)依据节点聚类的结果进行分组.若$|V_d|=t$,则从 $V_1,V_2,\cdots,V_{d-1},V_d$中各随机选取一个节点构成 t 个新的节点分组;若$|V_d|>t$,则从 $V_1,V_2,\cdots,V_{d-1}$中各随机选取一个节点,根据节点间的相似性从 V_d中选取 t 个节点随机划入 t 个分组,剩余$|V_d-t|$个节点不划入任何分组,新构成的 t 个分组中各包含 d 个节点,称 d 为分割参数.

(3)依据分组结果构建 $\boldsymbol{G}^S$的 t 个子图.将第 i 个分组中的节点及其该分组内节点间的边构建的子图记作:$\boldsymbol{G}^{Si}$,则可得到 $\boldsymbol{G}^S$ 的 t 个子图 G^{S1},G^{S2},…,G^{St},$|V(G^{Si})|=d$,$\bigcup G^{Si}\subset G^S(0<i\leqslant t)$;$G^S$经分割得到的 t 个子图的内部结构保持不变,子图间的结构保持不变.

(4)依据划分的子图结果构建 $\boldsymbol{G}^{SEr}$与 $\boldsymbol{G}^{SVr}$.$\boldsymbol{G}^{SEr}$表示 $\boldsymbol{G}^S$分割后未划入子图 $\boldsymbol{G}^{Si}(0<i\leqslant t)$中的边及其与这些边相关联的节点构成的子图;$\boldsymbol{G}^{SVr}$表示 $\boldsymbol{G}^S$分割后未划入子图 $\boldsymbol{G}^{Si}(0<i\leqslant t)$中的节点及其与这些节点相关联的边构成的子图.

2.3.2　向量集 $\boldsymbol{A}_S(\boldsymbol{G}^S)$的构建

本文提出的隐私保护方法主要针对 t 个子图展开,对 $\boldsymbol{G}^{SEr}$与 $\boldsymbol{G}^{SVr}$可采用已有的隐私保护方法进行处理(见 4.1 节).将 $\boldsymbol{G}^S$的向量集模型记作 $\boldsymbol{A}_S(\boldsymbol{G}^S)$,构建方法如下:

(1)构造一个包含 d 个节点的无向无权完全图 K_d,根据定义 1 构建 K_d的边空间 $\psi(K_d)$.

(2)构建 t 个子图在数域 P 上的向量.根据子图中节点在 $\boldsymbol{G}^S$中的编号从小到大重新由 1 到 d 进行编号,根据节点编号及 K_d中边的编号确定子图中边的编号,确保 K_d和子图中相同节点对间边的编号相同,$\boldsymbol{G}^{Si}$在 P 上的 k 维向量表示为$(\boldsymbol{e}_{i1},\boldsymbol{e}_{i2},\cdots,\boldsymbol{e}_{ik})$,$e_{ik}\in\{0,1\}$.

(3)构建 t 个子图在数域 R 上的向量.将子图的向量中分量为 1 的元素用对应边的权重值替换,则向量$(\boldsymbol{e}_{i1},\boldsymbol{e}_{i2},\cdots,\boldsymbol{e}_{ik})$中的元素取值为 0 或 $w_{ij}\in W(\boldsymbol{G}^S)(0<j\leqslant k)$,即当边 $\boldsymbol{e}_{ij}$在 $\boldsymbol{G}^{Si}$中时,对应 k 维向量的第 j 分量为 w_{ij},否则为 0.

(4)向量集 $\boldsymbol{A}_S(\boldsymbol{G}^S)$为 t 个子图的向量构成的集合,易知 $\boldsymbol{A}_S(\boldsymbol{G}^S)$中的向量维数 $k=d\times(d-1)/2$.

3　向量扰动方法

3.1　加权欧氏距离

将 $\boldsymbol{G}^S$的发布记作 $\boldsymbol{G}^P$,将 $\boldsymbol{A}_S(\boldsymbol{G}^S)$称为原始向量集,将 $\boldsymbol{A}_P(\boldsymbol{G}^P)$称为目标向量集,向量扰动利用向量相似和随机扰动相结合的方法构建 $\boldsymbol{A}_P(\boldsymbol{G}^P)$.通常采用欧氏距离度量两个向量的相似性,但欧式距离没有体现向量中每个元素的重要程度[11].我们提出采用加权欧氏距离作为度量 $\boldsymbol{A}_S(\boldsymbol{G}^S)$与 $\boldsymbol{A}_P(\boldsymbol{G}^P)$相似性标准,其定义见式(2).其中,$\boldsymbol{x}_i$和 $\boldsymbol{x}_j$表示维数为 η 的两个向量,ρ_η表示向量中元素的权重,$d(x_i,x_j)$表示加权欧氏距离.

$$d(\boldsymbol{x}_i,\boldsymbol{x}_j)=\Big(\sum_{\eta=1}^{l}\rho_\eta\,|\,x_{i\eta}-x_{j\eta}\,|^2\Big)^{1/2}\qquad(2)$$

本文采用边的介中性[12]作为向量中元素权重分配的依据,具体如下:

(1)计算子图中每条边的介中性.已知 $\boldsymbol{G}^S$ 的子图 $\boldsymbol{G}^{Si}$, $\forall\ e_{ij}\in E(\boldsymbol{G}^{Si})$, $0<i\leqslant t$, $0<j\leqslant M_i$ ($M_i=|E(\boldsymbol{G}^{Si})|$),边 e_{ij} 的介中性记作: b_{ij} ($b_{ij}>0$),得到 $\boldsymbol{G}^{Si}$ 中边的介中性集合,记作: $\{b_{i1},b_{i2},\cdots,b_{iMi}\}$.

(2)对非子图中边的介中性进行处理.若 $\boldsymbol{G}^{Si}$ 的向量 $(\boldsymbol{e}_{i1},\boldsymbol{e}_{i2},\cdots,\boldsymbol{e}_{ik})$ 中存在元素 $\boldsymbol{e}_{ij}=0(0<j\leqslant k)$,表明其对应的边 $e\notin E(\boldsymbol{G}^{Si})$ 但 $e\in E(K_d)$,则令 e 的边介中性取值为常数 $c(0<c\ll \mathrm{Min}(b_{ij}))$.

(3)采用归一化方法,对边介中性的取值进行处理.对经步骤(1)和步骤(2)计算得到的边介中性的正项序列 $b_{i1},b_{i2},\cdots,b_{ik}$ 进行变换,令: $c_{ij}=b_{ij}/\sum b_{ij}$,得到新序列 $\{c_{i1},c_{i2},\cdots,c_{ik}\}$, $c_{ij}\in[0,1]$, $\sum c_{ij}=1$, $1\leqslant j\leqslant k$.

(4)将集合 $\{c_{i1},c_{i2},\cdots,c_{ik}\}$ 中的元素作为 $\boldsymbol{G}^{Si}$ 的向量 $(\boldsymbol{e}_{i1},\boldsymbol{e}_{i2},\cdots,\boldsymbol{e}_{ik})$ 中对应元素的权重值,即 $\rho_\eta(e_{i\eta})=c_{i\eta}$ $(0<\eta<k)$.

3.2 候选向量集 CandSet($\boldsymbol{G}^S$)的构建

由于 $\boldsymbol{A}_S(\boldsymbol{G}^S)=\bigcup Vec(\boldsymbol{G}^{Si})(0<i\leqslant t)$ ($Vec(\boldsymbol{G}^{Si})$ 为 $\boldsymbol{G}^{Si}$ 的向量),候选向量集 $CandSet(\boldsymbol{G}^S)=\bigcup CandSet(\boldsymbol{G}^{Si})$. $\boldsymbol{G}^{Si}$ 的候选向量集 $CandSet(\boldsymbol{G}^{Si})$ 的构建步骤如下:

(1)根据 $\boldsymbol{G}^{Si}$ 中边权重的取值规律,在 $[Min(W(\boldsymbol{G}^{Si})), Max(W(\boldsymbol{G}^{Si}))]$ 范围内参照 $W(\boldsymbol{G}^{Si})$ 取值随机生成 k 个数值作为 K_d 中的边权重值,按 2.3.2 节向量集模型的构建得到 K_d 的向量集 $\boldsymbol{A}_{Si}(K_d)$.

(2)计算 $Vec(\boldsymbol{G}^{Si})$ 与 $\boldsymbol{A}_{Si}(K_d)$ 中向量的加权欧式距离.从 $\boldsymbol{A}_{Si}(K_d)$ 中随机选取一个向量,其对应的子图记作 S_j, $0<j\leqslant 2^k$,向量表示记作: $Vec(S_j)$,若 $E(\boldsymbol{G}^{Si})\neq E(S_j)$,则按式(2)计算 $Vec(\boldsymbol{G}^{Si})$ 与 $Vec(S_j)$ 的加权欧式距离,结果记作: $d(\boldsymbol{G}^{Si},S_j)$.

(3)设加权欧式距离阈值为 d_λ,构建向量集合 $CandSet(\boldsymbol{G}^{Si})$.若 $d(\boldsymbol{G}^{Si},S_j)\leqslant d_\lambda$,则将 $Vec(S_j)$ 加入候选集中.令 $\boldsymbol{A}_{Si}(K_d)\leftarrow\boldsymbol{A}_{Si}(K_d)-Vec(S_j)$,转至步骤(2)继续执行,直至 $\boldsymbol{A}_{Si}(K_d)=\varnothing$ 为止.

3.3 目标向量集 $\boldsymbol{A}_P(\boldsymbol{G}^P)$ 的构建

已知 $\boldsymbol{G}^S$ 的 t 个子图向量候选集 $\bigcup CandSet(\boldsymbol{G}^{Si})$ $(0<i\leqslant t)$,从每个子图的向量候选集中随机选取一个向量作为该子图的发布向量,将 $\boldsymbol{G}^{Si}$ 的发布向量记作: $PVec(\boldsymbol{G}^{Si})$,则 $\boldsymbol{A}_P(\boldsymbol{G}^P)=\bigcup PVec(\boldsymbol{G}^{Si})$ $(0<i\leqslant t)$.

3.4 性能分析

结论 1 设子图 $\boldsymbol{G}^{Si}$ 的候选集规模 $T_i=|CandSet(\boldsymbol{G}^{Si})|$,则 $\boldsymbol{G}^P$ 可能有 $\prod T_i(1\leqslant i\leqslant t)$ 种结果集.

证明 $\boldsymbol{A}_P(\boldsymbol{G}^P)$ 是由每个子图的候选向量集中的向量随机选取后组合而成.若 $\boldsymbol{G}^{Si}$ 的候选集中向量数量为 T_i,则每个子图对应的发布向量有 T_i 种可能,将每个子图的 T_i 个目标向量进行组合,则有 $T_1\times T_2\times\cdots\times T_i=\prod T_i(1\leqslant i\leqslant t)$ 种组合结果,而 $\prod T_i$ 种组合中的任何一种都可能成为 $\boldsymbol{A}_P(\boldsymbol{G}^P)$,使得 $\boldsymbol{G}^P$ 也有 $\prod T_i$ 种可能的结果集,故结论 1 成立.

设 $\Pr(x[y])$ 为在 $\boldsymbol{G}^P$ 中攻击者将节点 y 与目标节点 x 匹配的概率, $\Pr(\boldsymbol{G}^P)$ 为在可能的发布结果集中发布集为 $\boldsymbol{G}^P$ 的概率.

结论 2 $\Pr(x[y])$ 与 $\Pr(\boldsymbol{G}^P)$ 关系为: $\Pr(x[y])\propto\Pr(\boldsymbol{G}^P)$ ($\propto$ 为正相关).

证明 由结论 1 知攻击者在 $\prod T_i$ 种可能发布集中识别 $\boldsymbol{G}^P$ 的概率 $\Pr(\boldsymbol{G}^P)=1/(\prod T_i)$.设 $q=\prod T_i$,则可能发布集为 $\{\boldsymbol{G}^{P1},\boldsymbol{G}^{P2},\cdots,\boldsymbol{G}^{Pq}\}$.若攻击者尝试在 $\boldsymbol{G}^P$ 中识别目标节点 x,攻击者首先要根据自己掌握的关于目标节点的背景知识进行匹配候选集的构建,而与节点 x 相匹配的候选集是所有可能发布集中与其相匹配的节点 y 构成的集合.对于节点 x,设在 $\boldsymbol{G}^{Pi}$ 中与 x 匹配的节点集为 $Y_i(1\leqslant i\leqslant q)$,则节点 x 的完整匹配候选集的大小为 $|\bigcup Y_i|$,在精准计算的前提下,识别概率为 $\Pr(x[y])=1/|\bigcup Y_i|$.实际上在匹配候选集中必然存在一定数量的伪节点,因此最终的识别概率要低于 $1/|\bigcup Y_i|$.易知,随着 q 的增加和减少, $\Pr(x[y])$ 与 $\Pr(\boldsymbol{G}^P)$ 保持相同的变化趋势,故结论 2 成立.

向量扰动方法对社会网络分割时,在节点聚类过程中采用了基于共同邻居的贪心选择算法,而在具体分割时将每一个聚类集中的节点分散于不同的子图.具有共同邻居的节点更倾向于互相连接[13],在聚类时同一组内的节点连接概率较高,使得这些节点分散于不同的子图时,子图间各节点的连接概率降低.比较而言社会网络经分割后得到的子图都属于稀疏图.向量扰动方法主要是对子图结构进行变换,同时保证了变换后的子图与原图具有一定的相似性,仅通过对少量边的修改达到隐私保护的目的,最大限度提高了发布数据效用.

4 隐私保护算法

将本文提出的权重社会网络的隐私保护算法记作 RPVS(Random Perturbation Based on Vector Similarity).

4.1 算法实现

算法 1: RPVS

输入:社会网络 $\boldsymbol{G}^S$、分割参数 d、距离阈值 d_λ

输出:社会网络 $\boldsymbol{G}^S$ 的发布——$\boldsymbol{G}^P$

步骤:

Step1 对 G^S进行去标识处理,对节点从 1 至 N 进行标号;

Step2 计算 $V(G^S)$ 中的节点的邻居节点及节点间共同邻居;

Step3 选取共同邻居数量最多且共同邻居数量相差最小的前 t 个节点构建聚类$\cup Clu_j(1\leqslant j\leqslant d)$,若同等条件下的相似节点数量超过 t 个,则应尽量保证剩余节点的相似性最大;

Setp4 根据聚类$\cup Clu_j$构建 t 个分割子图,根据 E_r构建 G^{SEr},根据 V_r构建 G^{SVr};

Step5 按 2.3.2 节的方法构建向量集 $A_S(G^S)$;

Step6 采用 Newman 在文献[14]中提出的基于广度优先搜索的算法计算每个子图中边的介中性;

Step7 按 3.1 节的方法进行子图向量元素的权重分配;

Step8 根据式(2)构建每个子图的候选向量集 $CandSet(G^{Si})(1\leqslant i\leqslant t)$;

Step9 由子图的候选向量集$\cup CandSet(G^{Si})$构建 G^S的候选向量集 $CandSet(G^S)$;

Step10 从 t 个子图的候选向量集各随机选取一个向量构建 $PVec(G^{Si})$;

Step11 $A_P(G^P)\leftarrow\cup PVec(G^{Si})$,返回目标向量集 $A_P(G^P)$;

Step12 按文献[1]中的方法对未划入子图的节点进行同构处理,根据结果构建 G^{PEr};

Step13 按文献[8]中的方法对未划入子图的边进行 d-权重匿名处理,根据结果构建 G^{PVr};

Step14 基于 $PVec(G^{Si})$构建发布子图 G^{Pi};

Step15 构建发布网络 $G^P\leftarrow(\cup G^{Pi})\cup G^{PEr}\cup G^{PVr}$.

4.2 算法复杂度分析

由前述内容可知,RPVS 算法的主要时间复杂度为边的介中性计算和子图候选向量集的构建.实际上由于子图是稀疏图,计算介中性的边的数量相对整个网络的边集而言较少,但在最坏情况下边的介中性计算时间复杂度为 $O(td^3)$[14],而子图向量候选集的构建时间复杂度为 $O(2^m)$.由 4.1 节的算法描述可知,RPVS 算法时间复杂度约为 $O(N\mu+td^3+2^m+M)$,其中 μ 为网络的平均度值.当分割参数取值较小,即 $N\mu+M>>2^m$,时间复杂度约为 $O(N\mu+M)$;当分割参数取值较大,即 $2^m>>N\mu+M$,时间复杂度约为 $O(2^m)$.

5 实验

5.1 实验环境与数据

实验环境:Intel 酷睿 i3-3240 @ 3.40GHz 双核,4.00GB 内存,操作系统为 Microsoft Windows 7,编程语言为 C++与 MATLAB.

实验数据[15]:Karate 网络(34 个节点,78 条边),Lesmis 网络(77 个节点,254 条边),Prefuse 网络(129 个节点,161 条边),PowerGrid 网络(4941 个节点,6594 条边).其中,Prefuse 和 PowerGrid 公开发布的数据集是无权网络,为了实现权重值的隐私保护测试,对于 Prefuse 和 PowerGrid 的边权重采用随机数生成器生成区间分别在[1,50]和[1,100]的随机整数作为权重值.

5.2 实验结果及分析

实验选取文献[1]、文献[2]、文献[3]、文献[4]、文献[8]、文献[9]中的算法与 RPVS 算法进行比较,将上述文献中的算法依序记作 KN、KD、KM 、RGP、KEW 和 KW.

参与比较的六个算法均通过对社会网络的结构进行修改实现隐私保护,与 RPVS 算法属于同一类方法.其中,KN、KD、KM、KW 和 KEW 算法基于 K-匿名模型设计,针对某一确定发布场景,通过结构修改使得网络中的任意节点至少与其他 $K-1$ 个节点不可区分,取 K 为其性能参数;RGP 算法采取随机插入、删除相同数量边的策略实现网络结构地修改,可抵御基于子图查询地攻击,取干扰比率 γ 为其性能参数;RPVS 算法中设子图的向量候选集大小为 β,d_λ为 1 可满足 β 的取值要求.由于数据集规模不同,在隐私保护质量(5.2.2 节)和数据效用(5.2.3 节)测试实验,设定 Karate 和 Lesmis 网络分割参数 $d=5$,Prefuse 网络分割参数 $d=6$,PowerGrid 网络分割参数 $d=7$.

5.2.1 算法执行效率

本实验对 RPVS 算法的执行效率进行测试.由图 1 的实验结果可知,d_λ确定时,随着 d 取值增加,算法执行时间迅速上升;d 确定时,随着 d_λ增加,算法执行时间变化不大.

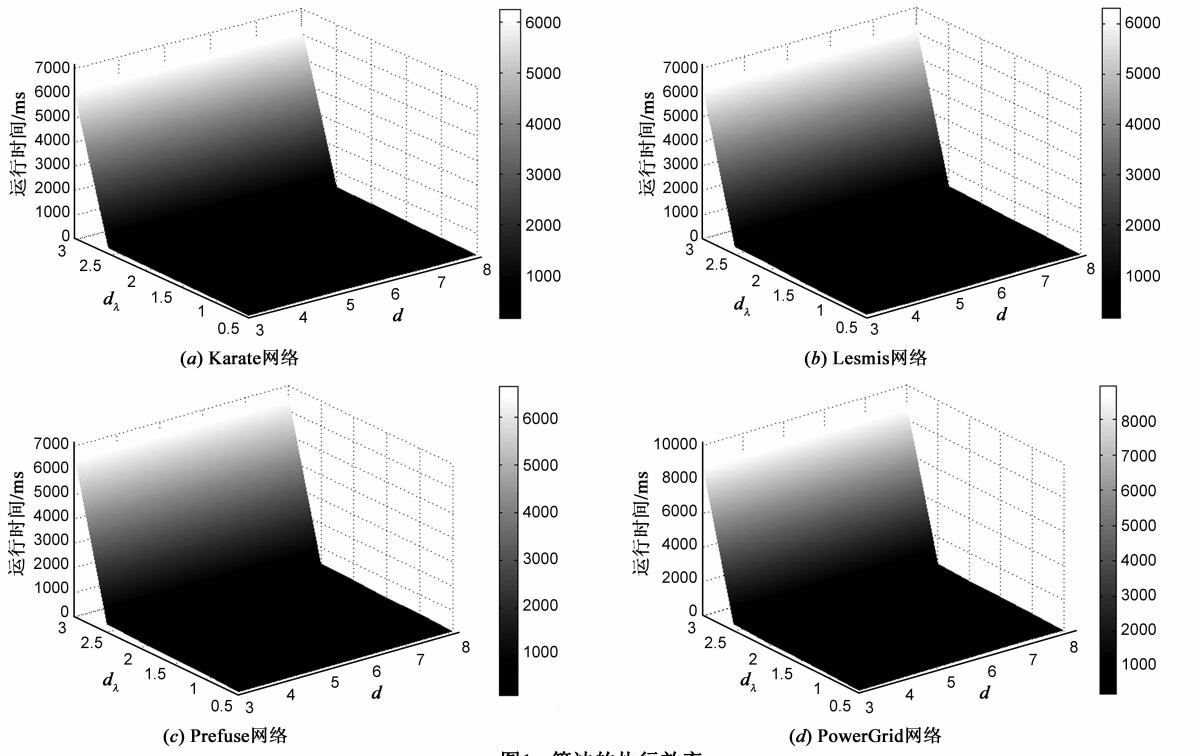
5.2.2 隐私保护质量

对基于度的节点识别攻击,选取 KD 算法进行对比.RPVS 算法随机选取 $5\times\beta$ 个结果集中的匹配节点构建的集合作为目标节点的匹配候选集,实验以平均匹配候选集的大小作为衡量标准,结果如图 2 所示.

对基于子图的节点识别攻击,选取 KN、KM 和 RGP 算法进行对比,各算法参数缺省值为 10.RPVS 算法在 Karate 和 Lesmis 网络中选取 10 个、Prefuse 和 PowerGrid 网络中选取 20 个可能结果集中的匹配子图构建匹配候选集,结果如图 3 所示.

对基于权重的节点识别攻击,选取 KW 和 KEW 算法进行对比,与基于度的节点识别攻击采用相同的实验方案,结果如图 4 所示.

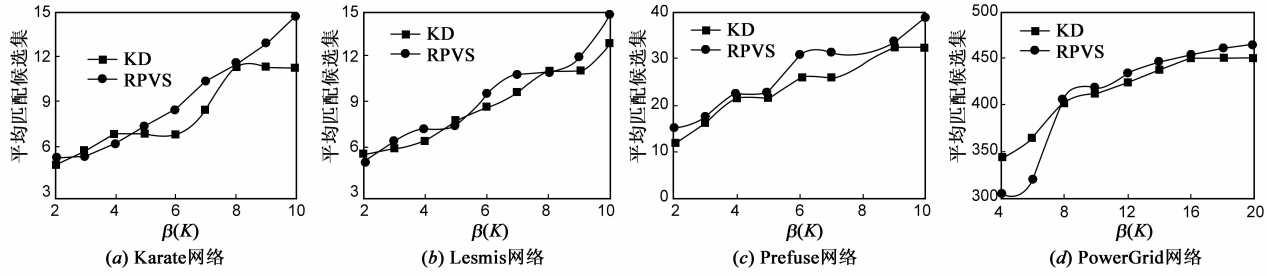
RPVS 算法使得攻击者在进行识别攻击时要考虑所有可能的发布结果,识别概率为 $1/|\cup Y_i|$.上述实验中 Y_i的取值为β,且 $\beta=K$,显然$|\cup Y_i|>K$.由上述实验结果可知,RPVS 算法在多数情况下其节点识别概率要低于同类方法.



(*a*) Karate网络 (*b*) Lesmis网络

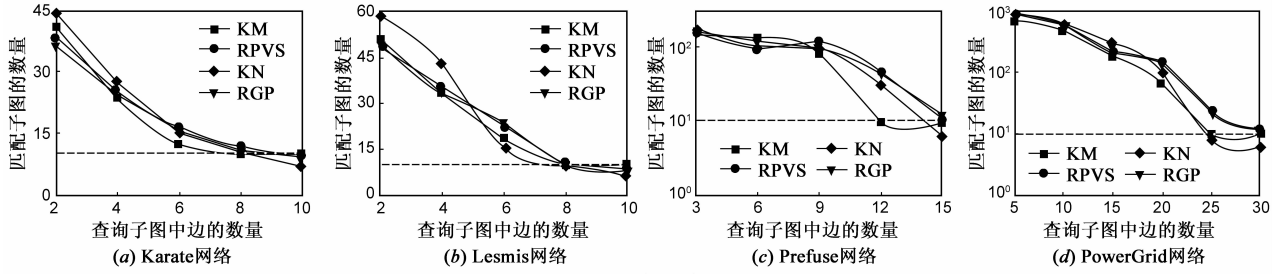
(*c*) Prefuse网络 (*d*) PowerGrid网络

图1 算法的执行效率



(*a*) Karate网络 (*b*) Lesmis网络 (*c*) Prefuse网络 (*d*) PowerGrid网络

图2 基于度的节点识别测试



(*a*) Karate网络 (*b*) Lesmis网络 (*c*) Prefuse网络 (*d*) PowerGrid网络

图3 基于子图的节点识别测试

5.2.3 数据效用

实验选取 KN、KD、KM 和 RGP 算法进行平均最短路径与平均聚类系数测试,选取 KW 算法进行节点对间最短距离测试,参数设置同隐私保护质量实验.对于最短路径的测试,随机选取 50% 节点对,取 10 个可能结果集相对误差绝对值的平均值作为度量标准.实验结果如图 5、图 6 和图 7 所示.

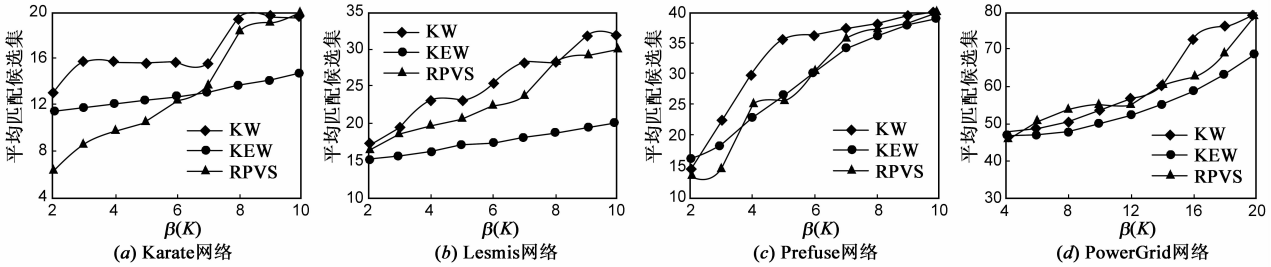


图4　基于权重的节点识别测试

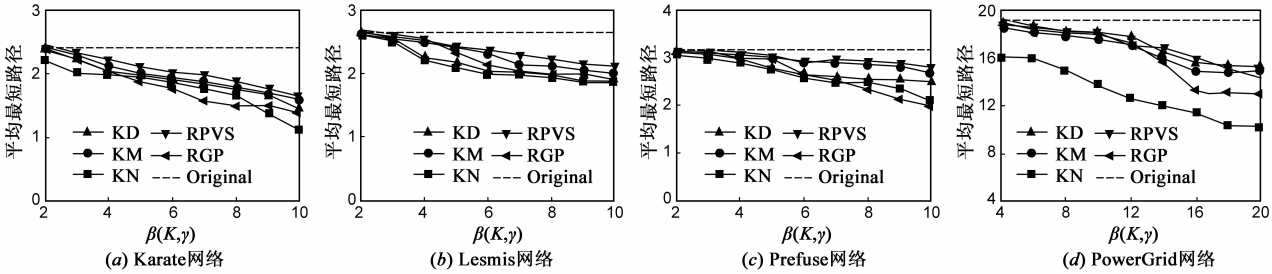


图5　平均最短路径

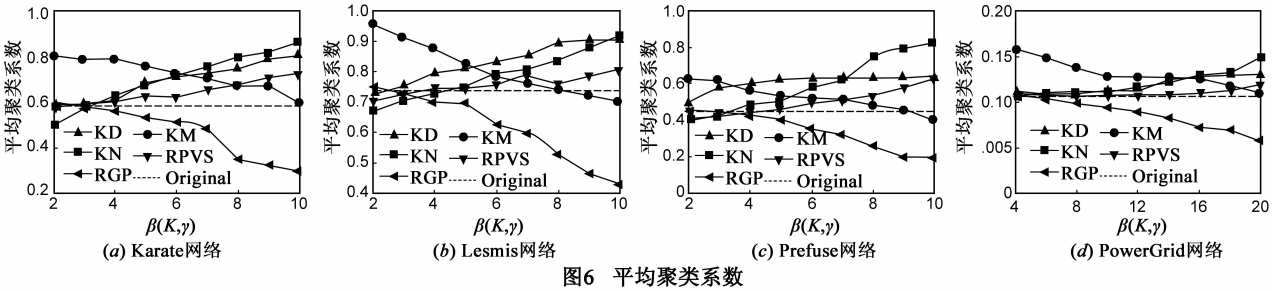


图6　平均聚类系数

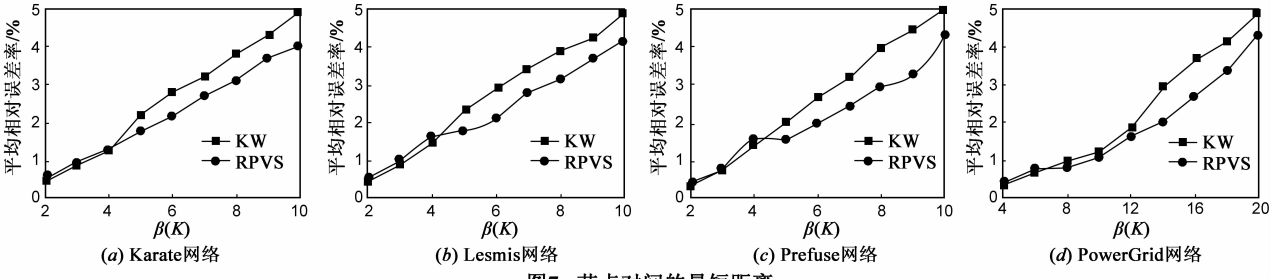


图7　节点对间的最短距离

由实验结果可知,尽管 RPVS 算法不能在 β 的任意取值都保持最好的数据效用,但多数情况下 RPVS 算法的发布性能要高于同类方法,可获得相对较高的效用.

6　结束语

本文针对权重社会网络发布,提出采用向量扰动方法解决多个发布场景下社会个体的隐私保护,与已有隐私保护方法有较大不同.首先,该方法不仅考虑了社会网络结构的隐私保护,还考虑了边权重的隐私保护;其次,该方法可抵御多种隐私攻击,适用于多个发布场景;再次,该方法采用了加权欧式距离作为向量相似的度量标准,提高了子图候选向量集中向量的相似性,进而最大限度的提高了发布社会网络与原始社会网络的相似性;最后,该方法将向量相似与随机扰动技术相结合,在保证社会个体隐私保护质量的基础上,提高了发布数据效用.

参考文献

[1] Zhou B, Pei J. Preserving privacy in social networks against neighborhood attacks[A]. 2008 IEEE 24th International Confer-

ence on Data Engineering[C]. Cancun, Mexico: IEEE Computer Society, 2008. 506 – 515.

[2] Liu K, Terzi E. Towards identity anonymization on graphs[A]. 2008 ACM SIGMOD International Conference on Management of Data[C]. Vancouver, Canada: Association for Computing Machinery, 2008. 93 – 106.

[3] 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.

[4] Hay M, Mikiau G, Jensen D, Weis P, et al. Anonymizing social networks[R]. University of Massachusetts Amherst, 2007.

[5] Ying X W, Wu X T. Randomizing social networks: A spectrum preserving approach[A]. 2008 8th SIAM International Conference on Data Mining[C]. Atlanta, United States: Society for Industrial and Applied Mathematics Publications, 2008. 739 – 750.

[6] Liu L, Wang J, Liu J et al. Privacy preserving in social networks against sensitive edge disclosure[R]. Department of Computer Science, University of Kentucky, 2008.

[7] Das S, Egecioglu Ö, Abbadi A E. Anónimos: An LP-based approach for anonymizing edge-weighted social network graphs [J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 4(4): 590 – 603.

[8] Li Y, Shen H. Anonymizinggraphs against weight-based attacks [A]. 2010 10th IEEE International Conference on Data Mining Workshops[C]. Sydney, Australia: IEEE, 2010. 491 – 498.

[9] Skarkala M E, Maragoudakis M, Gritzalis S et al. Privacyp-reservation by k-anonymization of weighted social networks [A]. 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining[C]. Istanbul, Turkey: IEEE Computer Society, 2012. 423 – 428.

[10] 王树禾. 图论(第二版)[M]. 北京: 科学出版社, 2009: 177 – 180.

[11] 吕锋, 杜妮, 文成林. 一种模糊-证据 kNN 分类方法[J]. 电子学报, 2012, 40(12): 2390 – 2395.
Lü Feng, Du Ni, Wen Cheng-lin. A fuzzy-evidential k nearest neighbor classification algorithm[J]. Acta Electronica Sinica, 2012, 40(12): 2390 – 2395. (in Chinese)

[12] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. Physical Review E, 2004, 69(2): 1 – 15.

[13] Lü L, Zhou T. Link prediction in complex networks: A survey [J]. Physica A: Statistical Mechanics and its Applications, 2011, 390(6): 1150 – 1170.

[14] Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences, 2002, 99(12): 7821 – 7826.

[15] Nexus. The network repository [DB/OL]. http://nexus.i-graph.org/.

作者简介

兰丽辉(通信作者) 女, 1976 年出生, 吉林乾安人, 沈阳大学信息工程学院副教授, 江苏大学计算机应用技术专业博士生, 主要研究方向为信息安全、隐私保护.
E-mail: syu_lanlihui@syu.edu.cn

鞠时光 男, 1955 年出生, 江苏南通人, 博士, CCF 高级会员, 江苏大学计算机科学与通信工程学院教授、博士生导师, 主要研究方向为空间数据库、信息安全理论与技术.
E-mail: jushig@ ujs.edu.cn