一种基于GSNPP算法的社交网络隐私保护方法研究

韦 伟 李 杨 张为群

(西南大学计算机与信息科学学院 重庆 400715) (重庆市智能软件与软件工程重点实验室 重庆 400715)

摘 要 随着网络信息技术的快速发展,社交网络迅速涌现。针对社交网络隐私保护问题,提出了一种基于GSNPP算法的隐私保护方法。它通过对社交网络中节点进行聚类,再对生成的簇进行簇内泛化及簇间泛化,来对社交网络进行匿名化处理,拟达到隐私保护的目的;同时量化了社交网络匿名化处理过程中所带来的不同类型信息的丢失。最后通过实验验证了该方法的可行性和有效性。

关键词 社交网络,隐私保护,聚类方法,k-匿名

Study on GSNPP Algorithm Based Privacy-preserving Approach in Social Networks

WEI Wei LI Yang ZHANG Wei-qun

(Faculty of Computer and Information Science, Southwest University, Chongqing 400715, China)

(Chongqing Intelligent Software and Software Engineering Laboratory, Chongqing 400715, China)

Abstract With the rapid development of network information technology, social networks have sprung up. For the protection of their privacy, we proposed a GSNPP algorithm-based privacy-preserving approach. By clustering the nodes in social networks, and then making generalization not only within cluster but also among clusters, the approach processes the social networks anonymously, to achieve the purpose of privacy protection, then quantifies the information loss of different types in the process of the anonymity of social networks. At last, the results of experiment demonstrate the feasibility and validity of the approach.

Keywords Social networks, Privacy preserve, Cluster approach, k-anonymity

1 引言

近年来,随着网络信息技术的蓬勃发展,互联网已经渗透到了人们生活以及生产的方方面面。尤其是Web2.0的迅速兴起,使得各种社交网站呈现出爆炸式的发展态势,但在使用社交网络的同时,也给广大的用户群体带来了严重的隐私安全问题[6,12]。目前,社交网络数据在隐私保护方面的研究还处于起步阶段,其主要通过修正图的方法,防止恶意者通过知道节点的度而重新识别出个体的身份[2,3]。该方法需要在社交网络图中增加或删除边来满足k-度的条件,代价比较大,匿名化后数据的有效性较低。本文提出了一种基于GSNPP(Greedy for Social Network Privacy-Preserving)算法的社交网络隐私保护方法。该方法通过GSNPP算法对原社交网络中节点进行聚类[11],生成多个簇,再通过簇内泛化及簇间泛化,对社交网络进行匿名化处理,量化社交网络匿名化处理过程中所带来的不同类型信息的丢失,使得该网络具有较好的隐私保护效果,平衡属性信息与结构信息的丢失,并且尽可能降低其信息丢失量,以提高数据有效性。

2 社交网络隐私模型

2.1 相关定义

定义1(社交网络) 令无向图$G=(V,E)$为社交网络[1],其中V为社交网络中节点的集合,每个节点表示一个实体;$E\subseteq V\times V$为社交网络中边的集合,每条边表示两个实体之间的关系,所有关系为相同类型。

定义2(簇内准标识符信息泛化) 令$clt=\{v_1,v_2,\cdots,v_\mu\}$为节点聚类后生成的一个簇,簇中节点准标识符属性$Q=Q_N\cup Q_C$,$Q_N=\{n_1,n_2,\cdots,n_s\}$为准标识符数值属性集合,$v_1(n_j)$为某一节点数值属性值,数值属性$n_j\in Q$泛化后为$Ge(n_j)=[\min\{v_1(n_j),v_2(n_j),\cdots,v_\mu(n_j)\},\max\{v_1(n_j),v_2(n_j),\cdots,v_\mu(n_j)\}]$;$Q_C=\{c_1,c_2,\cdots,c_t\}$为准标识符类属性集合,$v_1(c_j)$为某一节点类属性值,对类属性$c_j\in Q$泛化后,$Ge(c_j)$为$\{v_1(c_j),v_2(c_j),\cdots,v_\mu(c_j)\}$生成的泛化层$H_{c_j}$的最近公共祖先[8]。

定义3(匿名化社交网络) 给定一个社交网络$G=(V,E)$,以及对节点V进行聚类生成的簇集合$S_{dt}=\{clt_1,clt_2,\cdots,clt_m\}$,$\bigcup_j^m clt_j=V$,$clt_i\cap clt_j=\emptyset$,$i,j\in 1,2,\cdots,m$,$i\neq j$;生成的匿名化社交网络图$G^A=(V^A,E^A)$,$V^A=\{v_{dt_1},v_{dt_2},\cdots,v_{dt_m}\}$,其中$v_{dt_i}$为簇$clt_i\in S$泛化后生成的一个匿名节点,$v_{dt_j}=[Ge(Q),(|clt_j|,|E_{dt_j}|)]$;$E^A\subseteq V^A\times V^A$,$\forall v_{dt_i},v_{dt_j}\in V^A$,且$\exists v_1\in clt_i,v_2\in clt_j,(v_1,v_2)\in E$,则$(v_{dt_i},v_{dt_j})\in E^A$,表示为$|E_{clt_i,clt_j}|$。

定义4(k-匿名社交网络) 对$\forall j=1,\cdots,m$,满足$|clt_j|\geq k$,

到稿日期:2011-04-03 返修日期:2011-07-01 本文受重庆市信息产业发展资金项目(200921011)资助。

韦 伟(1983-),男,硕士生,主要研究方向为软件工程、信息安全等,E-mail:kobe198308@gmail.com;李 杨(1987-),女,硕士生,主要研究方向为信息安全等;张为群(1950-),男,教授,主要研究方向为软件工程、形式语言、信息安全等。

则 $G^A=(V^A,E^A)$ 是 k-匿名社交网络[10]，其中 $V^A=\{v_{clt_1},v_{clt_2},\cdots,v_{clt_m}\}$，$v_{clt_j}=[Ge(Q),(|clt_j|,|E_{clt_j}|)]$，$j=1,\cdots,m$。

定义 5(节点相异度) 给定图 $G(V,E)$，节点 $V=\{v_1,v_2,\cdots,v_n\}$，每个节点 v_i 的邻居节点表示为 n-维布尔向量 $B_i=(b_i^1,b_i^2,\cdots,b_i^n)$，当 $\exists b_i^j$ 为 1 时，则 v_i,v_j 之间存在一条边 $(v_i,v_j)\in E,\forall i=1\cdots n;i\neq j$；反之为 0。则

两节点 v_i,v_j 之间的相异度定义为[5]

$$d(v_i,v_j)=\frac{|\{\rho|\rho=1\cdots n\wedge\rho\neq i,j;b_i^\rho\neq b_j^\rho\}|}{n-2} \tag{1}$$

由式(1)得节点 v 与簇 clt 之间的相异度为

$$d(v,clt)=\sum_{v_j\in clt}d(v,v_j)/|clt| \tag{2}$$

2.2 隐私保护模型

社交网络数据发布后，其标识符属性会被移除，但攻击者仍可能通过记录关联技术，将准标识符与外部信息进行匹配，从而推理出用户身份，导致用户隐私泄漏。因此，本文主要考虑准标识符属性及准标识符属性关系。可使用如下隐私模型对社交网络进行匿名化处理。

令 n 为节点集合 N 中的节点数，使用聚类方法将节点划分为互不相交的 m 个簇，即 $clt_1,clt_2,\cdots,clt_m$。

簇内泛化：$\exists clt=\{v_1,v_2,\cdots,v_\mu\}$，$Q=\{n_1,n_2,\cdots,n_s,c_1,c_2,\cdots,c_t\}$，则 $Ge(E_{clt})(|clt|,|E_{clt}|)$，$Ge(clt)=\{Ge(c),Ge(n)\}\Rightarrow$一个匿名节点 v_{clt}。

簇间泛化：$\exists clt_i,clt_j,e$，若 $e\in E_{clt_i,clt_j}$ 且 $e\in E,e\in clt_i\times clt_j$，则 $Ge(E_{clt_i,clt_j})=|E_{clt_i,clt_j}|\Rightarrow$一条泛化边 E_{clt_i,clt_j}。

匿名化社交网络图：给定 $G^A=(V^A,E^A)$，对 $\forall j=1,\cdots,m$，满足 $|clt_j|\geqslant k$，$V^A=\{v_{clt_1},v_{clt_2},\cdots,v_{clt_m}\}$，$E^A\subseteq V^A\times V^A\Rightarrow$一个 k-匿名的社交网络图 G^A。

3 GSNPP 算法

GSNPP 算法表现为一个贪心算法，通过贪心聚类的过程对节点进行选取，来生成一个 k-匿名的社交网络。按照如下 3 个条件进行节点的选择：簇中节点相异度应该最小；每个簇至少应包含 k 个节点；所选节点带来的泛化信息丢失应是最少的。同时，为了获得较高的数据有效性，分别量化分析两种信息丢失量，即数据信息丢失量以及结构信息丢失量。

3.1 量化信息丢失量

3.1.1 数据信息丢失量

令 clt 为一个簇，$Q=\{n_1,n_2,\cdots,n_s,c_1,c_2,\cdots,c_t\}$ 为准标识符集，$Ge(Q)$ 为准标识符泛化集。泛化准标识符属性所产生的数据信息丢失量[8]为

$$QL(clt)=|clt|\left(\sum_{j=1}^{s}\frac{size(Ge(n_j))}{size(\min_{v\in N}\{v(n_j)\},\max_{v\in N}\{v(n_j)\})}+\sum_{j=1}^{t}\frac{height(\Lambda(Ge(c_j)))}{height(H_{c_j})}\right) \tag{3}$$

式中，$|clt|$ 为簇 clt 中节点数量；$size([i_1,i_2])$ 为 $[i_1,i_2]$ 的区间大小，即 i_2-i_1；$\Lambda(\nu),\nu\in H_{C_j}$ 为泛化层；$height(H_{C_j})$ 为泛化层 H_{C_j} 的高度。则其总数据信息丢失量定义为

$$TQL(G)=\sum_{j=1}^{m}QL(clt_j) \tag{4}$$

对式(4)进行归一化处理，得

$$NTQL(G)=TQL(G)/n\cdot(s+t) \tag{5}$$

3.1.2 结构信息丢失量

令 clt 为对节点 V 聚类后所生成的一个簇，$G_{clt}=(V_{clt},E_{clt})$ 为 $G=(V,E)$ 的一个子图。当簇 $clt\in S$ 泛化后生成一个匿名节点 v_{clt}，簇中任意两个节点之间存在一条边的概率为 $|E_{clt}|/\binom{|clt|}{2}$，那么将边存在标识为边不存在的概率为 $1-|E_{clt}|/\binom{|clt|}{2}$。

簇内结构信息丢失量 $INA\text{-}SL(clt)$ 为错误地将簇中任意两个节点之间标识为边存在或边不存在的概率。令簇 clt 内含 $|E_{clt}|$ 条边，则存在 $2\cdot\left(\binom{|clt|}{2}-|E_{clt}|\right)$ 个未连接的节点，则

$$INA\text{-}SL(clt)=2\cdot|E_{clt}|\cdot\left(1-|E_{clt}|/\binom{|clt|}{2}\right) \tag{6}$$

簇间结构信息丢失量 $INR\text{-}SL(clt)$ 为错误地将簇 clt_i，clt_j 间任意两个节点 (v_i,v_j) 之间标识为边存在或边不存在的概率，其中 $v_i\in clt_1,v_j\in clt_2$。令簇 clt_i,clt_j 之间存在 $|E_{clt_i,clt_j}|$ 条边，则存在 $2\cdot(|clt_i|\cdot|clt_j|-|E_{clt_i,clt_j}|)$ 个未连接的节点，则

$$INR\text{-}SL(clt_i,clt_j)=2\cdot|E_{clt_i,clt_j}|\cdot\left(1-\frac{|E_{clt_i,clt_j}|}{|clt_i|\cdot|clt_j|}\right) \tag{7}$$

由式(6)、式(7)得其总结构信息丢失量为

$$TSL(G)=\sum_{j=1}^{m}INA\text{-}SL(clt_j)+\sum_{i}^{m}\sum_{j=i+1}^{m}(INR\text{-}SL(clt_i,clt_j)) \tag{8}$$

对式(7)进行归一化处理，得

$$NTSL(G)=TSL(G)/(n\cdot(n-1)/4) \tag{9}$$

3.2 生成簇集合

基于以上的隐私保护模型及信息丢失量化方法，提出了基于 GSNPP 算法的簇生成方法。该算法一次只生成一个簇，只向簇中添加一个节点，当前簇中节点达到 k 时，再生成新的簇；如图中无节点，则算法结束。其流程如图 1 所示。

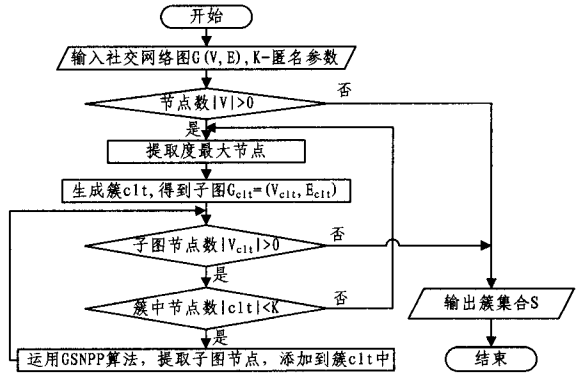

图 1 簇生成流程

对社交网络 $G=(V,E)$ 进行匿名化处理，首先选择网络中度最大的节点，生成簇，通过计算节点相异度及信息丢失量，选择节点，得到满足 k-匿名条件的簇集合 $S=\{clt_1,clt_2,\cdots,clt_m\}$。引入权重参数 $\alpha\in[0,1]$，来控制不同类型信息的丢失。其流程如下所示。

输入：社交网络图 $G=(V,E)$，k-匿名参数，权重参数 α

输出：簇集合 S

Step1 判断 G 中节点数，若 $|V|>0$，执行 Step2；否则，转至 Step9。

Step2 提取 V 中度最大节点 v_s。

Step3 生成簇 clt_j，得到子图 $G_{clt}=(V_{clt},E_{clt})$。

Step4　判断子图 G_{clt} 中节点数，若 $|V_{clt}|=0$，转至 Step9；否则，执行 Step5。

Step5　若簇 clt_j 中节点数 $|clt_j|\geqslant k$，满足 k-匿名条件，转至 Step2；否则，执行 Step6。

Step6　根据权重参数 α，由式(2)、式(5)计算要添加到簇中节点 v_i 的信息丢失量，即

$v_i.loss=\alpha\cdot NTQL(G_{clt})+(1-\alpha)\cdot(d(v,clt_j)),1\leqslant i\leqslant n$。

Step7　若 $v_i.loss>v_{i+1}.loss$，则转至 Step6；否则，执行 Step8。

Step8　向簇 clt_j 中添加节点 v_i，并执行 Step4。

Step9　输出簇集合 $S=\{clt_1,clt_2,\cdots,clt_m\}$。

当生成最后一个簇时，其簇内节点数可能小于 k，要对簇进行拆分，即将簇内节点添加到其他节点数大于等于 k 的簇中，按照如下流程进行拆分和添加。

```
While |clt|<k do
  S=S-{clt};
  For each v∈clt do
    clt_cand=null;v'.loss=∞;
      For each clt_i∈S do
        If
          α·NTQL(G_1)+(1-α)(d(v,clt_i))<v'.loss then
          v'.loss=α·NTQL(G_clt)+(1-α)(d(v,clt_i));
          clt_cand=clt_i;
        End if
      End for
    Return clt_cand
  End for
End while
```

4　模拟实验

本文采用美国加州大学欧文分校机器学习库(UC Irvine Machine Learning Repository)中的 Adult 数据库[9]进行模拟实验。该数据库是研究隐私保护匿名算法所经常用到的一个数据库，数据库大小为 3.9M。随机从数据库中抽取 500 条记录作为社交网络节点，将{Age, workclass, education, race, sex, native-county}作为准标识符属性，并使用 R-MAT graph model 以及 GTgraph suite 中的图生成器来生成一个满足幂律分布及小世界特征[7]的边集合，以此来模拟一个社交网络，其节点的平均度为 10。与文献[4]中的算法进行对比，其结果如图 2 和图 3 所示。

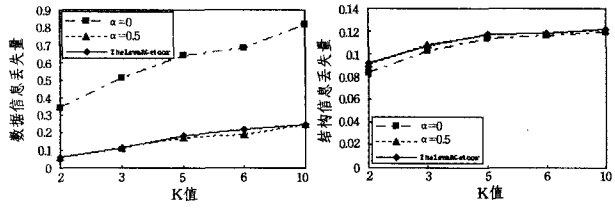

图 2　信息丢失量与 k 的关系图

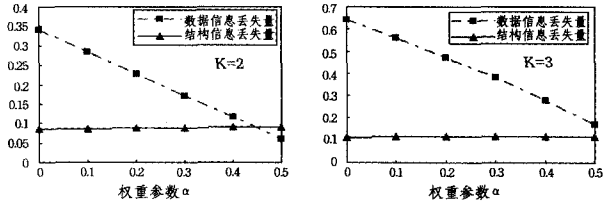

图 3　信息丢失量与 α 的关系图

通过此实验，可以看出：1)图 2 中，信息丢失量随 k 的增大而增加，其中当 $\alpha=0.5$ 时，数据信息丢失量略小于文献[4]中的算法；结构信息丢失量均小于文献[4]中的算法。2)图 3 中，当 k 一定时，数据信息丢失量随 α 的增大而减少，结构信息丢失量随 α 的增大而增大。3)社交网络数据发布者可以通过设置权重参数 α，来自定义其可以承受的数据信息丢失量及结构信息丢失量，即当 $\alpha\in(0,0.5)$ 时，可以保护更多的数据信息；当 $\alpha\in(0.5,1)$ 时，可以保护更多的结构信息。

结束语　本文提出的基于 GSNPP 算法的社交网络隐私保护方法，对社交网络中个体进行聚类分析，并且将数据信息与结构信息进行泛化处理，量化了两种信息的丢失量；同时引入了一个权重参数，有利于数据发布者平衡数据信息与结构信息的丢失，也有利于在实现隐私保护的同时，提高数据的有效性。下一步的研究工作还包括：(1)当图中加入边的权重时，如何量化其信息丢失量；(2)如何扩展现有的社交网络的隐私保护模型。

参 考 文 献

[1] Hay M, Miklau G, Jensen D, et al. Anonymizing social networks [R]. 07-19. University of Massachusetts Amherst, 2007

[2] Liu K, Terzi E. Towards identity anonymization on graphs [C]// Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data (SIGMOD'08). New York, NY, USA, ACM Press, 2008: 93

[3] Liu Kun, Das K, Grandison T, et al. Privacy-preserving data analysis on graphs and social networks[C]//Kargupta H, Han J, Yu P, et al., eds. Nexteneration Data Mining. CRC Press, 2008

[4] Zheleva E, Getoor L. Preserving the privacy of sensitive relationships in graph data[C]// Proceedings of the 1st ACM SIGKDD Workshop on Privacy, Security, and Trust in KDD (PinKDD'07). 2007: 153-171

[5] Han Jia-wei, Kamber M, Data Ming. Concepts and Techniques (第二版)[M]. 范明，阵小峰，译. 北京：机械工业出版社，2007: 255-259

[6] Zhou Bin, Pei Jian, Luk W-S. A Brief Survey on Anonymization Techniques for Privacy Preserving Publishing of Social Network Data[J]. ACM SIGKDD Explorations, ACM Press, 2008, 10 (2): 12-22

[7] Zhou B, Pei J. Preserving Privacy in Social Networks against Neighborhood Attacks[C]// IEEE International Conference on Data Engineering(ICDE). 2008: 506-515

[8] Byun J W, Kamra A, Bertino E, et al. Efficient k-Anonymization using Clustering Techniques[C]// International Conference on Database Systems for Advanced Applications(DASFAA). 2007: 188-200

[9] Blake E K C, Merz C J. UCI repository of machine learning databases[EB/OL]. http//www. ics. uci. edu/~mlearn/MLRepository. html, 1998

[10] 岑婷婷，韩建民，王基一，等. 隐私保护中 K-匿名模型的综述[J]. 计算机工程与应用，2008, 44(4): 130-134

[11] 姚瑶，吉根林. 一种基于隐私保护的分布式聚类算法[J]. 计算机科学，2009, 36(3): 100-105

[12] 杨秋伟，刘玲，李肯立，等. 一种支持隐私保护的角色访问控制模型[J]. 计算机科学，2010, 37(6): 46-121

[13] 毛韶阳，李肯立，等. k-means 初始聚类中心优化算法研究[J]. 重庆邮电大学学报：自然科学版，2007, 19(4): 422-425