基于 DSNPP 算法的社交网络隐私保护方法

张付霞,蒋朝惠
(贵州大学 计算机科学与技术学院,贵州 贵阳 550025)

摘 要:社交网络发展迅速,数据发布过程中存在的一个重要安全隐患就是隐私泄露。针对目前大多数社交网络隐私保护研究存在的“人员属性隐私保护”和“社区结构保护”之间没有实现真正结合的问题,就两者综合考虑,提出一种基于密度聚类算法的社交网络隐私保护方法(Density for Social Network Privacy-Preserving,DSNPP)。该算法通过对节点进行密度聚类分析,得到任意形状的簇,采用对簇内节点进行泛化、在簇内插入真实节点、增加相应边等技术来保护节点的信息和节点之间的关系信息,从而实现了人员属性隐私保护和社区结构保护两方面的真正结合。最后,通过实验表明,与 p-Sensitive k -匿名算法、GSNPP 算法相比,该算法信息丢失量上优势明显,可以获得更高的隐私保护。

关键词:社交网络;隐私保护;密度聚类;真实节点;泛化
中图分类号:TP309.2 **文献标识码**:A **文章编号**:1673-629X(2015)08-0152-04
doi:10.3969/j.issn.1673-629X.2015.08.032

Privacy-preserving Approach in Social Networks Based on DSNPP Algorithm

ZHANG Fu-xia,JIANG Chao-hui
(College of Computer Science and Technology,Guizhou University,Guizhou 550025,China)

Abstract:With the rapid development of social network,an important safety hazard that exists in the process of data publishing is leakage. For the questions that most researches on social network privacy protection do not realize existence of protecting privacy in property and community structures,considering the both,propose a method of social networking privacy,Density for Social Network Privacy-Preserving (DSNPP). The algorithm is based on density clustering method,which gets clusters in arbitrary shape through nodes cluster analysis,and it uses the technology of generalizing cluster nodes,inserting the real nodes in the cluster,increasing corresponding edges and so on to protect information of nodes and the relationship between nodes,which achieves purpose of social networks privacy protection. Finally,compared with p-Sensitive k -anonymous algorithm and GSNPP algorithm,the algorithm has the advantage in the amount of information loss,and it can obtain higher privacy protection.

Key words:social networks;privacy preservation;density clustering;real nodes;generalization

0 引 言

在应用需求的驱使下,社交网络隐私保护技术已经得到了广泛研究,并产生了大量研究成果。其中,k -匿名模型[1]、l -多样性模型[2]以及 (a,k) -匿名模型[3]对隐私保护研究产生了深远影响。后续研究中很多成果都是基于这些模型的改进,如个性化 (a,k) -匿名算法[4]、个性化 (P,α,k) 匿名模型[5]等。

在社交网络隐私保护研究领域,文献[6-8]提出了一种社交网络发布中敏感边的隐私保护算法和一种基于 k -同构的动态社交网络隐私保护算法。Liang Xiaohui 等[9]将社会网络隐私保护推向了实际应用领域;Liu Hua 等[10]研究了社交网络中移动用户的隐私保护问题等。但是目前大多数的研究都是在人员属性隐私保护和社区结构保护之一进行,并没有实现二者的结合。

故文中提出的基于密度聚类算法,不仅可以有效地保护人员的隐私信息,同时也对其交往频繁的朋友隐私进行了保护。最后,用户可以根据隐私保护程度,不仅可对簇的大小作相应处理,而且可以控制生成真实节点的数量,真正实现了用户隐私的个性化保护。

收稿日期:2014-09-18 **修回日期**:2014-12-23 **网络出版时间**:2015-07-21
基金项目:贵州省科学技术基金项目(黔科合 J 字[2012]2128 号);贵州大学研究生创新基金资助项目(校研理工 2015017)
作者简介:张付霞(1987-),女,硕士,研究方向为计算机应用技术;蒋朝惠,教授,研究方向为通信网络与信息安全。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20150721.1439.036.html

1　节点密度聚类分析

1.1　密度的相关定义

定义 1 密度:社交网络中任意一节点的密度是指以该节点为圆心,以 R 为半径的区域内所包含节点的数量。

定义 2 邻域:社交网络中的邻域是指以任意节点为圆心,以 R 为半径所包含区域中节点的集合。

定义 3 核心节点[11]:称一个节点 v 为核心节点,当且仅当,对于给定的邻域半径 R 和邻域密度阈值 Points, v 的邻域中包含的节点数大于或等于 Points 个节点。

定义 4 密度可达:给定一个节点集合 P,如果存在节点链 $v_1, v_2, \cdots, v_n$,对于 $v_i \in V(1 \leqslant i \leqslant n)$, v_{i+1} 在 v_i 的邻域内,而且 v_i 是核心节点,则节点 v_n 是从位置 v_1 关于 R 和 Points 密度可达。假设 Points = 4,图 1(左)是节点 v_1 到 v_3 关于 R 和 Points 密度可达。

定义 5 密度相连:对于节点集合 P,如果 $V \in P$,并且位置 v_i 和 v_j 都是从 V 关于 R 和 Points 密度可达的,则位置 v_i 到 v_j 是关于 R 和 Points 密度相连的,如图 1(右)所示。

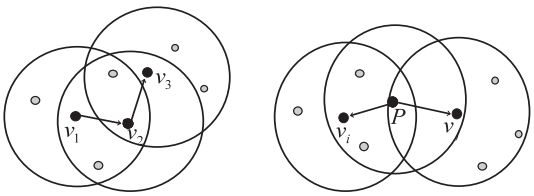

图 1　密度可达(左)和密度相连(右)

定义 6 簇(cluster):一个基于密度可达的最大密度相连节点对象的集合。

定义 7 噪声点:所有簇都不包含的位置点称为噪声点。

1.2　社交网络的相关定义

定义 8 社交网络:把无向图 $G = (V, E)$ 定义为社交网络,其中顶点集 $V = \{v_1, v_2, \cdots, v_n\}$ 定义为社交网络中的节点集合,边集合 E 为社交网络中的边集合。其中,对于每一个节点为一个实体,每条边表示两个实体之间的关系,且类型相同。

定义 9 准标识符属性泛化集[12]:假设社交网络最终生成的簇为 $\mathrm{Clt} = (\mathrm{Clt}_1, \mathrm{Clt}_2, \cdots, \mathrm{Clt}_n)$,簇中节点准标识符属性分为数值属性 Q_n 和非数值属性 Q_c,则 $Q = Q_n \cup Q_c$。

(1) $Q_n = \{n_1, n_2, \cdots, n_k\}$ 为准标识符中数值属性集合,对节点 V_i 的某一数值属性 A 泛化后的所有范围为 $\mathrm{Range}(A)$。

(2) $Q_c = \{c_1, c_2, \cdots, c_k\}$ 为准标识符中非数值属性集合,对节点 V_i 的某一非数值属性 A, $\mathrm{Sub}(A)$ 表示 A 的所有泛化范围。

2　社交网络隐私保护量化处理模型

数据在发布之前,为保护数据隐私,必须对数据进行相关的处理,通常能够唯一标识用户的标识符属性会被移除,但是攻击者仍然可以通过与外部信息进行匹配进行关联攻击,用户身份很容易被泄露。所以,文中主要从保护准标识符属性之间的关系出发,对准标识符进行泛化处理。以下是对社交网络进行匿名化处理的思想。

2.1　节点属性信息丢失量化方法

给定属性 A, A 上某个属性值 x 的信息损失计算方式如下:

若 A 为数值型属性,该信息损失为

$$\mathrm{NCP}_A(x) = \frac{\mathrm{Range}(x)}{\mathrm{Range}(R_A)} \tag{1}$$

若 A 为非数值型属性,该信息损失为

$$\mathrm{NCP}_A(X) = \frac{|\ \mathrm{Sub}(x)\ |}{|\ \mathrm{Sub}(R_A)\ |} \tag{2}$$

其中, $\mathrm{Range}(x)$ 和 $\mathrm{Sub}(x)$ 表示字段 x 泛化的范围; $\mathrm{Range}(R_A)$ 和 $\mathrm{Sub}(R_A)$ 表示元素 A 的所有泛化范围。在元组 t 上属性值的信息损失计算方式:

$$\mathrm{NCP}(t) = \sum_{i=1}^{n} \mathrm{NCP}_{A_i}(t[A_i]) \tag{3}$$

2.2　簇内节点间结构信息丢失量化方法

假设节点 V 的度为 k,则与其相连的这 k 个顶点可以形成的边的最大数目是 $T(n) = k(k-1)/2$,而在社交网络中实际数目是 $E(n)$,则节点 V 的聚集系数定义为 $C(n) = E(n)/T(n)$。簇内节点间结构信息丢失量采用泛化前后节点的聚集系数之和差值来量化。

3　DSNPP 算法

DSNPP 算法是基于密度聚类算法,可以根据社交网络的特征,生成符合大小的任意形状的簇,同时通过量化数据信息丢失量和结构信息丢失量来对数据进行较高的有效性分析,使得构造匿名化社交网络过程中的信息丢失量相对最少。最后对成功生成的簇,计算每个簇中应生成真实节点的数目,通过增加边,更好地保护了实体之间的联系信息。

3.1　量化信息丢失量

3.1.1　量化数据信息丢失量

假设 Clt 是一个簇,准标识符为 $Q = (n_1, n_2, \cdots, n_s, c_1, c_2, \cdots, c_t)$,则泛化准标识符所产生的数据信息丢失量为

$$\mathrm{QL}(\mathrm{Clt}) = |\ \mathrm{Clt}\ | * \left[\sum_{i=1}^{s}\left(\frac{\mathrm{Range}(n_i)}{\mathrm{Range}(n)}\right) + \sum_{j=1}^{t}\left(\frac{\mathrm{Sub}(n_j)}{\mathrm{Sub}(c)}\right)\right] \tag{4}$$

其中，|Clt| 为簇中节点的数量。假设共生成 m 个簇，则总数据信息丢失量为

$$TQL(G) = \sum_{i=1}^{m} (QL(Clt)) \tag{5}$$

3.1.2 结构信息丢失量

假设社交网络 $G = (V,E)$，$V = \{v_1, v_2, \cdots, v_n\}$，假设某一个簇 Clt 中节点集合 $V' = \{v_1, v_2, \cdots, v_r\}$，设该簇中节点 v_i 的度为 $k(v_i)$，与节点 v_i 相连接的其他节点可以形成的边的最大数目是

$$T(v_i) = k(v_i) * (k(v_i) - 1)/2 \tag{6}$$

在社交网络中与节点 v_i 相连接的其他节点形成边实际的数目是 $E(v_i)$，则节点 v_i 的聚集系数为

$$C(v_i) = E(v_i)/T(v_i) \tag{7}$$

故社交网络 G 的聚集系数之和为

$$FQL = \sum_{i=1}^{n} C(V_i) \tag{8}$$

簇 Clt 的聚集系数之和为

$$EQL = \sum_{k=1}^{m} \left(\sum_{i=1}^{|CL_i|} C(V_i) \right) \tag{9}$$

总结构信息丢失量为

$$NTQL = FQL - EQL \tag{10}$$

3.2 真实节点选取

插入虚拟节点有助于简化操作，但会引入过多额外信息，从而降低了数据的可用性，故 DSNPP 算法选择插入真实节点，这样有助于克服插入虚拟节点的不足。同时插入的节点从相邻簇中选取而不是从当前簇中产生，这有助于降低由于同一簇中集聚过多具有相同描述属性信息的节点，从而造成基于属性再识别攻击的隐私泄露的风险。

3.3 边的增加

由用户对隐私保护的要求，来确定生成真实节点的总数目，假定为 N，同时假定共生成 m 个簇，则每个簇中生成真实用户的数量为 N_i，n_i 表示第 i 个簇中的节点数目，则

$$N_i = \left(\frac{n_i}{n_1 + n_2 + \cdots + n_m}\right) * N, 1 \leqslant i \leqslant m \tag{11}$$

最后，在不改变其他节点连接情况的基础上，选取度数最少的节点，分别与新生成的真实节点相连接。

3.4 算法描述

第一阶段（生成簇集合）。

输入：社交网络图 $G = (V,E)$，邻域半径为 R，邻域密度阈值 Points，其中 $V = \{v_1, v_2, \cdots, v_n\}$；

输出：一组聚类簇 $Clt = \{Clt_1, Clt_2, \cdots, Clt_m\}$。

（1）将 V 中的所有节点设置为未被标记的状态；

（2）判断 V 中节点数，若 $|V| > 0$，执行步骤 3，否则跳转步骤 8；

（3）提取 V 中度最大节点 v，生成簇得到子图 Gclt = (Vclt, Eclt)；

（4）判断 v 是否已被标记，若已被标记则执行步骤 2，否则执行步骤 5；

（5）判断该节点是否为核心点，若是执行步骤 6，若不是则标记为噪声点执行步骤 2；

（6）找到所有由 v_i 关于 R，Points 密度可达的节点 v_{i+1}，根据重要性权重参数 $\alpha \in [0,1]$ 和式（5）及式（9）来计算要添加到簇中节点 v_i 的信息丢失量。

$$V_i loss = \partial * TQL(G) + (1 - \partial) * NTQL, 1 \leqslant i \leqslant n \tag{12}$$

（7）若 $V_i loss > V_{i+1} loss$，转至步骤 6，否则像簇 Gclt 中添加节点 v_i；

（8）输出簇集合 $Clt = \{Clt_1, Clt_2, \cdots, Clt_m\}$ 重复以上步骤，直到处理完所有的点，剩下的即为噪声点。

第二阶段（添加真实节点和边）。

（1）对于第一阶段得到的各个簇 $\{Clt_1, Clt_2, \cdots, Clt_m\}$ 计算每个簇内的节点数为 $n_1, n_2, \cdots, n_m$；

（2）根据式（2）计算在每个簇中生成真实节点的数量 N_i，在生成真实节点的数量不足 1 的簇中生成一个真实节点；

（3）在不改变其他节点连接情况的基础上，选取度数最少的节点，分别与新生成的真实节点相连接。

4 模拟实验

DSNPP 算法采用基于密度的方法对网络节点进行聚类分析，方法简洁，易于计算，同时对聚类后的簇进行插入真实节点和增加边，不仅增强了节点准标识符的保护，对节点之间的关系保护也得到显著提高。

在量化信息丢失量方面，实验采用 Pajek 软件生成满足幂律分布及小世界特征[13]的边集合，以此来模拟一个社交网络。使用美国机器学习库中的 Adult 数据库作为数据源，数据库大小为 3.9 M，随机抽取 500 条记录作为社交网络节点，同时对该记录做相应的处理，将{Age, workclass, education, sex}作为本次实验的准标识符。文中使用 MATLAB 开发的一个简单应用程序来验证 DSNPP 算法的有效性。

在不同准标识符个数的情况下，DSNPP 算法与 p-Sensitive k -匿名算法信息损失差异比较如图 2 所示。

由图 2 可知，在准标识符的个数不断增加的过程中，两种算法的信息损失量都在不断增大，但 DSNPP 算法[14]明显比 p-Sensitive k -匿名算法在相同标识符数量下的信息损失量低，即在相同隐私保护度前提下，DSNPP 在保存节点的属性信息和节点之间的结构信息两方面比 p-Sensitive k -匿名算法更具有优势。

当 R 固定，Points 为 2 时，在不同的重要性参数 α

的条件下,DSNPP 算法与 GSNPP 算法相比较,数据属性信息损失量与结构属性信息损失量的变化情况如图 3 所示。

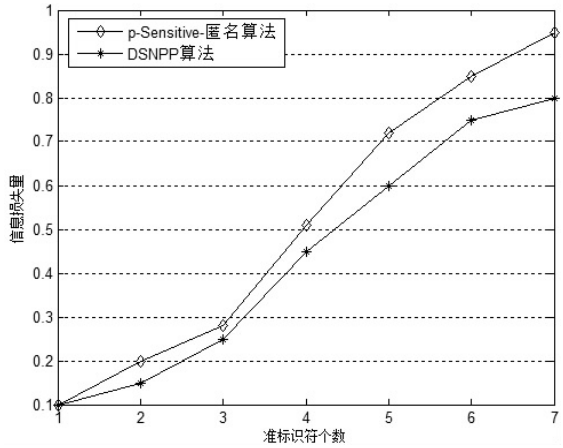

图 2　准标识符个数与信息损失差异关系图

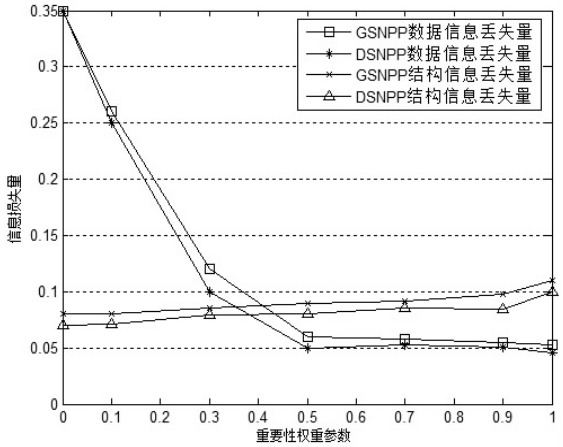

图 3　重要性权重参数 α 与信息丢失量的关系图

由图 3 可见,当重要性权重参数 α 逐渐增大时,两种算法准标识符的数据信息丢失量不断减少,而结构属性信息丢失量不断增大。但是 DSNPP 算法的数据信息丢失量和结构信息丢失量均比 GSNPP 算法在相同重要性权重参数下数据信息丢失量和结构信息丢失量略低。实验结果表明,DSNPP 算法是合理有效的。

5　结束语

DSNPP 算法是基于密度的聚类算法,用户可以根据自身隐私保护度要求,设置合适的 R 和 Points 来控制聚类簇中节点的数量和簇的大小,与其他算法比较可知,DSNPP 算法在社交网络构匿名化过程中信息丢失量最少。同时,为了在节点信息保护和节点之间信息保护两者之间进行合理的权衡,引入了重要性权重参数α,实现了个性化的隐私保护。最后在聚类后的各个簇中,插入相应的真实节点和增加节点之间的边,进一步增强了隐私保护力度。由实验得知,该算法在丢失较少信息量的基础上获得了更高的隐私保护。

参考文献:

[1] Sweeney L. k-anonmity:a model for protecting privacy[J]. International Journal of Uncertainty Fuzziness and Knowledge Based Systems,2002,10(5):557-570.

[2] Machanavajjhala A,Kifer D,Gehrke J,et al. l-diversity:privacy beyond k-anonymity[J]. ACM Transactions on Knowledge Discovery from Data,2007,1(1):1-52.

[3] Wong R,Li J,Fu A.(a,k)-anonymity:an enhanced k-anonymity model for privacy preserving data publishing[C]// Proc of 2006 12th ACM SIGKDD international conference on knowledge discovery and data mining.[s.l.]:ACM,2006:754-759.

[4] 韩建民,于　娟,虞慧群,等. 面向敏感值的个性化隐私保护[J]. 电子学报,2010,38(7):1723-1728.

[5] 孔庆江. 社交网络中个人信息与人际关系的隐私保护研究[D]. 杭州:浙江工业大学,2011.

[6] 兰丽辉,孙英慧,鞠时光. 社会网络发布中敏感边的隐私保护[J]. 吉林大学学报:信息科学版,2011,29(4):324-331.

[7] 张晓琳,李玉峰,刘立新,等. 社会网络隐私保护中 K-同构算法研究[J]. 微电子学与计算机,2012,29(5):99-103.

[8] 张晓琳,李玉峰,王　颖. 动态社会网络隐私保护方法研究[J]. 计算机应用研究,2012,29(4):1434-1437.

[9] Liang Xiaohui,Barua M,Lu Rongxing,et al. HealthShare:achieving secure and privacy-preserving health information sharing through health social networks[J]. Computer Communications,2012,35(15):1910-1920.

[10] Liu Hua,Krishnamachari B,Annavaram M. Game theoretic approach to location sharing with privacy in a community-based mobile safety application[C]//Proc of 2008 4th ACM international symposium on QoS and security for wireless and mobile networks.[s.l.]:ACM,2008:27-31.

[11] 郭晓丽. 基于位置服务的移动对象隐私保护技术研究[D]. 哈尔滨:哈尔滨工程大学,2013.

[12] 韦　伟. 基于贪心算法的社会网络隐私保护方法研究[D]. 重庆:西南大学,2012.

[13] van Dongen S. Graph clustering by flow simulation[D]. Utrecht:Univ of Utrecht,2000.

[14] 韦　伟,李　杨,张为群. 一种基于 GSNPP 算法的社交网络隐私保护方法研究[J]. 计算机科学,2012,39(3):104-106.