doi:10.11959/j.issn.1000-436x.2015290

# 基于聚类的动态社交网络隐私保护方法

谷勇浩[1]，林九川[2]，郭达[3]

（1. 北京邮电大学 计算机学院 智能通信软件与多媒体北京市重点实验室，北京 100876；
2. 公安部第三研究所，上海 201204；3. 北京邮电大学 电子工程学院，北京 100876）

**摘 要**：由于社交网络图结构的动态变化特性，需要采用有效的动态隐私保护方法。针对现有动态数据发布隐私保护方法中存在的攻击者背景知识单一、对图结构动态变化适应性较低等问题，提出基于聚类的动态图发布隐私保护方法。分析表明，该方法能抵御多种背景知识攻击，同时对社交网络图结构动态变化具有较好的适应性。

**关键词**：动态社交网络；隐私保护；聚类；信息损失度；隐匿率

**中图分类号**：TP393 **文献标识码**：A

# Clustering-based dynamic privacy preserving method for social networks

GU Yong-hao[1], LIN Jiu-chuan[2], GUO Da[3]

(1. Beijing Key Laboratory of Intelligent Telecommunications software and Multimedia, School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. The Third Research Institute of Ministry of Public Security, Shanghai 201204, China;
3. School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China)

**Abstract:** Due to the dynamic characteristics of the social network graph structure, an effective dynamic privacy preserving method was needed. To solve the problems of the existing dynamic privacy preservation methods, such as attacker's too little background knowledge and the low adaptability to the dynamic characteristics of graph structure, a clustering-based dynamic privacy preservation method was provided. The analysis shows that the proposed method can resist many kinds of background knowledge attacks and has good adaptability to the dynamic characteristics of the social network graph structure.

**Key words:** dynamic social networks; privacy preserving; clustering; information loss degree; anonymization rate

## 1 引言

社交网络作为新兴的互联网应用模式受到广泛关注，社交网络中的数据对经济预测、舆情分析等具有重要意义，然而数据中含有大量的个人隐私信息，直接发布会造成用户隐私泄露。如何在发布数据时对用户隐私进行有效保护，已经成为一个重要的研究方向[1]。

数据发布隐私保护的研究内容涉及到保护方法、并行性和动态性等多方面[2]。其中，隐私保护方法包括聚类、数据扰动、泛化、随机化以及推演控制[3]等。并行性体现在海量的社交网络数据分析和处理需要有效的并行算法来实现[4]。同时，现有大多数隐私保护方法只针对单次数据发布进行研究，即数据发布后不再进行任何改变，而社交网络不断发展变化的特性决定了社交网络的动态性，无时无刻不在发生着数据的更新迭代。社交网络的动态性决定了单次数据发布的隐私保护方法（静态隐私保护方法）不能保证动态社交网络的隐私安全。在数据更新迭代过程中，如何进行有效的数据发布

**收稿日期**：2015-10-29

**基金项目**：国家自然科学基金资助项目（61173017）；工业和信息化部通信软科学基金资助项目（2014-R-42）；信息网络安全公安部重点实验室开放课题基金资助项目（C14613）

**Foundation Items:** The National Natural Science Foundation of China(61173017); Communication Soft Science Foundation of Ministry of Industry and Information(2014-R-42); Key Lab of Information Network Security Foundation of Ministry of Public Security (C14613)

隐私保护具有重要的理论意义和实用价值，社交网络中动态数据发布隐私保护方法的研究成为一个重要研究课题[5]。

本文在对社交网络动态数据发布隐私保护方法研究的基础上，着重解决现有方法中攻击者背景知识单一、对社交网络图结构动态变化适应性低等问题，提出基于聚类的动态图发布隐私保护方法。

## 2　相关工作

社交网络数据发布隐私保护技术中多采用聚类的方法，这类方法是将社交网络中发布的节点和边划分到不同的聚类组中，每个组被称为超级顶点，每组中会有该组所划分的节点和边构成一个子图，每个聚类组中的信息对外不可见。基于聚类的社交网络数据发布隐私保护技术主要包括：节点聚类技术[6,7]、边聚类技术[8,9]、节点和边聚类技术[8,10]以及节点属性映射聚类技术[11]。

社交网络的动态性体现在网络结构图的不断更新迭代，包括网络节点和边的增加或删除以及属性值的更新，随之出现了多种动态隐私保护方法。Viswanath 等[12]对社交网络的动态性进行了分析，提出匿名方法应该满足网络数据随时间变化的性质。Cheng 等[13]采用节点标识符(ID)泛化处理动态网络的多重发布，并没有给出如何处理节点增加或删除的情况，而且数据的可用性和匿名质量取决于子图的划分和节点标识的泛化。Zou 等[14]采用自同构方法，通过增加、删除边和节点使得网络结构图满足 *k*-自同构，同时通过节点泛化满足动态图结构数据隐私保护的需求，但是该方法不能有效阻止全部的结构化攻击。张晓琳等[15]针对攻击者基于背景知识的结构化攻击方法，提出一种优化的图划分算法。采用 *k* 同构算法将原始图划分为 *k* 个同构子图，然后通过节点 ID 泛化方法，Karwa 等[16]防止攻击者利用网络图多重发布时的关联攻击，该方法未解决属性值更新过程中的隐私保护。文献[16~18]将差分隐私保护方法应用到社交网络中，由于图中节点间的高度相关性，当数据规模较大时隐私保护算法的复杂度较高，如何降低复杂度是关键。张伟等[19]针对攻击者具有不同时刻节点邻域子图的背景知识，提出基于 *k*-邻域同构的动态隐私保护方法，该方法采用下三角矩阵表示社交网络图中节点的 l-邻域子图，通过对矩阵的修改和异或运算实现匿名保护。但是，该方法的时间复杂度较大。

综上所述，现有动态数据发布隐私保护方法主要不足包括：对社交网络图结构动态变化适应性不高(只解决节点和边增加、删除以及属性值更新的部分变化)，算法执行效率低，未考虑攻击者具有背景知识的攻击（或只考虑单一背景知识的攻击）。

## 3　基于聚类的动态图发布隐私保护方法

为解决现有动态数据发布隐私保护方法中存在的不足，本文在基于图结构相似的 *K* 匿名算法基础上，提出基于聚类的动态图发布隐私保护方法。

### 3.1　基于图结构相似的 *K* 匿名算法

基于图结构相似的 *K* 匿名算法[20]包括 3 部分：节点聚类、倒排表索引的构建[21]和聚类边匹配[21]，使得匿名后发布的结构图满足 *K* 匿名要求。首先，对社交网络中的相似顶点进行聚类，节点聚类可以很好地隐藏节点的属性特征，将大量具有相似属性的节点聚类，从而保护节点属性值和节点度。其次，基于节点边的权值构建一个倒排序表索引，并且根据权值出现次数的多少进行排序。然后得到聚类中基于不同权值的所有等值边的匹配，最后处理聚类间的边，由此得到聚类隐私保护后的社交网络图。

基于图结构相似*K*匿名算法可抵御攻击者具有顶点度和边权值等背景知识的攻击，同时该方法数据可用性较好。但是，该方法存在如下问题。首先，在节点聚类后，该算法不能进行节点删除，否则会减少聚类中的节点数，造成隐私保护程度降低，导致删除节点的聚类中节点隐私泄露。其次，该算法每次聚类前选取的第一节点会随着图结构的更新迭代而变化，导致聚类隐私保护时产生的匿名图结构改变，这在单次数据发布中虽然没有影响，但是进行多次数据更新后，会造成节点隐私的泄露。

### 3.2　基于聚类的动态图发布隐私保护方法

为解决上述问题，本文先引入节点度的信息损失度和隐私参数 2 个概念，提出隐私参数的节点聚类算法。然后引入聚类间距离和代替聚类的概念，解决基于图结构相似的 *K* 匿名算法无法进行节点删除的不足，同时对原有节点修改操作中的保护方式进行改进，形成基于聚类的动态图发布隐私保护方法。

**定义 1**　节点度的信息损失度。社交网络图 $G(V,E,W)$，$V$ 是 $G$ 中的节点集，$E$ 是 $G$ 中的边集，$W$ 是 $G$ 中的边权重集。用 $d(v)$表示节点 $v$ 的度, $D(c_i)$ 表示聚类 $c_i$ 中所有节点度的和，$n$ 为聚类 $c_i$ 中节点数目。如果将节点 $v$ 隐匿到聚类 $c_i$ 中，节点 $v$ 的度

的信息损失度定义为

$$NLoss(v)=\left|d(v)-\frac{D(c_i)}{n}\right| \tag{1}$$

**定义 2** 隐私参数。社交网络图 $G(V,E,W)$，$v$ 是 $G$ 中的节点，$c_i$ 是 $G$ 中第 $i$ 个聚类，$C$ 是 $G$ 中聚类的集合。寻找隐私参数 $k$，使得 $G$ 中所有聚类满足条件 $\forall c_i \in C, k-1<n\leqslant k+1$，其中，$n$ 为聚类 $c_i$ 中节点的个数。

基于隐私参数 $k$ 的节点聚类算法如下。

**算法 1** 基于隐私参数 $k$ 的节点聚类算法

输入：节点集合 $V$，隐私参数 $k$

输出：聚类集合 $C$

1）寻找 $V$ 中节点度最大的节点,记为第一节点 $v_1$。

2）计算 $v_1$ 与其他节点间的节点度的信息损失度。

3）按照节点度的信息损失度从小到大排列。

4）将前 $k$–1 个节点与第一节点 $v_1$ 聚类为 $c_1$。

5）重复步骤 1~步骤 4，直到得到所有聚类。

6）将所有聚类保存到聚类集合 $C$ 中。

7）输出 $C$。

**定义 3** 聚类间距离。社交网络图 $G(V,E,W)$，$V$ 是 $G$ 中的节点集。假设 $c_1,c_2,\cdots,c_t \subset V$ 为 $t$ 个不相交的节点聚类。$c_i$ 和 $c_j$ 为 $V$ 中任意 2 个聚类，$D(c_i)$ 和 $D(c_j)$分别表示 $c_i$ 和 $c_j$ 中所有节点度的和，$n_i$ 和 $n_j$ 分别表示 $c_i$ 和 $c_j$ 中各自的节点数。聚类 $c_i$ 和 $c_j$ 间的距离定义为

$$d(c_i,c_j)=\left|\frac{D(c_i)}{n_i}-\frac{D(c_j)}{n_j}\right| \tag{2}$$

**定义 4** 代替聚类。社交网络图 $G(V,E,W)$，假设 $c_1,c_2,\cdots,c_t \subset V$ 为 $t$ 个不相交的聚类。定义 $c_i$ 的代替聚类是与 $c_i$ 距离最近的聚类 $c_j$。

基于聚类的动态图发布隐私保护算法如下。

**算法 2** 基于聚类的动态图发布隐私保护算法

输入：$t$ 时刻社交网络图 $G_t$，$t$ 时刻发布的匿名社交网络图 $G_t'$，$t$+1 时刻社交网络图 $G_{t+1}$ 和隐私参数 $k$

输出：$t$+1 时刻的匿名社交网络图 $G_{t+1}'$

1) 遍历 $G_t$ 和 $G_t'$；

2) 得到 $G_t$ 所有节点的集合 $V_t$;

3) 根据算法 1 得到 $G_t'$ 所有的聚类集合 $C_t$;

4) 遍历 $G_{t+1}$;

5) 得到 $G_{t+1}$ 上所有节点的集合 $V_{t+1}$;

6) 通过对 $V_{t+1}$ 和 $V_t$ 比较得到添加、删除和修改的节点集合分别为 $V_a$，$V_d$，$V_{ts}$;

7) foreach 添加的节点 $v_a \in V_a$

8) 使用式(1)计算 $v_a$ 与各聚类之间的节点度的信息损失度;

9) 添加节点 $v_a$ 到信息损失度最小的聚类 $c_i$ 中;

10) if $c_i$ 中节点个数满足 $n$==$k$+1

11) 使用式(2)计算聚类间距离 $d(c_i,c_j)$ 确定 $c_i$ 的代替聚类 $c_j$;

12) 添加节点 $v_a$ 到 $c_j$ 中;

13) end if

14) end for

15) foreach 删除的节点 $v_d \in V_d$

16) if $v_d$ 所属聚类 $c_i$ 中节点数目小于 $k$

17) 利用式(2)找到与 $c_i$ 聚类间距离最小的聚类 $c_j$;

18) 合并 $c_i$ 和 $c_j$ 为 $c'$；

19) if $c'$ 的节点个数大于等于 $2k$

20) 将 $c'$ 分为 2 个大于等于 $k$ 的聚类 $c_i'$ 和 $c_j'$；

21) $c_i=c_i'$， $c_j=c_j'$；

22) end if

23) end if

24) end for

25) foreach 修改的节点 $v_{ts} \in V_{ts}$，修改后的节点记为 $v_{ts}'$

26) $v'=v_{ts}'$；

27) 执行步骤 16~步骤 23，对节点 $v_{ts}$ 进行删除操作;

28) 执行步骤 8~步骤 13，添加 $v'$；

29) end for

30) return $C_{t+1}$;

31) 根据文献[21]构建倒排表索引;

32) 利用文献[21]中的算法对新的聚类集合 $C_{t+1}$ 进行聚类边匹配;

33) return $G_{t+1}'$。

## 4 实验分析

实验环境使用 CPU 为主频 2.4 GHz Core i5,内存 4 GB，操作系统使用虚拟机下的 Windows 7 系统，开发语言使用 Visual C++6.0 和 MATLAB 8.5。

实验数据是美国 Enron 公司的邮件数据集，其中包含 150 名用户、50 万封以上邮件，放在 3 500 个左右的文件夹中。为体现动态性，本文选取

2004.3.2，2009.8.21，2011.4.2 和 2015.5.7 的数据。其中，有部分邮件重复，删除重复数据后，将其中不同发件人作为社交网络中的节点，邮件往来作为节点间的边，通信次数作为边权重，统计可知该社交网络图中共有 74 000 多个节点，260 000 多条边。

本节将从算法执行时间，防御不同背景知识攻击时隐匿率的变化，以及算法动态适应效果等方面对所提方法进行分析。

### 4.1　算法执行时间的分析

在节点更改数量变化时，比较 3 种更新方式（添加、修改、删除）执行时间的差异。如图 1 所示的更改节点数目分别为 100、200、300、400，可以发现，在对节点进行删除时，执行时间变化最大，节点添加时，执行时间变化最小，3 种数据变动的执行时间均随着节点更改数目的增多而增加。

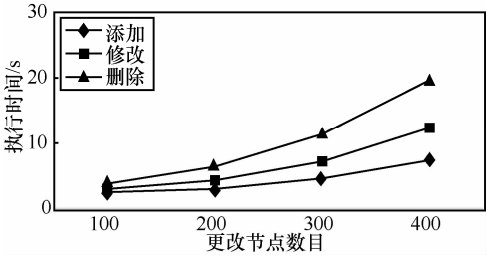


图 1　更新节点数目对执行时间的影响

### 4.2　防御不同背景知识攻击下隐匿率变化的分析

**定义 5**　隐匿率。隐匿的节点数/边数占社交网络图中总节点数/边数的比例。

从图 2 可以看出，隐私参数 $k$ 越大，隐匿率越小，隐私泄露程度越小。因为 $k$ 为聚类中包含相似节点的数量，$k$ 越大，相似节点数量越多，攻击者利用节点度攻击得到的相似节点也越多，发现目标节点的概率就越小。而且，本文所提方法对攻击者具有边权重背景知识攻击的防御效果要好于攻击者具有节点度背景知识攻击。

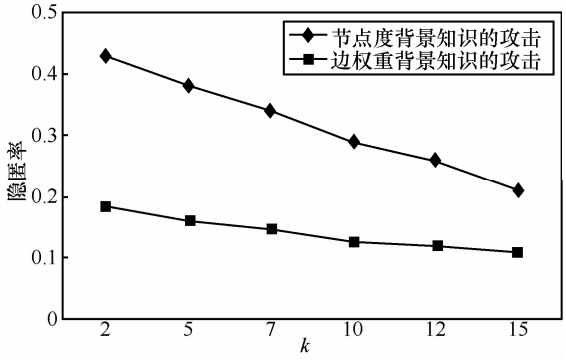


图 2　防御不同背景知识攻击下 $k$ 取值对隐匿率的影响

### 4.3　算法动态适应效果的分析

从图 3 可以看出，在多次迭代后，基于图结构相似的 $K$ 匿名算法的隐匿率呈几何形式增长，在迭代 3 次后已经无法满足隐私保护对隐匿率（小于 0.5）的要求。而本文所提方法在隐匿率上虽然也有升高，但是并不明显，而且始终在隐匿率要求（小于 0.5）范围内。说明本文所提方法在多次迭代更新的情况下具有较好的隐私保护效果，动态适应效果较好。

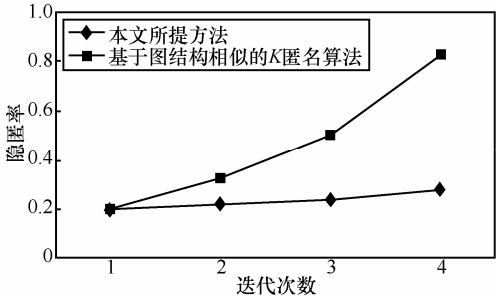


图 3　2 种方法中隐匿率随迭代次数变化趋势的对比

## 5　结束语

为解决现有动态数据发布隐私保护方法中存在的攻击者背景知识单一、对社交网络图结构动态变化适应性低等问题，本文在基于图结构相似的 $K$ 匿名算法基础上，提出基于聚类的动态图发布隐私保护方法。实验分析结果表明，本文所提方法具有较好的隐私保护效果（隐匿率小于 0.5），而且能在抵御多种背景知识攻击的同时，具有较好的动态适应效果。今后工作重点是如何降低算法复杂度，提高算法执行效率。

### 参考文献：

[1] 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 Y Y, ZHANG M, FENG D G，*et al*. Attribute privacy preservation in social networks based on node anatomy[J]. Journal of Software, 2014, 25(4): 768−780.

[2] 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 X Y, WANG B, YANG X C. Survey on privacy preserving techniques for publishing social network data[J]. Journal of Software, 2014, 25(3):576-590.

[3] LIU X, YANG X. Protecting sensitive relationships against inference attacks in social networks[A]. Proc of the 17th Int'l Conf on Database Systems for Advanced Applications[C]. 2012. 335-350.

[4] GAO J, XU J Y, JIN R. Neighborhood-privacy protected shortest distance computing in cloud[A]. Proc of the 2011 ACM SIGMOD Int'l

Conf on Management of Data[C]. 2011. 409-420.

[5] BHAGAT S, CORMODE G, SRIVASTAVA D. Privacy in dynamic social networks[A]. Proceedings of the 19th International Conference on World Wide Web[C]. ACM, 2010.1059-1060.

[6] YING X, WU X, YING X. Randomizing social networks: a spectrum preserving approach[J]. SDM, 2008:739-750.

[7] THOMPSON B, YAO D. The union-split algorithm and cluster-based anonymization of social networks[A]. Proceedings of the 4th International Symposium on Information, Computer, and Communications Security[C]. 2009. 218-227.

[8] ZHELEVA E, GETOOR. L. Preserving the privacy of sensitive relationships in graph data[A]. Privacy, Security, and Trust in KDD[C]. 2008. 153-171.

[9] ANDERSEN S K, JUDEA P. Probabilistic reasoning in intelligent systems: networks of plausible inference [J]. Artificial Intelligence, 1991, 48(91):117-124.

[10] CAMPAN A, TRUTA T M. A clustering approach for data and structural anonymity in social networks[J]. Privacy, Security, and Trust in KDD Workshop (PinKDD), 2008. 33-54.

[11] CORMODE G, SRIVASTAVA D, Yu T, *et al*. Anonymizing bipartite graph data using safe groupings[J]. VLDB Journal, 2010, 19(1):115-139.

[12] VISWANATH B，MISLOVE A，CHA M, *et al*．On the evolution of user interaction in facebook[A]. Proceedings of the 2nd ACM Workshop on Online Social Networks[C]. 2009. 37-42.

[13] CHENG J, FU AWC, LIU J. *K*-isomorphism: privacy preserving network publication against structural attacks[A]. Proc of the 2010 ACM SIGMOD Int'l Conf on Management of Data[C]. 2010. 459-470.

[14] ZOU L, CHEN L, OZSU MT. *K*-automorphism: a general framework for privacy preserving network publication[A]. Proc of the 35th International Conference on Very Large Databases[C]. 2009. 946-957.

[15] 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.
ZHANG X L, LI Y F, WANG Y. Research on privacy preserving method for dynamic social network[J]. Application Research of Computers, 2012, 29(4):1434-1437.

[16] KA R WA V，SOFYA R，SMITH A, *et al*． Private analysis of graph structure[A]. Proc of the 37th International Conference on Very Large Databases[C]. 2011. 1147-1157.

[17] CHEN S，ZHOU S．Recursive mechanism: towards node differential privacy and unrestricted joins[A]. Proc of the International Conference on Management of Data[C]. 2013. 653-664.

[18] CHEN R，FANG B C M，PHILIP S Y, *et al*．Correlated network data publication via differential privacy[J]．The VLDB Journal，2014，23(4):653-676.

[19] 张伟, 王旭然, 王珏,等. 基于 *k*-邻域同构的动态社会网络隐私保护方法[J]. 南京邮电大学学报(自然科学版), 2014, 34(5):9-16.
ZHANG W, WANG X R, WANG J, *et al*. Privacy preservation in dynamic social networks based on *k*-neighborhood isomorphism[J]. Journal of Nanjing University of Posts and Telecommunications(Natural Science), 2014, 34(5):9-16.

[20] 孙浩月. 防止图结构攻击的社会网络隐私保护技术研究[D]. 东北大学, 2011.
SUN H Y. Research on the techniques of social network privacy preserving against graph structural attacks[D]. Northeastern University, 2011.

[21] AGGRAWAL G, FEDER T, KENTHAPADI K, *et al*. Achieving Anonymity via Clustering[A]. Proceedings of the 25th ACM SIGACT-SIGMODSIGART Symposium on Principles of Database Systems [C]. 2006, 6(3):153-162.

**作者简介：**

**谷勇浩**（1980-），男，山西太原人，博士，北京邮电大学讲师，主要研究方向为网络安全、隐私保护等。

**林九川**（1980-），男，江苏盐城人，硕士，公安部第三研究所助理研究员，主要研究方向为信息安全。

**郭达**（1976-），男，江西南昌人，北京邮电大学电子工程学院博士后，主要研究方向为物联网、移动互联网等。