基于边介数模型的差分隐私保护方案

黄海平[1,2]，王凯[1,2]，汤雄[1,2]，张东军[1,2]
（1. 南京邮电大学计算机学院，江苏 南京 210023；2. 江苏省无线传感网高技术研究重点实验室，江苏 南京 210023）

摘 要：随着社交网络应用的不断发展，用户社交关系等个人隐私数据的安全保护问题亟待解决。为显著减小社交网络数据的敏感度，提出了一种基于边介数模型的差分隐私保护方案 BCPA。基于 *dK* 模型捕获图结构对应的 2*K* 序列，根据边中介中心性系数对 2*K* 序列重新排序；依据排序结果将 2*K* 序列聚类成多个子序列，再利用 *dK* 扰动算法对各子序列分别进行加噪；根据整合后的新 2*K* 序列生成满足差分隐私的社交网络发布图。基于真实数据集，通过模拟仿真将所提方案与其他经典方案进行比较，实验结果表明，所提方案在保证较强隐私保护性的同时，提高了发布数据的准确性和可用性。

关键词：社交网络；隐私保护；差分隐私；*dK* 模型；聚类；分组扰动
中图分类号：TP309
文献标识码：A
doi: 10.11959/j.issn.1000−436x.2019095

Differential privacy protection scheme based on edge betweenness model

HUANG Haiping[1,2], WANG Kai[1,2], TANG Xiong[1,2], ZHANG Dongjun[1,2]
1. College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
2. High Technology Research Key Laboratory of Wireless Sensor Network of Jiangsu Province, Nanjing 210023, China

Abstract: With the continuous development of social network application, user's personal social data is so sensitive that the problem of privacy protection needs to be solved urgently. In order to reduce the network data sensitivity, a differential privacy protection scheme BCPA based on edge betweenness model was proposed. The 2*K* sequence corresponding to the graph structure based on the *dK* model was captured, and 2*K* sequences based on the edge betweenness centrality were reordered. According to the result of reordering, the 2*K* sequence was grouped into several sub-sequences, and each sub-sequence was respectively added with noise by a *dK* perturbation algorithm. Finally, a social network graph satisfying differential privacy was generated according to the new 2*K* sequences after integration. Based on the real datasets, the scheme was compared with the classical schemes through simulation experiments. The results demonstrate that it improves the accuracy and usability of data while ensuring desired privacy protection level.

Key words: social network, privacy protection, differential privacy, *dK* model, clustering, group perturbation

1 引言

快速发展的社交网络应用促使大量社交数据被广泛使用，对用户隐私构成了威胁[1]。因此，社交网络的研究者通过在节点参数和图结构中添加噪声来实现数据隐私保护。然而，即使是引入噪声的图数据仍然可以实现去匿名化，尤其是在攻击者具备相应网络背景知识的前提下。随着社

收稿日期：2018–10–11；**修回日期**：2019–03–21
基金项目：国家自然科学基金资助项目（No.61672297）；江苏省重点研发计划（社会发展）基金资助项目（No.BE2017742）；江苏省六大人才高峰基金资助项目（No.DZXX-017）
Foundation Items: The National Natural Science Foundation of China (No.61672297), The Key Research and Development Program of Jiangsu Province (Social Development Program) (No.BE2017742), The Sixth Talent Peaks Project of Jiangsu Province in China (No.DZXX-017)

交网络在全球范围内的普及，攻击者将更容易获得各种类型的背景知识，隐私泄露的风险也随之增加。

为了加强隐私保护的效果，差分隐私技术[2]被应用于社交网络中。最早的差分隐私只是对社交网络数据进行一些简单的统计分析，大多只涉及属性信息而忽略了结构信息，许多重要的图属性都可以通过子图计数查询得到。其后，关系差分隐私框架出现，它保证了社交网络中任何一条边的改变都不会对查询结果造成太大的影响。然而，尽管关系差分隐私框架将需要引入的噪声量从网络中节点数量的多项式级别减少到了多项式对数级别，但相对于查询函数中节点的数目，关系差分隐私需要引入的噪声量依然是超指数级别的，查询结果的数据可用性依然很难保证。为了在不泄露隐私的情况下安全地发布真实的社交网络图数据，可应用一个更加强健的匿名技术框架[3]，其以精妙的方式修改图结构以提高隐私保护程度，而保留大部分的原始图结构。然而，这类方法通常仅能针对一种特定类型的攻击方式，而对一些新型的去匿名化攻击技术效果不佳。

鉴于此，本文将寻求一种提供图数据隐私保护且能保留图结构的方案，该方案将“重要性”相同的边划分成为单独的组进行加噪处理，在提供了较高强度的隐私保护性的同时，保证了社交网络图的数据可用性。本文的主要贡献如下。

1) 结合 dK−匿名结构将排序后的 $2K$ 序列根据条件进行聚类操作，划分成一组组子序列，使 $2K$ 序列敏感度上限显著减少，从而大大降低了所需引入的总噪声量。

2) 提出基于边介数模型的差分隐私保护方案 BCPA（betweenness centrality protection algorithm），根据介数排序的 dK 序列将原始图中“重要性”相同或相近的边聚类在同一子序列中，分组加噪时能保证不改变各条边的“重要性”，在保留了更多结构特征的同时也具有更好的数据可用性。

3) 提出一种基于邻接度的隐私保护性衡量算法，可以有效地衡量扰动对原始图数据的影响程度。

2　相关工作

近年来，针对社交网络图结构的差分隐私，研究者们取得了很多成果。Hay 等[4]提出 k−边差分隐私，即 2 个邻图最多相差 k 条边，以求扰动程度与隐私保护强度之间的平衡，与常用的节点差分隐私相比，其减小了校准噪声的添加，使图结构分析在 k−边差分隐私时较为准确。Mir 等[5]试图用随机 Kronecker 图[6]生成与原始图具有相似敏感度的差分隐私拓扑图，提供较为准确的真实图像，同时保护其中参与个体的隐私。Day 等[7]引入基于边加法的图投影方法，提出 2 种基于聚合和累加直方图的方法来发布满足差分隐私的扰动图，通过降低敏感度来生成逼近真实度分布的拓扑图。

在处理差分隐私发布图时，研究者广泛运用了 dK 图隐私模型。Sala 等[8]提出一种 dK 图隐私模型，设计了 dK-PA（dK-perturbation algorithm）处理差分隐私的图发布，使用 dK 图查询并获取图结构的 dK 序列，对查询结果 dK 注入差分隐私噪声获得扰动的 dK 序列。其后，Sala 等[8]又设计出 DRC（divide randomize and conquer）算法，通过对加噪之前的 dK 序列进行聚类分组，以降低算法复杂度，但是由于全局敏感度较大，所提出的 $2K$ 图隐私模型实用性比较低。Wang 等[9]结合 dK 图模型与随机矩阵图模型，设计了针对图的 $1K$、$2K$、$3K$ 模型处理算法注入扰动参数，获取网络图的边差分隐私。兰丽辉等[10]针对权重社交网络，结合 dK 图模型设计 LWSPA（protection algorithm based on Laplace noise for weighted social network）对查询结果集中的三元组序列进行分割，再对每个子序列构建满足差分隐私的算法，将查询结果集映射为一个实数向量，通过在向量中注入 Laplace 噪声实现隐私保护。然而，这些方法在数据可用性方面仍显不足，除非隐私参数 ε 被设置成不合理的大值。

通过对上述相关工作的研究与分析，本文提出了基于边介数模型的差分隐私处理方案。在该方案中，选用 dK 图模型来实现图的差分隐私。根据给定的原始网络图，生成对应的 $2K$ 序列；根据边中介中心性系数对 $2K$ 序列进行重新排序；将排序后的序列聚类为一个个子序列，并分组进行扰动[11]以满足差分隐私；将扰动后的各子序列整合成完整的新 $2K$ 序列，并还原成图进行发布。最后，本文将该隐私保护方案与已有经典方案进行比较，实验结果表明，在较高隐私需求的情况下，其隐私保护性和大部分数据可用性都具有一定优势。

3 模型与算法描述

3.1 差分隐私定义

差分隐私是由 Dwork[2]提出的一种隐私保护模型。该模型可以保证即使攻击者获得了所能掌握的最大背景知识，仍无法识别某一条数据记录是否存在于该数据集中。同时，该模型还提出了一种量化分析方法来表示隐私保护强度。

定义 1 若给定 2 个数据集 D_1 和 D_2，有且只有一条记录不同，表示为 $|D_1 \Delta D_2|=1$，则称 D_1 和 D_2 为兄弟数据表。对于随机算法 A，Range(A)用来表示该算法的输出结果集合，即对于 D_1 与 D_2 的算法输出结果 $S \in \text{Range}(A)$，若满足式(1)，则称算法 A 满足 ε–差分隐私。

$$\Pr[A(D_1) \in S] \leqslant e^{\varepsilon} \Pr[A(D_2) \in S] \tag{1}$$

其中，概率 Pr 表示隐私被披露的风险；ε 为调节算法 A 隐私保护强度的参数，即隐私预算参数，ε 越小，数据的安全性越高[12]。

定理 1 非交互输出扰动[13]。设有函数集 F，其敏感度为 $S(F)$，K 为向 F 中每一个函数 f 的输出添加独立噪声的算法。若该噪声为参数值取 $\frac{S(F)}{\varepsilon}$ 的 Laplace 分布，则算法 K 满足 ε–差分隐私。其中，对以兄弟数据表为输入的查询函数，敏感度 $S(F)$为其查询结果的最大曼哈顿距离。

证明 见文献[13]。

定理 2 组合性质[14]。假设有 n 个随机算法，其中 A_i 满足 ε_i–差分隐私，且任意 2 个算法的操作数据集没有交集，则$\{A_i\}(1 \leqslant i \leqslant n)$组合后的算法满足 $\max(\varepsilon_i)$–差分隐私。

证明 见文献[14]。

由定理 1 可知，敏感度 $S(F)$和差分隐私参数 ε 决定了实现差分隐私需要添加的 Laplace 噪声量。

3.2 *dK* 序列及 2*K* 敏感度分析

dK 序列是一组描述图属性的数据集合，它可以将不同细节级别的图结构表示为统计信息。随着 d 值的增加，描述的图属性也更加具体。

对于给定的图 G，0*K* 分布只是图的平均节点度；1*K* 分布则是图的度数分布；2*K* 分布是图的联合度分布，即具有不同节点度组合的 2 节点子图的数目；3*K* 分布表示具有不同节点度组合的 3 节点子图的数目，即聚类系数分布。在极限情况下，*nK* 分布（其中 n 是图中的节点数）可以捕获完整的图结构。图 1 给出一个详细的示例，其中列出了所给图的 2*K* 序列和 3*K* 序列。

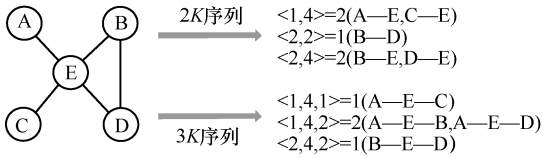


图 1 *dK* 序列示例

dK 模型是一个图构造模型，它利用 *dK* 序列捕获的结构性统计数据集合，生成与原始图结构类似的图。对于给定的图 G，*dK* 图模型可以对其进行分析以产生相应的 *dK* 序列，然后使用相应的生成器来构造使用 *dK* 序列作为输入的合成图。

本文方案主要运用 2*K* 图模型，因此 2*K* 序列敏感度在其中起到至关重要的作用。设图 $G=(V, E)$，其中 V 是节点的集合，E 是连接 V 中节点对的边的集合。2*K* 序列形式上是元组$\{d_x, d_y; k\}$的集合，其中，每个元组表示具有度 d_x 和 d_y 且连接分量为 2 的节点对的数量为 k。设 m 为元组的总数目，$d_{\max}$ 是图 G 中的最大节点度数，有 $m \leqslant \sum_{i=1}^{d_{\max}} i$，且在图 G 为完全图时取得最大值，因此 $m=O(d_{\max}^2)$。

函数的敏感度被定义为当函数域中的一个元素被修改时，函数输出的最大曼哈顿距离。2*K* 序列的域是图 G，G 的邻图是所有与 G 相差至多一条边的图 G'。改变 G 中的一条边将导致在相应的 2*K* 序列中改变一个或多个条目。因此，2*K* 序列的敏感度等于所有 G 的邻图 G'中的 2*K* 序列的最大变化数。

定理 3 2*K* 序列敏感度上限[8]。2*K* 序列的敏感度 S_{2K} 的上限为 $4d_{\max}+1$。

证明 见文献[8]。

由定理 1 可知，对于给定的差分隐私参数 ε，敏感度的上限决定实现差分隐私所必要的噪声添加量，因此在这里只讨论 2*K* 序列敏感度的上限。如果使用更高阶的 *dK* 序列，将造成更高的敏感度值，这必然会降低生成的合成图的精确度。

3.3 基于边介数模型的差分隐私处理方案

3.3.1 方案算法流程

若直接对图 G 的 2*K* 序列进行随机扰动以满足差分隐私，需引入大量的噪声。根据定理 3，任意图 G 上 2*K* 序列的敏感度 S_{2K} 的上限为 $4d_{\max}+1$。为

了显著减少实现给定级别差分隐私所必须添加的噪声量，本文方案根据各元组中边中介中心性系数的平均值大小对 $2K$ 序列进行重新排序，将数据分割成一组组子序列，然后独立地对每个子序列进行扰动以实现 ε−差分隐私。下面本文将证明如果在每个划分出的子序列上实现 ε−差分隐私，在整个 $2K$ 序列也将实现 ε−差分隐私。具体算法流程如算法 1 所示。

算法 1　基于边中介中心性系数的分组差分隐私

输入　由给定图 G 生成的 $2K$ 序列，用 $\{d_x, d_y; k\}$ 表示，子序列数量 n

输出　满足差分隐私的新 $2K$ 序列，用 $\{d'_x, d'_y; k'\}$ 表示

1) 求出 $2K$ 序列的元组数量 q；
2) for i = 1 to q
3) 求出各元组中边中介中心性系数的平均值 $\sum_{i=1}^{k}\frac{C_i^B}{k}$；
4) end for
5) 根据各元组中边中介中心性系数平均值 $\sum_{i=1}^{k}\frac{C_i^B}{k}$ 的大小将 $2K$ 序列重新排序；
6) 将排序后的 $2K$ 序列聚类为一组组子序列；
7) for i = 1 to n
8) 利用 dK 扰动算法向子序列注入噪声；
9) end for
10) 将扰动的子序列合成为新的序列 $\{d'_x, d'_y; k'\}$；

具体的实现过程将在 3.3.2 节~3.3.4 节介绍。

3.3.2　边介数排序算法

边中介中心性系数（edge betweenness centrality）定义为网络中所有最短路径中经过该条边的路径的数目占最短路径总数的比例[15]，即网络中包含成员 i 的所有最短路径数占所有最短路径数的百分比。它表示边 i 的控制能力，其值越大就代表有越多的节点需要通过它才能与其他节点发生联系[16]。对于给定图 G，N 为其节点数目，i 是图 G 中任意一条边，则边 i 的中介中心性系数 C_i^B 的计算式为

$$C_i^B = \frac{1}{(N-1)(N-2)}\sum_{s,t\in G, s\neq t\neq i}\frac{n_{st}(i)}{n_{st}} \tag{2}$$

其中，n_{st} 表示节点 s 与节点 t 之间最短路径的数目，而 $n_{st}(i)$ 表示节点 s 与节点 t 之间通过边 i 的最短路径的数目。

由 3.2 节可知，$2K$ 序列是元组 $\{d_x, d_y; k\}$ 的集合，其中每个元组表示具有度 d_x 和 d_y 且连接分量为 2 的节点对的数量为 k，即度为 d_x 的节点和度为 d_y 的节点之间连接边的数量为 k。对于每个元组，求出其中包含的边的中介中心性系数平均值，并根据中介中心性系数平均值由小到大对 $2K$ 序列重新排序。具体算法流程如算法 2 所示。

算法 2　边中介中心性系数平均值计算

输入　由给定图 G 生成的 $2K$ 序列，用 $\{d_x, d_y; k\}$ 表示

输出　$2K$ 序列的边中介中心性系数平均值

1) 求出 $2K$ 序列的元组数量 q；
2) for i = 1 to q
3) 筛选出符合 $\{d_x, d_y\}$ 度分布的 k 条边；
4) 对于每一条边，根据式(2)求出 C_i^B；
5) 求出该组边中介中心性系数平均值 $\sum_{i=1}^{k}\frac{C_i^B}{k}$；
6) end for
7) 根据边中介中心性系数平均值由小到大对 $2K$ 序列重新排序；

边中介中心性系数可以量化表示某一条边在图中的“重要性”，通过引入边中介中心性系数，对 $2K$ 序列重新排序，将“重要性”相同或相近的边进行聚类分组，在扰动时能够有效地保留原始图的属性与结构特征，使加噪后的图数据具有较好的研究意义。

3.3.3　分组差分隐私噪音添加

将排序后的 $2K$ 序列 R 根据数据集大小划分为 n 个子序列，第 i 个命名为 R_i，$i\in[1,n]$。n 的大小由数据集大小决定，若数据集较大，则选择相对较大的 n；若数据集较小，则选择相对较小的 n。依照下面 2 条规则对 R 序列进行聚类。

1) 每个子序列只能获取 R 序列中的连续元组。

2) 每个元组必须出现且仅可以出现在一个子序列中。

根据上述 2 条规则，可以获得连续且相互不相交的子序列 R_i。由定理 3 可知，这 2 条规则对于子序列敏感度性质是非常重要的。

定理 4　子序列独立性。对于排序后的 $2K$ 序列 R，当 $i\neq j$ 时，R 序列的任何子序列 R_i 与 R_j 的敏感度相互独立。

证明　该定理是基于以下假设利用反证法进

行证明的。假设 R_i 的敏感度受到发生在 R_j 中的变化的影响，$i\neq j$。在不失一般性的前提下，假设 $i<j$，并且 $T(i')$是 R_i 中的元组，$T(j')$是 R_j 中的元组。假设在节点 v_1 和节点 v_2 之间形成一条新边，其中节点 v_1 的相应元组$<T(i),T(i+1),\cdots>\in R_i$且节点 v_2 的相应元组$<T(j),T(j+1),\cdots>\in R_j$。由于此事件可能发生的最大变化次数受 v_1 和 v_2 的度值的限制。令 d_1 为 v_1 的新度值，R_i 中变化的元组的最大数目为 d_1-1 个被删除的元组和 d_1 个添加的元组，即小于 $2d_1$。对称地，令 d_2 为 v_2 的新度数，因此 R_j 中变化的元组的最大数目小于 $2d_2$。即使 d_1 和 d_2 等于它们的子序列中的最大度值 $d_{\max}$，如在定理 3 中规定的，每个子序列中涉及的变化的数目是 $2d_{\max}<4d_{\max}+1$，这意味着 R_i 和 R_j 的敏感度不会相互影响，与该假设相矛盾。

证毕。

利用 *dK* 扰动算法，计算要注入 2*K* 序列的噪声以满足 ε–差分隐私。*dK* 扰动算法是基于 Laplace 分布 Lap(λ)中的随机变量改变 2*K* 序列的每个元组的噪音添加算法。

定义 2 设 DK 是对图 *G* 生成的 2*K* 序列执行差分隐私机制，使 $\mathrm{DK}(G)=dK(\mathrm{G})+\mathrm{Lap}(\frac{S_{2K}}{\varepsilon})^{\mu}$。对于至多相差一条边的任何图 *G* 和 *G*′，如果满足式(3)，则 DK 满足 ε–差分隐私。

$$\left|\ln\frac{\Pr\mid DK(G)=O}{\Pr\mid DK(G')=O}\right|\leqslant\varepsilon \tag{3}$$

其中，S_{2K} 为 2*K* 序列的 Laplace 机制敏感度，*O* 为 DK(*G*)可能的输出，μ 为 2*K* 序列的元组个数。

根据定理 3，若直接对 2*K* 序列 *R* 添加噪声，该噪声为参数取值 $\frac{4d_{\max}+1}{\varepsilon}$ 的 Laplace 分布，且 μ 值为整个 2*K* 序列 *R* 的元组个数。若如算法 2 所述，先对 2*K* 序列进行聚类分组，获得连续且相互不相交的子序列 R_i，则每个子序列的敏感度为 S_{R_i}，再运用定义 2 对每个子序列 R_i 注入参数为 $\frac{S_{R_i}}{\varepsilon}$ 的 Laplace 噪声，这将大幅减小扰动噪声的添加量，理由如下。

设 d_i 为相应子序列 R_i 中节点的最大度数，μ_i 为相应子序列 R_i 的元组个数，则该噪声为参数 $\frac{4d_i+1}{\varepsilon}$，且 μ_i 值为子序列 R_i 的元组个数。由于 $d_i\leqslant d_{\max}$ 且 $\mu_i\leqslant\mu$，因此对噪声添加量 $\mathrm{Lap}\left(\frac{S_{2K}}{\varepsilon}\right)^{\mu}$ 的缩减是指数级别的，这意味着在获得相同隐私保护强度的情况下，聚类操作将很大程度上减小对原始图结构的扰动。

随后利用 *dK* 扰动算法向每个子序列注入噪声使每个子序列都满足 ε–差分隐私，再将加噪后的子序列整合为一个新的 2*K* 序列。下面将证明如果在每个划分出的子序列上实现 ε–差分隐私，在整合之后的 2*K* 序列上也将实现 ε–差分隐私。

定理 5 子序列组合性质。给定 *n* 个满足 ε–差分隐私的不同的 R_i 子序列，$i\in[i,n]$，它们合成的 2*K* 序列 *R* 也满足 ε–差分隐私。

证明 对于 *n* 个满足 ε–差分隐私的子序列 R_i。根据定理 4，当 $i\neq j$ 时，任何子序列 R_i 与 R_j 的敏感度相互独立，即每个 R_i 引入噪声量相互独立。根据定理 2，合成的 2*K* 序列 *R* 满足 ε–差分隐私。

证毕。

3.3.4 2*K* 随机图生成算法

给定无向图 $G=(V, E)$，其节点数 $n=|V|$，边数 $m=|E|$。设 deg(v)为节点 v 的度数，V_k 为度数为 k 的节点集合。

联合度矩阵 **JDM** 定义为

$$\mathbf{JDM}(k,l)=\sum_{v\in V_k}\sum_{w\in V_l}1_{\{(v,w)\in E\}} \tag{4}$$

此矩阵描述连接度为 k 的节点和度为 l 的节点之间的边的数量，2*K* 序列可以容易地由联合度矩阵推导得出。根据同一个联合度矩阵，2*K* 序列可以构造出不止一个在结构和属性上有略微差异的图。

2*K* 随机图生成算法如算法 3 所示。

算法 3 2*K* 随机图生成

输入 联合度矩阵 **JDM**′

输出 满足 **JDM** = **JDM**′的随机图

1) 创建$|V|$个节点，每个节点拥有 deg(v)个可用端口；

2) 对于所有$(k,l)\in$**JDM**′，令 **JDM** $(k,l)=0$；

3) for $(k,l)\in$**JDM**′(k,l)

4) while **JDM** $(k,l)<$**JDM**′(k,l)

5) 任选不相连的 2 个节点 $v\in V_k$ 及 $w\in V_l$；

6) if v 没有可用端口

7) 调用算法 4 NeighborSwitch(v)；

8) end if

9)　　if w 没有可用端口

10)　　调用算法 4 NeighborSwitch(w)；

11)　　end if

12)　　**JDM** (k,l)++；**JDM** (l,k)++；

13)　end while

14)　end for

初始化过程中，创建$|V|=n$ 个节点，分别以它们的度数作为标号。对每个节点 $v\in V$，根据它们各自的度数分配 deg(v)个可用端口。初始状态下全部节点都尚未连接，所以将 **JDM** (k,l)中所有度数对(k,l)的值初始化为 0。随后算法开始进行迭代，在每次迭代中，选择 2 个不相连的节点 v 和 w，度数分别为 k 和 l。当 **JDM** (k,l)没有构造完全时，将 v 的一个可用端口与 w 的一个可用端口连接起来以创建边(v,w)，并且将相应的 2 个项 **JDM** (k,l)和 **JDM** (l,k)的值增加 1。直到当前 **JDM** 的所有条目已经达到其在 **JDM**′(l,k)中的目标值，2K 图构造完成。

2K 随机图生成算法的核心在于：当 **JDM** (deg(v), deg(w))尚未达到其目标值时，即使不相连的 2 个节点 v 和 w 中的任一个或两者都没有可用端口时，也总是可以在它们之间添加一条边。在添加边(v,w)之前，对于没有可用端口的每个节点，对边执行重新布线。这一操作被称为“邻节点转换”。

将没有可用端口的节点定义为饱和节点，将至少有一个可用端口的节点定义为不饱和节点。邻节点转换算法如下。

算法 4　邻节点转换算法（NeighborSwith）

输入　饱和节点 i

输出　生成边(i', j)

1) 选择不饱和节点 i'；

2)　if　deg(i') == deg(i)

3)　选择与 i 相连且与 i'不相连的节点 j；

4)　删除边(i, j)；

5)　添加边(i', j)；

6)　end if

对一个给定节点 i 进行邻节点转换，可以在不改变当前 **JDM** 的情况下为节点 i 释放一个可用端口。由于 deg(i') == deg(i)，邻接点转换保证了 **JDM** (deg(v),deg(w))的值不会改变。

4　实验仿真与性能分析

本节主要通过仿真实验来分析本文方案 BCPA 的隐私保护性和数据可用性，并将其与同样使用差分隐私机制的 dK-PA 方案和 DRC 方案进行比较。其中，dK-PA 模型是直接对 2K 序列进行加噪处理，而另外 2 种模型都是基于 2K 序列排序聚类加噪算法。不同的是，DRC 模型是基于节点度大小对 2K 序列进行排序，而 BCPA 是基于边中介中心性系数进行重新排序。

为了验证本文方案是否适用于社交网络图结构，如表 1 所示，选取 wiki-Vote 和 ego-Twitter 这 2 个真实社交网络数据集作为测试数据集，它们分别是维基百科 who-votes-on-whom 网络数据和 Twitter 社交网络数据。这 2 个数据集都具有以下 2 个显著特征：1) 节点度的统计个数符合幂律分布；2) “小世界”的特征，更接近真实世界。

表 1　　**测试数据集**

名称	节点数	边数
wiki-Vote	7 115	103 689
ego-Twitter	81 306	1 768 149

4.1　隐私保护性评估

为了量化差分隐私保护算法对原始数据图的具体扰动，针对社交网络图结构的特点，本文基于 dK 模型设计了一种基于邻接度的隐私保护性衡量算法，具体如算法 5 所示。

算法 5　隐私保护性衡量算法

输入　原始图 $G(V,E)$，扰动图 $G'(V',E')$

输出　隐私保护性 P

1)　计算原始图中的边个数 m，计算扰动图中的边个数 m'，$D_{\text{diff}}= 2m_{\max}$；

2)　对于节点 u，$u\in G$，找出与 u 所连接的其他节点的度，降序排列并存储在 link(u)=(d_1, d_2, ... , d_n)，其中 n 表示节点 u 的度；

3)　反复执行步骤 2)和步骤 3)，直至遍历所有节点；

4) 对于节点 u'，$u'\in G'$，找出与 u'所连接的其他节点的度，降序排列并存储在 link(u')=(d_1', d_2', ... , $d_{n'}'$)，其中 n'表示节点 u'的度；

5)　if　id(u)==id(u')

6)　　找出 link(u)和 link(u')相同的度关系，每找出一个，则 $D_{\text{diff}}= D_{\text{diff}}-1$；

7)　end if

8)　反复执行步骤 5)~步骤 8)，直至遍历所有节点；

9)　计算隐私保护性 P；

分别生成原始图相邻节点度序列和扰动图相邻节点度序列，比较 2 个序列，得到 2 个序列之间不相同的值的个数 D_{diff}，取原始图的边数量 m 和扰动图的边数量 m'。

2 组数据中边数量的最大值为

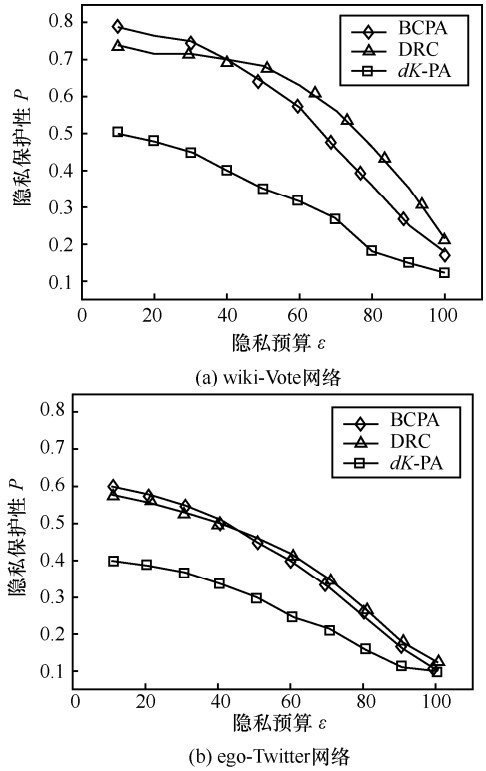
$$m_{\max}=\max(m,m') \tag{5}$$

隐私保护性为

$$P=\frac{D_{\text{diff}}}{2m_{\max}} \tag{6}$$

隐私保护性 P 的取值范围为 0~1，值越大，隐私保护程度越高。当隐私保护性 P 较小时，表示邻接节点的度改变较小，此时攻击者仍然有很大概率实施结构攻击和度攻击。

根据算法 5，实验分别对 BCPA、DRC 和 dK-PA 这 3 种方案进行了仿真，以达到比较 3 种算法的隐私保护效果，如图 2 所示。



(a) wiki-Vote网络

(b) ego-Twitter网络

图 2 隐私保护性变化趋势

由图 2 可以看出，在不同 ε 值和不同大小的数据集中，相比于直接对 $2K$ 序列进行扰动的 dK-PA 方案，聚类加噪的 BCPA 方案和 DRC 方案都表现出较优异的隐私保护效果。这印证了本文中的理论分析，即聚类后加噪可以更有效地利用隐私预算以达到提高隐私保护强度的作用。BCPA 方案和 DRC 方案在不同的 ε 值范围下，隐私保护效果各有优劣。无论是在体量较小的 wiki-Vote 网络中，还是在体量较大的 ego-Twitter 网络中，在 ε 值较小的情况下($\varepsilon\leqslant 40$)，即隐私保护强度较高时，BCPA 方案表现更好；而当 ε 值较大时($\varepsilon\geqslant 50$)，隐私保护强度较弱，此时 DRC 方案的效果更胜一筹。这种结果差异性是排序算法不同造成的，2 种排序算法基于不同的图属性重组 $2K$ 序列，导致在根据加噪后的 $2K$ 序列生成新图时，新图结构的差异在隐私保护性衡量算法中反映出来。值得注意的是，当 ε 值增加到一定值后($\varepsilon>90$)，3 种方案的隐私保护效果差别较小，可见引入噪声量较小时，3 种方案的隐私保护性能差异不大。综上，在隐私预算较小时，BCPA 方案具有更高的隐私保护强度。

4.2 数据可用性评估

对于社交网络图数据而言，数据可用性分析主要是对社交网络的结构特征参数进行分析。本实验通过将原始图与扰动图的特征参数进行比较分析，验证在不同的隐私预算下 BCPA 方案在数据可用性方面的优势。本实验选择平均聚集系数、平均最短路径和平均度分布这 3 个重要参数，用于衡量图结构特征。

由于 $2K$ 序列非常敏感，其需要添加高水平的噪声以提供高级别的隐私保护。然而，非常小的 ε 值需要引入非常大的噪声值，虽然隐私强度很高，但会导致产生与原始图结构极不相似的合成图。当 $\varepsilon<1$ 时，对于较大的图，所需的噪声水平过高，以至 dK 图生成器很难生成与所得到的 dK 分布匹配的合成图。因此，为了取得较好的实验效果，本文分别生成 ε=5、ε=10 和 ε=100 的满足 ε−差分隐私的扰动图。

4.2.1 平均聚类系数

在社交网络图中，平均聚类系数的定义为

$$\overline{C}=\frac{1}{n}\sum_{i=1}^{n}C_i \tag{7}$$

其中，n 为图中节点总数，i 为图中某一节点，C_i 为该节点的聚类系数。网络的平均聚类系数可以很好地表示网络中节点与其相邻节点之间彼此连接的程度，即由 3 条边连接三点形成的子图三角形在当前完整网络中的密集程度。

由图 3 可知，在 ε 值相同的情况下，BCPA 方案生成图的平均聚类系数与原始图更相近，而加噪

的差分隐私生成图都呈现出同样的趋势：即当 ε 值不断减小时，生成图的平均聚类系数逐渐增大。这是由于隐私保护需求的增大，引起更多的噪声边和噪声节点的加入，三角形子图密度增大，导致平均聚类系数值变大。

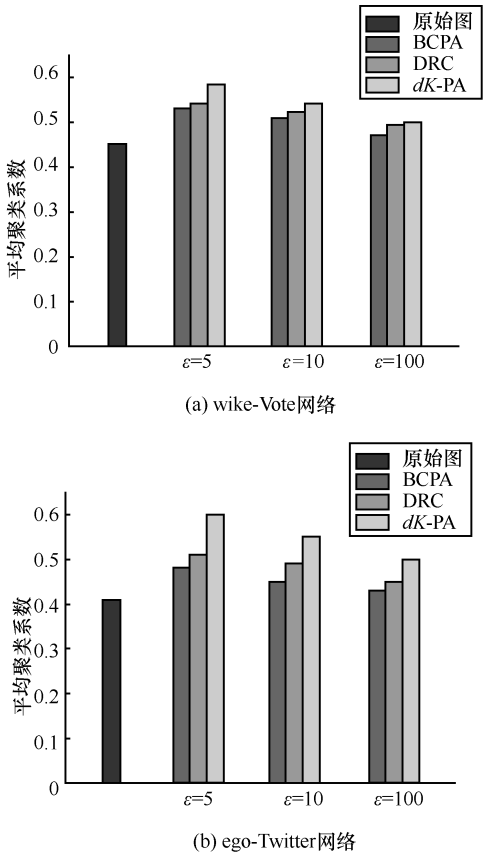


图 3　平均聚类系数柱状图

但即使是在 ε=5 的情况下，2 个体量各异的真实社交网络图与 BCPA 方案生成图之间的差异也没有超过 20%，因此，在保证较高隐私保护强度的情况下，BCPA 较好地保留了原始图属性。

4.2.2　平均路径长度

由于真实社交网络图数据满足“小世界”特征，即图平均路径长度值为 6 左右。通过仿真实验比较 wiki-Vote 网络和 ego-Twitter 网络的平均路径长度值与其满足 ε–差分隐私的扰动生成图的平均路径长度值，其结果如图 4 所示。

图 4 表明，2 个真实社交网络图数据的平均路径长度均符合“小世界”特征，而其差分隐私生成图随着 ε 值的减小，平均路径长度也随之变短。即，随着 ε 值的减小，隐私保护需求不断增大，所需引入的噪声也随之增多，添加了较大数量的扰动边，导致节点之间的连通路径有了更多选择。因此，平均路径长度随着噪声量的增大而不断减小。

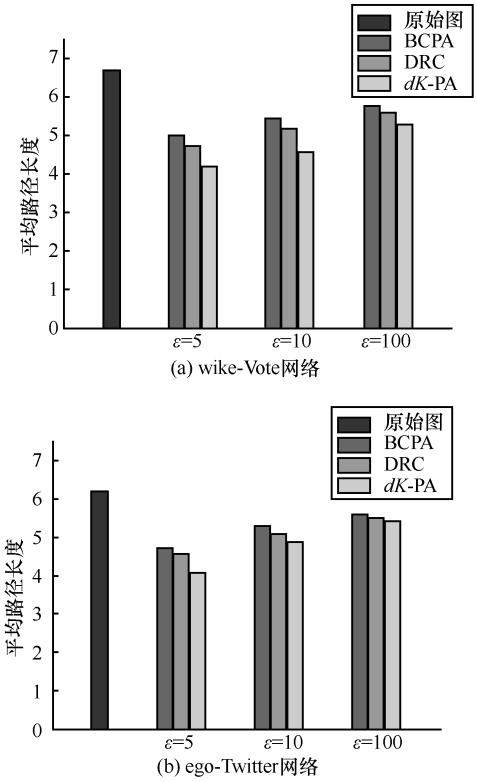


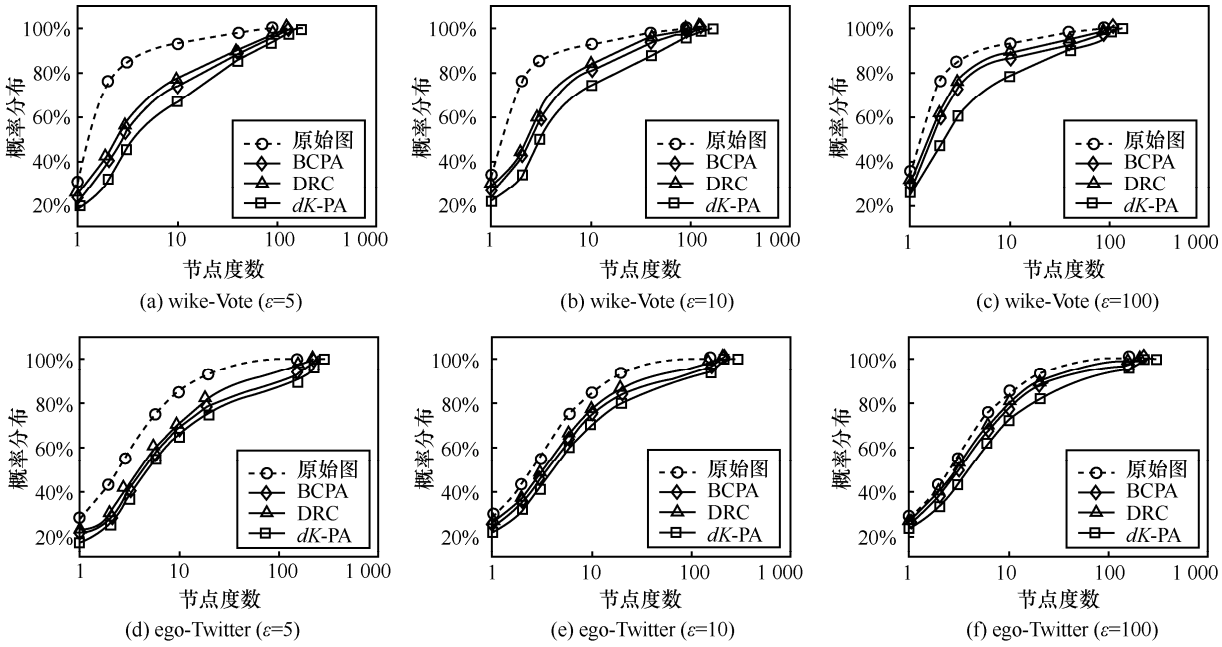
图 4　平均路径长度柱状图

由图 4 可知，对于数据量不同的 2 个社交网络图数据，BCPA 加噪生成图的平均路径长度均更加接近于原始图。因此，在给定隐私预算的情况下，BCPA 能够更好地保留原始图的属性，使扰动数据具有更好的数据可用性。

4.2.3　节点度分布

节点度分布是对一个图中节点度数的总体描述，对于本文方案所采用的 2K 随机图而言，节点度分布就是图中顶点度数的概率分布。节点分布对比如图 5 所示。

图 5(c)和图 5(f)表明，当 ε=100 时，由于所引入的噪声量较小，原始图与 BCPA 和 DRC 的加噪生成图的节点分布相差不大。而当 ε=5 和 ε=10 时，wiki-Vote 网络和 ego-twitter 网络的加噪生成图都与原始图结构有较大差距。出现这种差异的原因是少数高度数节点连接大多数其他节点。因此，当高度数节点引入了噪声时，它会产生向图的其余部分传递波动的结构变化。可见当数据集所含节点数较多时，为实现较好的隐私保护效果，引入的大量噪声显著改变了图结构。



(a) wike-Vote (ε=5) (b) wike-Vote (ε=10) (c) wike-Vote (ε=100)

(d) ego-Twitter (ε=5) (e) ego-Twitter (ε=10) (f) ego-Twitter (ε=100)

图 5 节点度分布对比

此外，由图 5 可知，在节点度分布属性上，BCPA 和 DRC 的效果明显优于 *dK*-PA，且 DRC 方案稍好于 BCPA 方案，但差距不明显。这是由于 DRC 是根据节点度大小对 2*K* 序列进行聚类分组，使加噪生成图的节点度分布更接近原始图。而 BCPA 是根据中介中心性系数对 2*K* 序列进行聚类分组，在原始图中“重要性”相同或相近的节点被聚类在同一子序列中，分组加噪时依然能保证不改变各节点的“重要性”。因此 BCPA 生成的差分隐私图在平均路径长度和平均聚类系数这 2 项重要属性上保留了更多的原始图结构特征。

5 结束语

本文针对基于差分隐私的社交网络图数据发布问题，提出一种基于边介数模型的差分隐私处理方案 BCPA，该方案结合 *dK* 模型对边序列进行聚类划分，同时考虑到各边在图中的影响力程度，引入边中介中心性系数来显著减少 2*K* 序列敏感度。将 BCPA 方案与 DRC 和 *dK*-PA 方案进行了实验仿真，通过对平均聚类系数、平均路径长度以及节点度分布等指标的对比和分析，在较高隐私需求情况下，BCPA 方案的隐私保护性和大部分数据可用性都优于 DRC 方案和 *dK*-PA 方案。下一步可进行以下 2 个方面的工作：1) 在本文所提算法的基础上，进一步改进 2*K* 序列构造算法而提升运行效率；2) 在隐私需求较高的情况下，在引入噪声量增大的同时，保证数据可用性维持在较高水准。

参考文献：

[1] WANG Q, ZHANG Y, REN K, et al. Real-time and spatio-temporal crowd-sourced social network data publishing with differential privacy[J]. IEEE Transactions on Dependable and Secure Computing, 2018, 15(4):591-606.

[2] DWORK C. Differemtial privacy[C]//Proceedings of the 33rd International Colloquium on Automata, Language and Programming(ICALP). 2006:1-12.

[3] CASAS-ROMA J, HERRERA-JOANCOMARTI J, TORRA V. *k*-Degree anonymity and edge selection: improving data utility in large networks[J]. Knowledge & Information Systems, 2016, 50(2):1-28.

[4] HAY M, LI C, MIKLAU G, et al. Accurate estimation of the degree distribution of private networks[C]//Ninth IEEE International Conference on Data Mining. 2009:169-178.

[5] MIR D J, WRIGHT R N. A differentially private graph estimator[C]//IEEE International Conference on Data Mining Workshops. IEEE, 2009:122-129.

[6] LESKOVEC J, FALOUTSOS C. Scalable modeling of real graphs using Kronecker multiplication[C]//Machine Learning, Proceedings of the Twenty-Fourth International Conference. 2007:497-504.

[7] DAY W Y, LI N, MIN L. Publishing graph degree distribution with node differential privacy[C]//ACM International Conference on Management of Data(SIGMOD). 2016:123-138.

[8] SALA A, ZHAO X, Wilson C, et al. Sharing graphs using differentially private graph models[C]//The 11th ACM SIGCOMM Internet

Measurement Conference. 2011:81-98.

[9] WANG Y, WU X. Preserving differential privacy in degree-correlation based graph generation[J]. Transactions on Data Privacy, 2013, 6(2): 127-145.

[10] 兰丽辉, 鞠时光. 基于差分隐私的权重社会网络隐私保护[J]. 通信学报, 2015, 36(9): 145-159.
LAN L H, JU S G. Privacy preserving based on differential privacy for weighted social networks[J]. Journal on Communications, 2015, 36(9): 145-159.

[11] FU Y, CHEN Z, KORU G, et al. A privacy protection technique for publishing data mining models and research data[J]. ACM Transactions on Management Information Systems, 2010,1(1):1-20.

[12] 熊文君, 徐正全, 王豪. 基于滤波原理的时间序列差分隐私保护强度评估[J]. 通信学报, 2017, 38(5): 172-181.
XIONG W J, XU Z Q, WANG H. Privacy level evaluation of differential privacy for time series based on filtering theory[J]. Journal on Communications, 2017, 38(5): 172-181.

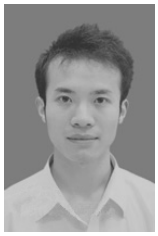
[13] DWORK C, MCSHERRY F, NISSIM K, et al. Calibrating noise to sensitivity in private data analysis[C]//The Third Theory of Cryptography Conference, 2006:265-284.

[14] MCSHERRY F. Privacy integrated queries: an extensible platform for privacy-preserving data analysis[C]//The 2009 ACM SIGMOD International Conference on Management of Data. 2009:19-30.

[15] JAMOUR F, SKIADOPOULOS S, KALNIS P. Parallel algorithm for incremental betweenness centrality on large graphs[J]. IEEE Transactions on Parallel and Distributed Systems, 2018, 29(3):659-672.

[16] MUHONGYA K, MAHARAJ M. Visualising and analysing online social networks[C]//International Conference on Computing, Communication and Security(ICCCS). 2016: 1-6.

[作者简介]



黄海平（1981− ），男，福建三明人，博士，南京邮电大学计算机学院教授、副院长，主要研究方向为物联网安全和数据隐私保护等。

王凯（1994− ），男，江苏扬州人，南京邮电大学硕士生，主要研究方向为物联网安全和数据隐私保护。

汤雄（1992− ），男，江苏沭阳人，南京邮电大学硕士生，主要研究方向为物联网安全和数据隐私保护。

张东军（1993− ），男，江苏徐州人，南京邮电大学硕士生，主要研究方向为物联网安全和数据隐私保护。