

谱半径和特征显著性约束的随机化社会网络方法

许黎明*, 强小强, 宋 转

(浙江师范大学 数理与信息工程学院, 浙江 金华 321004)

(*通信作者电子邮箱 limingx@zjnu.cn)

摘要:为了保护社会网络的安全性, 保证扰动后社会网络的可用性, 提出谱半径和特征显著性(非随机化性)约束的多点扰动社会网络的方法。在扰动社会网络过程中, 将社会网络的谱半径和特征显著性控制在一定的约束范围内, 从而在保证扰动后社会网络的可用性同时, 提高扰动后社会网络的隐私保护程度。理论上分析了该方法的安全性更好, 并给出相应的算法。最后通过实验比较随机化后社会网络的调和平均最短距离、传递系数和特征显著性结构性质的变化情况, 表明该方法能有效地保护社会网络的结构性质, 提高扰动后的可用性。

关键词: 社会网络; 匿名化; 谱半径; 无符号拉普拉斯矩阵; 社会网络的特征显著性

中图分类号: TP393.08 **文献标志码:** A

Random method of social network based on spectral spectrum and feature significant constraints

XU Li-ming*, JIANG Xiao-qiang, SONG Zhuan

(College of Mathematics Physics and Information Engineering, Zhejiang Normal University, Jinhua Zhejiang 321004, China)

Abstract: To protect the security of social network, ensure the availability of social network after perturbation, the paper proposed perturbation method of social network based on the signless Laplacian matrix and the social network non-randomness. In the perturbation process, this method controlled the social network spectral radius and the social network non-randomness by certain constraints, thus ensuring the usability and improving the privacy protection degree of the social network. The paper analyzed the security of this method in theory, and provided corresponding algorithm. At last, the experimental results on comparison of harmonic mean of the shortest distance of the social network, subgraph centrality and the social network non-randomness of change, show that the proposed method effectively protects the structural feature of social network and improving the availability of the social network.

Key words: social network; anonymous; spectral radius; signless Laplacian matrix; significant features of social network

0 引言

随着社会网络(social network)的发展,它在市场预测、社会心理分析、社会疾病分析等领域得到了广泛的应用,并发挥着越来越重要的作用。社会网络的发布和分析已成为数据库和数据挖掘领域研究的热点。然而,社会网络发布与分析会对社会网络中个体的隐私构成威胁。人们非常关注社会网络的隐私保护问题,但是简单地匿名化节点信息,已经不能保证社会网络的安全性。为此文献[1-3]提出了随机化扰动社会网络的方法,该方法的思想是随机添加/删除社会网络的边或随机交换社会网络的边,扰动后的社会网络与原始的社会网络是不同的,从而可以保护社会网络的隐私,但是随机扰动后社会网络的可用性较差。为此,文献[3-4]提出了K-degree匿名模型,通过增加/删除边的操作来构造K-degree匿名图,在K-degree匿名图中任何一个节点至少有K-1个其他节点的度与它相同。K-degree匿名图比随机化扰动模型更能保持原始社会网络的结构性质,但隐私保护程度是1/K,并不是很好,而且很难抵制结构攻击。文献[5]提出的算法产生的图的结构性质与原始图的结构性质非常接近,虽然大大增加了可用性,但同时也增加了隐私泄露,这是因为降低了约束条件。为此文献[6]提出了所有扰动后的图需满足度序列

和某一结构特征属性的约束,从而提高了隐私保护程度,但不能抵制被动攻击。文献[1]提出了基于谱约束的随机化社会网络扰动方法。该方法在谱约束下通过随机添加/删除边(Spectrum Add/Delete)的方法和谱约束的随机交换边(Spectrum Switch)的方法来扰动社会网络。Spectrum Add/Delete的思想是:在满足谱约束的情况下,随机添加两条原本不存在的边,再随机删除两条边,重复K次。类似的, Spectrum Switch的思想是:在满足谱约束的条件下,随机选取两条边 (t, w) 、 (u, v) ,且满足边 (t, u) 、 (v, w) 不存在,则将边 (t, w) 、 (u, v) 交换为 (t, u) 、 (v, w) ,重复K次。基于谱约束的随机化社会网络扰动方法随机化后的社会网络的可用性较好,但是隐私保护程度不高,因为该方法边的扰动比较固定。为此,文献[7]提出了基于无符号拉普拉斯谱扰动交换方法(S-Spectrum Switch),此方法扰动后社会网络的可用性与Spectrum Switch相近,隐私保护较好,这是因为此方法增加了扰动边的方法。目前,社会网络大都包含可随机化的边和非随机化的边,而一般的匿名化扰动技术并没有考虑图的特征显著性在扰动过程中的变化。文献[8]提出了图的特征显著性的度量方法,并说明谱具有一定的欺骗性,在谱不变的情况下,图的特性已经发生了比较大的变化。为此,本文提出图的特征显著性和谱半径约束的扰动方法,此方法可以很好地保

收稿日期:2011-07-14;修回日期:2011-09-10。

作者简介: 许黎明(1966-),男,浙江东阳人,讲师,硕士,主要研究方向:粗糙集、网络安全; 强小强(1985-),男,陕西榆林人,硕士研究生,主要研究方向:智能计算、信息安全; 宋转(1987)女,河南安阳人,硕士研究生,主要研究方向:图像处理。

护社会网络的结构特性,通过分析扰动过程中社会网络的谱半径、传递系数、调和平均最短距离和图的特征显著性的变化来分析扰动后社会网络的安全性和可用性。实验表明此方法可以更好地保护社会网络的隐私,同时提高了扰动后社会网络的可用性。

1 基本理论

1.1 符号表示

一般的社会网络节点代表个体,边代表两个个体之间的关系,这样社会网络就可以用一个图来表示。则社会网络可用图 $G(V, E)$ 表示,节点集 $V = \{1, 2, \dots, n\}$, 边集 $E = \{(i, j) | a_{ij} = 1\}$, 其中 $|E| = m$ 。为方便描述,本文引入以下符号:矩阵 A 表示图 G 的邻接矩阵,其大小为:当节点 i 与 j 有边时, $a_{ij} = a_{ji} = 1$; 相反, $a_{ij} = a_{ji} = 0$ 。对角矩阵 D 表示图 G 的度序列,其中 d_{ii} 表示第 i 个节点的度数,当 $i \neq j$ 时, $d_{ij} = 0$ 。

谱半径是描述图的一个重要结构性性质,反映了图的谱性质。下面给出谱半径的定义:设矩阵 A 为图 G 的邻接矩阵, λ_i 为矩阵 A 的特征值, e_i 是对应于 λ_i 的特征向量,且 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$, 则有 $A = \sum_{i=1}^n \lambda_i e_i e_i^T$, $A e_1 = \lambda_1 e_1$ 。集合 $\{\lambda_1, \lambda_2, \dots, \lambda_n\}$ 叫作图的谱,其中 λ_1 是矩阵的最大特征值,叫作图的谱半径, e_1 为 λ_1 对应的主特征向量,且 $e_1 = (x_1, x_2, \dots, x_n)^T$ 。

矩阵 $Q = D + A$ 表示图 G 的无符号拉普拉斯矩阵。同理, q_i 是矩阵 Q 的特征值,且 $q_1 \geq q_2 \geq \dots \geq q_n$, 集合 $\{q_1, q_2, \dots, q_n\}$ 叫作图的无符号拉普拉斯谱(也叫拟拉普拉斯谱)。其中 q_1 是无符号拉普拉斯矩阵的最大特征值, $2 \min d_i \leq q_1 \leq 2 \max d_i$ [7-8], d_i 表示节点 i 的度数, μ_1 是 q_1 对应的特征向量,且有 $\mu_1 = (y_1, y_2, \dots, y_n)$ 。

k -维谱空间是前 k 个最大特征值对应特征向量组成的空间。 k -维谱空间下节点 u 的坐标表示如下:

$$\alpha_u \rightarrow \begin{pmatrix} x_{11} & x_{21} & \dots & x_{k1} & \dots & x_{n1} \\ x_{12} & x_{22} & \dots & x_{k2} & \dots & x_{n2} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{1u} & x_{2u} & \dots & x_{ku} & \dots & x_{nu} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{1n} & x_{2n} & \dots & x_{kn} & \dots & x_{nn} \end{pmatrix}$$

其中: x_{ij} ($j = 1, 2, \dots, n$) 表示第 i 个最大特征值对应的特征向量, k 为取前几个最大特征值的个数。

1.2 图的结构特性

目前描述图扰动前后结构性质的变化还没有唯一的标准,本文从三个具有代表性的图的结构特性为指标分析社会网络。

1.2.1 调和平均最短距离

图的调和平均最短距离反映的是图中任意两节点之间最短距离的调和平均值,定义为

$$h = \left\{ \frac{1}{n(n-1)} \sum_{i \neq j} \frac{1}{d_{ij}} \right\}^{-1}$$

其中: d_{ij} 表示节点 i 与节点 j 的最短路径, n 为图中的节点总数。

1.2.2 传递性系数

图的传递性系数(一种聚类系数)指少量节点循环的概率,在大的社会网络中它趋于0,这些小的循环更能反映出社会网络的真实结构性性质。定义为 $C = 3N_{\Delta}/N_3$, 其中 N_{Δ} 表示图中任意三个节点构成三角形的数目, N_3 表示图中任意三个

节点连通的三元组的数量。

1.2.3 特征显著性

图的特征显著性是描述图的结构性质的重要指标,它比社会网络的谱更能反映社会网络的结构性质,定义为 $R_G =$

$$\sum_{(u,v) \in E} R(u,v) = \sum_{i=1}^k \lambda_i, \text{ 其中 } R(u,v) = \alpha_u \alpha_v^T = \sum_{i=1}^k x_{iu} x_{iv} \text{ 为边的特征显著性, } \lambda_i \text{ 为邻接矩阵的第 } i \text{ 个最大特征值。}$$

2 基于图约束的扰动理论

2.1 社会网络的谱约束扰动原理

本文提出图谱和特征显著性约束的多点扰动方法主要是通过改变原始图的特征值来修改图,如删除或增加边。设 \tilde{A} 表示扰动后图的邻接矩阵, $\tilde{\lambda}_1$ 表示 \tilde{A} 的最大特征值,且 $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \dots \geq \tilde{\lambda}_n$ 。本文方法通过图谱扰动原理将图的谱半径控制在一定的范围内,从而可以保证扰动后社会网络的可用性。

定理1 [9] 设 $\lambda_1, \tilde{\lambda}_1$ 分别是图 G 和图 \tilde{G} 的最大特征值,从 G 中删除一条边变为 \tilde{G} , 则有 $\tilde{\lambda}_1 \leq \lambda_1$; 同理有从 G 中增加一条边变为 \tilde{G} , 有 $\tilde{\lambda}_1 \geq \lambda_1$ 。

定理2 [9] 盖尔圆盘定理。设 $A = [a_{ij}]$, 又设 $R_i(A) = \sum_{j=1}^n |a_{ij}| (1 \leq i \leq n)$, λ_i 是矩阵 A 的特征值,对 A 进行扰动,则扰动后的矩阵 $A + E$ 的特征值包含在诸圆盘 $\{z \in C | z - \lambda_i \leq R_i(E)\}$ 中。

定理3 [9] 设 $\lambda_1, \tilde{\lambda}_1$ 分别是图 G 和图 \tilde{G} 的最大特征值,若 $\varepsilon_1 \geq \varepsilon_2 \geq \dots \geq \varepsilon_n$ 是扰动矩阵 E 的特征值,则有 $\varepsilon_n \leq |\lambda_i - \tilde{\lambda}_i| \leq \varepsilon_1$ 。

定理3给出了扰动矩阵特征值的界值变化范围。

2.2 社会网络的谱约束扰动条件分析

随着扰动程度 k 的增加,谱约束的随机化社会网络的结构性质跟着变化,最后逐渐趋于稳定。

无符号拉普拉斯 Non-randomness Spectrum Switch 方法扰动还用文献[1,7,10-11]中给出的约束关系式。

1) 若 $(x_i - x_u)(x_v - x_w) > 0$, 则 $\tilde{\lambda}_1 > \lambda_1$;

2) 若 $(x_i - x_u)(x_v - x_w) < 0$ 且 $\lambda_1 - \lambda_2 > \frac{(x_i - x_u)}{(x_w - x_v)} + \frac{(x_w - x_v)}{(x_i - x_u)}$, 则 $\tilde{\lambda}_1 < \lambda_1$;

3) 若 $(y_i - y_u)(y_v - y_w) > 0$, 则 $\tilde{q}_1 > q_1$;

4) 若 $x_s > x_t$, 则 $\tilde{\lambda}_1 > \lambda_1$;

5) 若 $y_t > y_s$, 则 $\tilde{q}_1 > q_1$;

6) 若 $x_t < x_s$, 则 $\tilde{\lambda}_1 < \lambda_1$;

7) 若 $y_t < y_s$, 则 $\tilde{q}_1 < q_1$;

8) 若 $(y_i - y_u)(y_v - y_w) < 0$ 且 $q_1 - q_2 > \frac{(y_i - y_u)}{(y_w - y_v)} + \frac{(y_w - y_v)}{(y_i - y_u)}$, 则 $\tilde{q}_1 < q_1$ 。

3 社会网络随机化扰动方法

3.1 社会网络中节点的分类方法

一般社会网络中节点分为保守点、中性点和自由点^[8],本文为了更好地表达特征显著性的扰动社会网络将自由点和中性点归为一类,将社会网络分为保守型和自由型子图。如此分类后有两个好处:一是可以提高扰动后社会网络的可用性,因为图的特征显著性能更好地描述图的扰动情况^[5];二是降低了算法的复杂度(3.3节将具体分析)。具体分类方法如下。

输入:给出社会网络图的邻接矩阵 A ;

输出:分类后的子图。

步骤:

1) 计算出所有节点的 k -维谱空间下的坐标 α_u 。

2) 通过前 k 个特征向量将所有节点分为保守型和自由型两类。具体做法是将第 k 个特征向量的分量大于 ε 的节点分为保守类,否则分为自由类,其中 ε 是给定的阈值。也可以通过计算节点的特征显著性,并给定阈值对节点进行分类。

3.2 基于谱半径及图的特征显著性的扰动算法

算法思想:由 3.1 节的分类算法将原始的图分为保守型的节点和自由型节点两个子图,然后在子图内进行扰动。本文首先选择自由节点一簇内边之间的扰动,如自由簇中没有满足条件的要求,再查看保守点一簇。这样扰动后可以更好地保持原始图的结构性质。由于图谱的大小具有一定的欺骗性(例如当图谱的大小并没有发生大的变化时,图的结构却发生了比较大的变化,此时图的特征显著性可以比较准确地描述图的结构变化),为此,本文在谱约束的情况下,引入图的特征显著性及其他一些结构性性质共同描述扰动前后图的结构性质,从而可以更好地说明扰动前后社会网络的变化情况。

下面给出社会网络的无符号拉普拉斯特征显著性扰动方法(Non-randomness Spectrum Switch)扰动社会网络的算法。

输入:社会网络分类后子图的序列和原始图的邻接矩阵 A 、度序列矩阵 D 及扰动程度 K ;

输出:随机化后的图。

计算出原始图的邻接矩阵 A 的特征值与特征向量 $(\lambda_1, \lambda_2, e_1)$,

无符号拉普拉斯矩阵 Q 的特征值与特征向量 (q_1, q_2, μ_1) 。

While ($i < K$)

{ 从子图 G_1 中随机选出一条边 (t, w)

if ($i \% 4 == 0$ && $\tilde{\lambda}_1 - \lambda_1 \geq \|E\|$)

{ 找出自由型子图 G_1 所有满足 $\tilde{\lambda}_1 > \lambda_1$ 且 $\tilde{q}_1 > q_1$ 的边,否则,从保守型子图 G_2 中找出满足条件的节点

从满足以上条件的边中随机选出一条边 (u, v) , 则将边 (t, w) 与 (u, v) 换为 (t, u) 与 (v, w)

}

else if ($i \% 4 == 1$ && $\tilde{\lambda}_1 - \lambda_1 \leq \|E\|$)

{ 找出自由型子图 G_1 所有满足 $\tilde{\lambda}_1 < \lambda_1$ 且 $\tilde{q}_1 < q_1$ 的边,否则,从保守型子图 G_2 中找出满足条件的节点

从满足以上条件的边中随机选出一条边 (u, v) , 则将边 (t, w) 与 (u, v) 换为 (t, u) 与 (v, w)

}

else if ($i \% 4 == 2$ && $\tilde{\lambda}_1 - \lambda_1 \geq \|E\|$)

{ 找出自由型子图 G_1 所有满足 $\tilde{\lambda}_1 > \lambda_1$ 且 $\tilde{q}_1 > q_1$ 的点,否则,

从保守型子图 G_2 中找出满足条件的节点

从满足以上条件的边中随机选出一边 s 则将边 (t, w) 换为 (s, w)

}

else if ($i \% 4 == 3$ && $\tilde{\lambda}_1 - \lambda_1 \leq \|E\|$)

{ 找出自由型子图 G_1 所有满足 $\tilde{\lambda}_1 < \lambda_1$ 且 $\tilde{q}_1 < q_1$ 的点,否则,从保守型子图 G_2 中找出满足条件的节点

从满足以上条件的边中随机选出一边 s 则将边 (t, w) 换为 (s, w)

}

else 随机交换自由型子图 G_1 中的边 $(t, w), (u, v)$ 为 $(t, u), (v, w)$

$i++$

}

3.3 算法复杂度分析

基于谱半径和特征显著性约束的扰动方法首先对图中节点进行了分类,且只在类内部扰动社会网络,从而大大降低了算法的复杂度。该算法每次扰动后都要计算矩阵的最大特征值 $\tilde{\lambda}_1$ 。而计算矩阵特征值的算法复杂度为 $O(n^2)$, 扰动程度为 K 和每次扰动后计算满足条件的边的复杂度为 $O(M)$, 所以算法复杂度为 $O(KMn^2)$, 其中 M 为图 1 中边的条数,当扰动程度 K 给定时就是常数,所以复杂度为 $O(Mn^2)$ 。实际运行时,为优化算法可以每扰动几次计算一次矩阵的最大特征值 $\tilde{\lambda}_1$ 。

3.4 隐私保护分析

本文方法与文献[1]的方法都是从边隐私保护分析社会网络的隐私保护。文献[1]中给出了详细的分析,得出攻击

边 (u, v) 成功的概率为 $p = \frac{1}{c_u c_v}$ 。根据图 1 结合算法分析当

节点集 $\{t, u, v, w\}$ 中每个节点相连错误的边数都大于 1 时,边 (u, v) 存在的可能性是所有与节点 t 、节点 w 、节点 v 和节点 u 相连的错误边的数目的积的倒数,攻击边 (u, v) 成功的

概率为 $p = \frac{1}{c_t c_w c_u c_v}$ (c_i 是与 i 节点关联的错误边的个数)。当

节点集中与某些节点相连错误边的数目为 0 时,攻击边 (u, v) 成功的概率为 $p = \min\left\{\frac{1}{c_u}, \frac{1}{c_t c_u}, \frac{1}{c_t c_w c_u}\right\}$ 。当旋转和交换的

次数比较大时,任何一条边错了,扰动后攻击 (u, v) 都有可能不成功。同理攻击边 (t, u) 成功的概率是 $p =$

$\min\left\{\frac{1}{c_u}, \frac{1}{c_t c_u}, \frac{1}{c_t c_w c_u}, \frac{1}{c_t c_w c_u c_v}\right\}$, 显然比一般的方法要好得多。

说明无符号拉普拉斯特征显著性扰动方法的隐私保护比无符号拉普拉斯谱扰动交换方法(S-Spectral Switch)强。

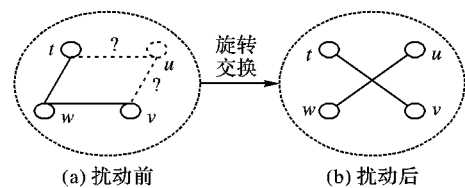


图1 扰动分析

4 实验结果与可用性分析

4.1 实验数据来源描述

本文所使用的实验数据是社会网络分析经常用到的经典的美国政治书籍数据^[1],它包含 105 个节点,441 条边。

4.2 实验结果分析

本文通过3.1节的分类算法将美国政治书籍的数据集进行分类,结果如图2。

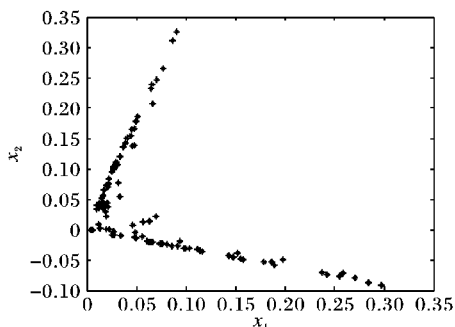


图2 图中节点的分类

分类后图的点大都分布在两条正交线上。取 $\varepsilon = 0.05$ (经验估计值),将 $x_2 > \varepsilon$ 的节点分为保守型一簇, $x_2 < \varepsilon$ 的节点分为自由点一簇,目的是为了可以更好地扰动社会网络,提高扰动后社会网络的可用性。

图3利用传递系数和调和平均最短距离2个指标分析扰动前后社会网络的结构性质,比较无符号拉普拉斯特征显著性扰动方法(Non-randomness Spectral Switch)与无符号拉普拉斯谱扰动交换方法(S-Spectral Switch)的优劣。

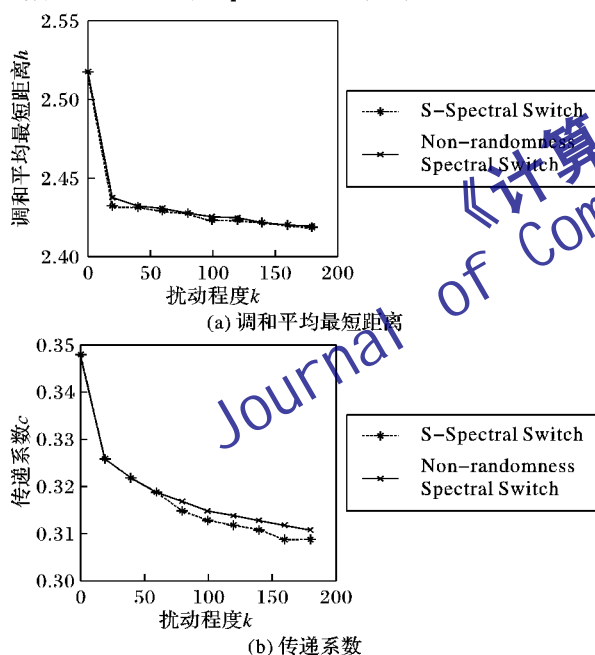


图3 图特性的扰动变化

从图3可以看出,无符号拉普拉斯特征显著性扰动方法和无符号拉普拉斯谱扰动交换方法随着扰动程度 k 的变化, h (调和平均最短距离)和 c (传递系数)的变化基本相似,当 k 增加到80后, h 和 c 的变化趋于平缓。这说明无符号拉普拉斯特征显著性扰动方法比无符号拉普拉斯谱扰动交换方法扰动后更能保护社会网络的结构性质,因此扰动后社会网络的可用性有所提高,这得益于图的特征显著性和分类扰动的优点。

图4从图的特征显著性描述社会网络的扰动效果,比较无符号拉普拉斯特征显著性扰动方法与无符号拉普拉斯谱扰动交换方法的优劣。

从图4的特征显著性可以看出,基于无符号拉普拉斯特

征显著性扰动方法扰动社会网络的方法比无符号拉普拉斯谱扰动交换方法随机化扰动社会网络效果好。基于无符号拉普拉斯特征显著性扰动方法更好地保持原始社会网络的结构性质,从而更好地保护了原始社会网络的结构性质。

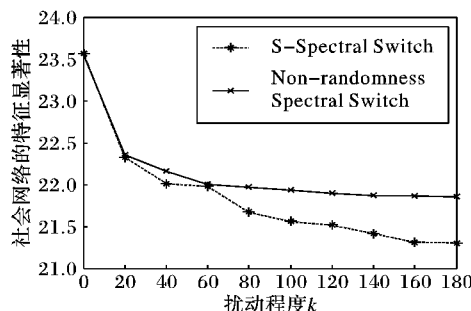


图4 图的非随机化扰动程度的变化

5 结语

本文通过无符号拉普拉斯矩阵对图谱保护随机化社会网络进行了改进。通过谱的界值及特征显著性的约束对社会网络进行随机化扰动,提高了它的可用性和隐私保护程度,这是因为图谱及特征显著性与图的结构性质密切相关。下一步工作可以通过其他拉普拉斯矩阵对图做类似的随机扰动,保护随机化社会网络。另外,进一步改进谱半径的界值也可以改进图谱保护随机化社会网络。

参考文献:

- [1] YING X, WU X. Randomizing social networks: a spectrum preserving approach [C]// SDM 08: Proceedings of the 8th SIAM Conference on Data Mining. Atlanta: SIAM, 2008: 739 - 750
- [2] YING X, WU X. On link privacy in randomizing social networks [C]// PAKDD '09: Pacific-Asia Conference on Knowledge Discovery and Data Mining, LNCS 5476. Berlin: Springer-Verlag, 2009: 28 - 39.
- [3] LIU K, TERZI E. Towards identity anonymization on graphs [C]// SIGMOD 08: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2008: 93 - 106.
- [4] YING X, PAN X, WU X, et al. Comparisons of randomization and k-degree anonymization schemes for privacy preserving social network [C]// SNA-KDD '09: The 3rd Workshop on Social Network Mining and Analysis. New York: ACM, 2009: 96 - 116.
- [5] HANHJÄRVI S, GARRIGA G C, PUOLAMÄKI K. Randomization techniques for graphs [C]// SDM '09: Proceedings of the 9th SIAM Conference on Data Mining. Paris: SIAM, 2009: 780 - 791.
- [6] YING X, WU X. Graph generation with prescribed feature constraints [C]// SDM '09: Proceedings of the 9th SIAM Conference on Data Mining. Sparks: SIAM, 2009: 966 - 977.
- [7] 强小强, 何小卫, 韩建民, 等. 基于谱约束的随机化社会网络多点扰动方法[J]. 计算机工程, 2011, 37(9): 98 - 103.
- [8] YING X, WU X. On randomness measures for social networks [C]// SDM '09: Proceedings of the 9th SIAM Conference on Data Mining. Sparks: SIAM, 2009: 709 - 720.
- [9] 孙继广. 矩阵的扰动分析[M]. 2版. 北京: 科学出版社, 2001.
- [10] CVETKOVI D, ROWLINSON P, SIMIC S. Signless Laplacians of finite graphs [J]. Linear Algebra and Its Applications, 2007, 423 (1): 155 - 171.
- [11] CVETKOVIC D, ROWLINSON P, SIMIC S. Eigenspaces of graphs [M]. Cambridge: Cambridge University Press, 1997.