基于奇异值分解的含权网络匿名化的安全性分析

曾勇，周灵杰，蒋忠元，刘志宏，马建峰
（西安电子科技大学网络与信息安全学院，陕西 西安 710071）

摘 要：分析基于奇异值分解（SVD）的匿名方法在加权社交网络隐私保护中的安全性，给出在含整数权重网络中的重构方法和在含任意权重网络中的非精确重构方法，定义 $\frac{\varepsilon}{N}$-容忍性来衡量其安全性，指出目前谱分析理论得到的ε（可重构系数）上界过于保守因而缺乏指导性。通过实验来测试随机网络、Barabasi-Albert 网络、小世界网络以及实际网络的可重构系数，同时测试了基于 SVD 的双重扰动策略的可重构系数。实验结果表明，加权社交网络对谱的丢失具有不同的容忍性，其容忍性与网络参数之间存在密切的关系。

关键词：奇异值分解；含权社交网络；隐私保护
中图分类号：TP309
文献标识码：A
doi: 10.11959/j.issn.1000-436x.2018074

Security analysis of weighted network anonymity based on singular value decomposition

ZENG Yong, ZHOU Lingjie, JIANG Zhongyuan, LIU Zhihong, MA Jianfeng
School of Cyber Engineering, Xidian University, Xi'an 710071, China

Abstract: The security of anonymous method based on singular value decomposition (SVD) in the privacy preserving of weighted social network was analyzed. The reconstruction method in network with integer weights and the inexact reconstruction method in network with arbitrary weighted were proposed. The $\frac{\varepsilon}{N}$-tolerance was definited to measure its safety. It was also pointed out that the upper bound of ε (the reconfigurable coefficient) obtained in current spectral theories was so conservative that lacks of guidance. The reconfigurable coefficients of random networks, Barabasi-Albert networks, small world networks and real networks were calculated by experiment. Moreover, the reconfigurable coefficients of double perturbation strategies based on SVD were also tested. Experimental results show that weighted social networks have different tolerances on spectrum loss, and there is a close relationship between its tolerance and network parameters.

Key words: singular value decomposition, weighted social networks, privacy preserving

1 引言

目前，大数据已成为信息技术领域的研究热点。但是，大数据的发展依然面临许多问题，安全和隐私保护是目前人们公认的关键问题之一[1]。为了保护用户的隐私信息，数据拥有者一般会在发布数据之前对数据进行匿名化处理。数据发布匿名保护是实现数据隐私保护的核心技术和基本手段。数据的隐私保护要保证数据可用性和安全性之间的平衡。

社交网络的隐私保护不同于关系数据的隐私

收稿日期：2017-10-12；**修回日期**：2018-04-04
通信作者：曾勇，yzeng@mail.xidian.edu.cn
基金项目：国家重点研发计划基金资助项目（No.2016YFB0800601）；国家自然科学基金资助项目（No.U1405255）；111 基地基金资助项目（No.B16037）；中央高校基本科研业务费专项资金资助项目（No.BDZ011402）
Foundation Items: The National Key Research and Development Program of China (No.2016YFB0800601), The National Natural Science Foundation of China (No.U1405255), China 111 Project Foundation (No.B16037), The Central University Basic Business Expenses Special Funding for Scientific Research Projects (No.BDZ011402)

保护，它需要综合考虑网络的图结构、节点间的关系强度（权值）以及节点的重要性等信息。*K*-匿名[2]是关系数据隐私保护的经典方法之一，文献[3,4]将其推广到社交网络的数据隐私保护，分别提出了节点 *K*-匿名和子图 *K*-匿名，它们保证了对目标节点或子图再识别时，至少有 *K* 个候选者，从而使目标隐私泄露的概率小于 $\frac{1}{K}$。数据扰动[5~8]也是社交网络匿名化处理的一类常用方法，它主要对社交网络进行随机修改，使攻击者不能准确地推测出原始的真实数据，例如，文献[6]在图中加入高斯噪声，对网络中边的权值进行扰动，在指定节点之间最短路径及其序列保持不变的情况下，保证其最短路径长度的隐私性。但是这些方法都存在一定的不足，例如，文献[9]明确指出 *K*-匿名方法存在信息损失问题，最重要的一点是它们往往对网络图结构破坏比较大，使扰动后图数据的可用性较低。

为了保护图数据的可用性，研究人员引入人工智能、图谱分析等理论，提出了基于降秩的方法，这些方法在图数据的谱空间可用性上远远优于上述方法。基于特征值分解方法[10]和基于奇异值分解（SVD, singular value decomposition）方法[11]是其中的 2 类经典方法，特征值与奇异值也称为图的谱。这 2 类方法通过舍弃范数小的谱及其对应的谱向量来保证图数据的隐私与谱空间主纬度上的数据可用性。文献[11]中给出了经典的奇异值分解扰动策略和稀疏化的奇异值分解扰动策略，并表明了此类方法能较好地保证图数据的可用性。Wu 等[12]提出用特征值（奇异值）分解对扰动网络进行处理来提高数据的可用性。文献[13,14]将降秩方法引入网络的链路预测中，提出了一种基于低秩矩阵的全局信息预测算法。

基于降秩的方法虽能较好地保护图数据的可用性，但它并不总是安全的，仍然存在数据隐私泄露的风险。由于在基于降秩的方法中涉及舍弃谱，那么舍弃谱的数量是这类方法的关键。目前，在谱分析理论及其他相关理论中对舍弃谱的数量并没有公认的结论。虽然在原网络中舍弃谱越少，扰动网络的数据可用性越好，但在文献[10]对无权网络的特征值分解测试中发现，当舍弃谱的数量较少时，可以从扰动网络中恢复原网络。文献[15]采用文献[10]中所提的方法对城市电路网络的拓扑结构进行了测试，进一步表明了基于降秩的方法并不总是安全的，且其安全性与舍弃谱的数量有关。虽然文献[12]提出了用特征值（奇异值）相似性来决定所舍弃谱的数量，但这样只关注了数据的可用性而忽略了安全性，并且文中实验也表明了此方法会引入新的安全问题。文献[13,14]中给出了舍弃谱的数量的衡量方法，并且实验也表明该方法具有较好的链路预测效果，但其主要针对边缺失的预测问题及网络的动态变化问题所提出。此外，目前对基于降秩的社交网络隐私保护方法的安全性分析主要针对无权网络，而在含权网络中的安全性分析很少。

针对在含权社交网络中基于降秩的隐私保护方法的安全性问题，本文对其进行了分析与测试。目前，降秩的隐私保护方法主要基于奇异值分解和特征值分解，所以本文选择了具有代表性的基于奇异值分解的方法进行分析，并主要分析了奇异值分解扰动和稀疏化的奇异值分解扰动用于加权社交网络数据隐私保护时抵御重构攻击的能力。为了限定舍弃谱的数量，本文提出了 $\frac{\varepsilon}{N}$-容忍性，它表示网络中冗余谱所占的比例。其中，ε为网络的可重构系数，表示在网络可被重构的情况下，可被舍弃谱的最大数量，$\frac{\varepsilon}{N}$ 为网络的可重构比例系数。当算法中舍弃谱的数量小于或等于ε时，网络可被重构，隐私数据将被泄露。本文指出目前的谱分析理论能给出的ε上界过于保守，并进行了大量实验来获取其数值解。实验发现这些网络对丢失的谱具有不同的容忍性，且ε、$\frac{\varepsilon}{N}$ 与网络参数之间存在一定的关系。同时测试了基于 SVD 的双重扰动策略，用于加权社交网络隐私保护时的安全性。实验结果表明，基于 SVD 的扰动策略对信息丢失具有一定的容忍性，直接用于社交网络的隐私保护存在被重构的风险，即使进行了双重扰动，也必须注意舍弃谱的数量。

2 基于奇异值分解的扰动策略

本节主要介绍了在加权社交网络中 2 种典型的基于奇异值分解的扰动策略。

2.1 奇异值分解扰动

文献[11]中作者将人工智能领域的奇异值分解用作数据扰动策略，来保护数据的隐私性。

一个含有 *N* 个节点、*M* 条边的无向加权网络可以表示为 $G=(V,E)$，其中，V 代表节点的集合，E

代表边的集合，顶点数 $N=|V|$，边数 $M=|E|$。若在边集 E 中存在元素 e_{ij}，表示节点 v_i 与节点 v_j 之间存在的边。在加权网络 G 中每一条边 e_{ij} 都有一个权重 $w_{ij}=w_{ji}$ 与之相对应。假设 G 无自环路且无多重边，则 G 可由其邻接矩阵 $\boldsymbol{A}\in R^{N\times N}$ 表示，其中

$$\boldsymbol{A}(i,j)=a_{ij}=\begin{cases}0\ ,\ e_{ij}\notin E\ 且\ e_{ji}\notin E\\ w_{ij},\ e_{ij}\in E\ 或\ e_{ji}\in E\end{cases} \tag{1}$$

对于一个网络 G 的邻接矩阵 $\boldsymbol{A}\in R^{N\times N}$，必然存在一个完全奇异值分解

$$\boldsymbol{A}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} \tag{2}$$

其中，$\boldsymbol{U}$、$\boldsymbol{V}$ 均为 $N\times N$ 阶的酉矩阵，$\boldsymbol{\Sigma}=\mathrm{diag}[\sigma_1,\sigma_2,\cdots,\sigma_{N-1},\sigma_N]$ 为对角矩阵，σ_i 为 $\boldsymbol{A}$ 的奇异值。注：$\boldsymbol{\Sigma}$ 的对角元素非升序排列，即 $\sigma_1\geqslant\sigma_2\geqslant\cdots\geqslant\sigma_{N-1}\geqslant\sigma_N\geqslant 0$。

SVD 扰动主要通过舍弃其部分较小的奇异值及对应的谱向量来实现对数据的扰动。令对角矩阵 $\boldsymbol{\Sigma}_j$ 为舍弃 $j(j\geqslant 0)$ 个较小的奇异值后的对角矩阵，则有

$$\boldsymbol{\Sigma}_j=\mathrm{diag}[\sigma_1,\sigma_2,\cdots,\sigma_{N-j-1},\sigma_{N-j},0,\cdots,0] \tag{3}$$

则 SVD 扰动后的邻接矩阵为

$$\boldsymbol{A}_j=\boldsymbol{U}\boldsymbol{\Sigma}_j\boldsymbol{V}^{\mathrm{T}} \tag{4}$$

2.2　稀疏化的奇异值分解扰动

为了加强对数据的保护，文献[11]在 SVD 扰动的基础上对 $\boldsymbol{U}$、$\boldsymbol{V}$ 矩阵做了进一步处理，并将这种对数据进行双重扰动的方法称为稀疏化奇异值扰动（SSVD，sparsified singular value decomposition）。SSVD 的具体过程如下。

对邻接矩阵 $\boldsymbol{A}$ 进行奇异值扰动，得 $\boldsymbol{A}_j$。然后令

$$\overline{\boldsymbol{U}}(i,j)=\begin{cases}0,|\boldsymbol{U}_k(i,j)|<\alpha\\ \boldsymbol{U}(i,j),|\boldsymbol{U}_k(i,j)|\geqslant\alpha\end{cases} \tag{5}$$

$$\overline{\boldsymbol{V}}(i,j)=\begin{cases}0,|\boldsymbol{V}_k(i,j)|<\alpha\\ \boldsymbol{V}(i,j),|\boldsymbol{V}_k(i,j)|\geqslant\alpha\end{cases} \tag{6}$$

其中，α 是 $\boldsymbol{U}$、$\boldsymbol{V}$ 矩阵中数值的绝对值的下界。则扰动后的矩阵为

$$\overline{\boldsymbol{A}_j}=\overline{\boldsymbol{U}}\boldsymbol{\Sigma}_j\overline{\boldsymbol{V}}^{\mathrm{T}} \tag{7}$$

α 是对矩阵中元素大小的限定，在不同类型的网络中，$\boldsymbol{U}$、$\boldsymbol{V}$ 矩阵的值分布差异可能较大，从而导致在 α 相同的条件下，在 $\boldsymbol{U}$、$\boldsymbol{V}$ 矩阵中舍弃的值的比例无法评估。为了更好地适应不同类型的网络，使 $\boldsymbol{U}$、$\boldsymbol{V}$ 中舍弃的值的数量更可控，本文定义 β 为 $\boldsymbol{U}$、$\boldsymbol{V}$ 中删除的值的比例，即

$$\beta=\frac{n_\alpha}{N\times N} \tag{8}$$

其中，n_α 为 $\boldsymbol{U}$ 或 $\boldsymbol{V}$ 矩阵中舍弃值的个数。

3　重构方法及 $\frac{\varepsilon}{N}$-容忍性

本节主要对无权网络特征值分解的重构方法[10]进行推广，提出在含正整数权重网络中的重构方法。对于含任意权重的社交网络，提出非精确重构的概念以及 2 种重构方法。为了分析降秩方法的安全性，提出了 $\frac{\varepsilon}{N}$-容忍性，它反映了隐私数据何时会被泄露。同时，在 3.4 节对重构方法以及 ε 进行了理论上的分析。

3.1　重构的描述

在 SVD 扰动中，显然，$j=0$ 时 $\boldsymbol{A}_j=\boldsymbol{A}$，$j\neq 0$ 时 $\boldsymbol{A}_j\neq\boldsymbol{A}$。当 $j\neq 0$ 时，$\boldsymbol{A}_j$ 中的值分布服从一定的规律。图 1 给出了在 ER 网络（N=100，p=0.5）中 j=0、j=10、j=20 以及 j=30 时 $\boldsymbol{A}_j$ 中值的分布情况（在其他模型网络中具有相同的结果），其中，$\boldsymbol{A}$ 中的权值服从[1,5]之间的均匀分布。由图 1 可以看出，$\boldsymbol{A}_j$ 中的值在[0,5]中各整数值的附近分布。

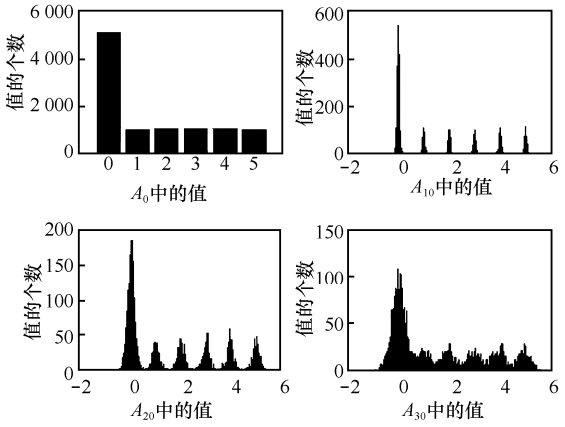

图 1　A_j 中值的分布情况

当 $j\neq 0$ 时，$\boldsymbol{A}_j$ 中的值含有非整数以及负值，本文试图用 $\boldsymbol{A}_j$ 重构 $\boldsymbol{A}$，以获得原网络的全部信息。令重构后的矩阵为 $\overline{\boldsymbol{A}_j}$，假设攻击者无任何背景知识，对于含有正整数权值的网络，定义 $\overline{\boldsymbol{A}_j}$ 为

$$\overline{\boldsymbol{A}_j}(i,j)=\left\lfloor \boldsymbol{A}_j(i,j)+\frac{1}{2}\right\rfloor \tag{9}$$

假设攻击者已知网络权重的分布规律，为了能在 j 取值更大的情况下重构出网络，可以根据网络

中值的分布情况对重构操作进行调整。例如，已知网络中的权值仅为偶数，为取得更佳的还原效果，则重构操作可以调整为

$$\overline{\boldsymbol{A}_j}(i,j)=\begin{cases}\left\lfloor \boldsymbol{A}_j(i,j)\right\rfloor, \left\lfloor \boldsymbol{A}_j(i,j)\right\rfloor \text{为偶数}\\ \left\lfloor \boldsymbol{A}_j(i,j)+1\right\rfloor, \text{其他}\end{cases} \tag{10}$$

3.2 非精确重构

3.1 节所提到的重构方法都是对含整数权重网络的精确重构，而对于含有任意数权重的网络来说，权值可能不全为整数，可能含有多位小数，这时若想精确地重构出原来的网络，代价是非常大的，有时甚至是计算不可行的。考虑到对某些值允许一定误差的存在，本节给出了非精确重构方法。

对于一个加权网络 G，其邻接矩阵为 $\boldsymbol{A}$，若它的重构网络 $\overline{\boldsymbol{A}}$ 与它本身之间满足

$$\frac{\left\|\boldsymbol{A}-\overline{\boldsymbol{A}}\right\|_{\mathrm{F}}}{\left\|\boldsymbol{A}\right\|_{\mathrm{F}}}\leqslant\theta \tag{11}$$

则称 $\overline{\boldsymbol{A}}$ 非精确重构出了 $\boldsymbol{A}$，记为 $\overline{\boldsymbol{A}}\cong\boldsymbol{A}$。其中，$\|\boldsymbol{A}\|_{\mathrm{F}}$ 表示矩阵 $\boldsymbol{A}$ 的 F（Frobenius）范数[16]，定义为

$$\left\|\boldsymbol{A}\right\|_{\mathrm{F}}=\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N}\left|a_{ij}\right|^2} \tag{12}$$

式(11)左边部分称为网络之间的值差[11]，$\boldsymbol{A}-\overline{\boldsymbol{A}}$ 为重构噪声矩阵，它们表示 $\overline{\boldsymbol{A}}$ 相较 $\boldsymbol{A}$ 失真的程度。$0\leqslant\theta\leqslant1$ 为用户阈值，用来限定非精确重构的程度。θ 越小，$\overline{\boldsymbol{A}}$ 与 $\boldsymbol{A}$ 就越相似，当 $\theta=0$ 时，$\overline{\boldsymbol{A}}$ 重构出 $\boldsymbol{A}$，即 $\overline{\boldsymbol{A}}=\boldsymbol{A}$。

本文提出了 2 种非精确重构方法。假设原邻接矩阵 $\boldsymbol{A}$ 及其扰动邻接矩阵 $\boldsymbol{A}_j$，且 $\boldsymbol{A}$ 中权值最多含有 $n_{\max}$ 位小数，则这 2 种方法分别介绍如下。

方法 1 首先将原网络中权值整数化：$\boldsymbol{A}m=m\times\boldsymbol{A}$，其中，$m=10^{n_{\max}}$（注：在小数位数较多的情况下，可以选择牺牲某些位，从而避免 m 太大引起的权值过大、处理困难等问题）。然后利用第 2 节的方法对 $\boldsymbol{A}m$ 进行 SVD 扰动和重构。

方法 2 对重构方法进行调整以适应对权值为小数的匹配。重构方法调整为

$$\overline{\boldsymbol{A}_j}(i,j)=\frac{\left\lfloor\left(\boldsymbol{A}_j(i,j)+\frac{1}{2\times10^{n_{\max}}}\right)\times10^{n_{\max}}\right\rfloor}{10^{n_{\max}}} \tag{13}$$

2 种重构方法本质上是相同的，但由于 2 种方法在对原网络的处理上要求不同，就导致方法 1 必须准确地知道原网络的数据，而方法 2 根据原网络的特性来直接调整重构方法，所以只需要了解原网络的部分特性。方法 1 实质上是权重全部为整数时的非精确重构，所以对 2 种方法的测试可以说明含任意权值的网络的非精确可重构性。

3.3 $\frac{\varepsilon}{N}$-容忍性及可重构系数

当 j 在一定条件下对扰动网络进行重构时，可以重构出原邻接矩阵 $\boldsymbol{A}$，即 $\overline{\boldsymbol{A}_j}=\boldsymbol{A}$。为了衡量网络对信息丢失的容忍程度，本文定义了 $\frac{\varepsilon}{N}$-容忍性。

定义 1 $\frac{\varepsilon}{N}$-容忍性。设网络 G 及其邻接矩阵 $\boldsymbol{A}$，其规模为 N，$\boldsymbol{A}$ 的扰动矩阵为 $\boldsymbol{A}_j$，重构矩阵为 $\overline{\boldsymbol{A}_j}$，令 ε 为

$$\varepsilon=\{\max\{j\}\mid\overline{\boldsymbol{A}_j}=\boldsymbol{A}\} \tag{14}$$

称 ε 为 G 的可重构系数，称 $\frac{\varepsilon}{N}$ 为 G 的可重构比例系数。

在删除网络中 ε 个谱及对应谱向量时，网络依然可以被重构，即 $\frac{\varepsilon}{N}$-容忍性反映了网络中无用谱（或称冗余谱）所占的比例。一个网络中 ε 越大，基于降秩方法中需要舍弃的谱的数量就越大（保证网络不被重构的情况下），则在相同扰动条件下，网络被重构的概率越大。

3.4 ε 取值的分析

对于邻接矩阵 $\boldsymbol{A}$ 的奇异值分解可以表示为

$$\begin{aligned}\boldsymbol{A}&=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}\\&=[u_1,u_2,\cdots,u_n]\begin{bmatrix}\sigma_1&&&\\&\sigma_2&&\\&&\ddots&\\&&&\sigma_n\end{bmatrix}\begin{bmatrix}v_1^{\mathrm{T}}\\v_2^{\mathrm{T}}\\\vdots\\v_n^{\mathrm{T}}\end{bmatrix}\\&=\sum_{i=1}^{n-j}\sigma_iu_iv_i^{\mathrm{T}}+\sum_{i=n-j+1}^{n}\sigma_iu_iv_i^{\mathrm{T}}\end{aligned} \tag{15}$$

对于矩阵 $\boldsymbol{A}$ 中的任意一个元素 a_{ij} 可以表示为

$$a_{ij}=\sum_{i=1}^{n-j}\sigma_i(u_iv_i^{\mathrm{T}})_{ij}+\sum_{i=n-j+1}^{n}\sigma_i(u_iv_i^{\mathrm{T}})_{ij} \tag{16}$$

那么，当满足

$$\left|a_{ij}-\sum_{i=1}^{n-j}\sigma_i(u_iv_i^{\mathrm{T}})_{ij}\right|<\frac{1}{2} \tag{17}$$

可以还原出原数据为

$$a_{ij}=\begin{cases}\left\lfloor\sum_{i=1}^{n-j}\sigma_i(u_iv_i^{\mathrm{T}})_{ij}\right\rfloor, \sum_{i=n-j+1}^{n}\sigma_i(u_iv_i^{\mathrm{T}})_{ij}<\dfrac{1}{2}\\ \left\lfloor\sum_{i=1}^{n-j}\sigma_i(u_iv_i^{\mathrm{T}})_{ij}\right\rfloor, \sum_{i=n-j+1}^{n}\sigma_i(u_iv_i^{\mathrm{T}})_{ij}\geqslant\dfrac{1}{2}\end{cases} \tag{18}$$

由此可见重构的条件为式(17)，即

$$\left|\sum_{i=n-j+1}^{n}\sigma_i(u_iv_i^{\mathrm{T}})_{ij}\right|<\frac{1}{2} \tag{19}$$

考虑到社交网络的无向特性，则 $\boldsymbol{A}$ 为实对称矩阵。设 λ_i 为 $\boldsymbol{A}$ 的特征值，且$|\lambda_1|\geqslant|\lambda_2|\geqslant\cdots\geqslant|\lambda_n|$，由奇异值分解的性质[17,18]可得结论 1。

结论 1　$\sigma_i=|\lambda_i|$，$v_i=\mathrm{sign}(\lambda_i)u_i$，若 $\lambda_i=0$，则 $\mathrm{sign}(\lambda_i)=1$。

由此，式(19)左边的值可以表示为

$$\begin{aligned}\left|\sum_{i=n-j+1}^{n}\sigma_i(u_iv_i^{\mathrm{T}})_{ij}\right|&\leqslant\sum_{i=n-j+1}^{n}\left|\sigma_i\right|\left|(u_iv_i^{\mathrm{T}})_{ij}\right|\\&\leqslant\sum_{i=n-j+1}^{n}\left|\lambda_i\right|\left|(u_iu_i^{\mathrm{T}})_{ij}\right|\end{aligned} \tag{20}$$

又因为 $\boldsymbol{U}$ 为酉矩阵[17]，那么它的 n 个行向量是两两正交的单位向量。由此可知$\|u_iu_i^{\mathrm{T}}\|_2\leqslant1$，继而可得$|(u_iu_i^{\mathrm{T}})_{ij}|\leqslant1$，所以利用文献[10]求解可重构系数的方法，最终可得舍弃奇异值的数量的一个上界为

$$\begin{aligned}\left|\sum_{i=n-j+1}^{n}\sigma_i(u_iv_i^{\mathrm{T}})_{ij}\right|&\leqslant\sum_{i=n-j+1}^{n}\left|\lambda_i\right|\left|(u_iu_i^{\mathrm{T}})_{ij}\right|\\&\leqslant\sum_{i=n-j+1}^{n}\left|\lambda_i\right|\end{aligned} \tag{21}$$

式(18)中的界是比较保守的，因为$(u_iu_i^{\mathrm{T}})_{ij}$ 可能存在负值。同时考虑到大多情况下网络将近一半的特征值为负值，由结论 1 可知，$(u_iu_i^{\mathrm{T}})_{ij}$ 为负值的可能性比较大，这就导致这个界是大于实际值的，而其下确界在矩阵谱分析理论中还没有公认的结论[19]，导致可重构系数的计算误差过大。换句话说，降秩方法中用此理论值作为舍弃的谱数量的依据可以保证网络不被重构，但此值不是最小值。接下来，本文通过实验来计算可重构系数的精确值。

4　网络的可重构系数

$\frac{\varepsilon}{N}$-容忍性反映了算法对网络中谱丢失的容忍程度，ε体现了算法中最少需要舍弃的谱的数量。ε、$\frac{\varepsilon}{N}$ 在一定程度上代表了网络的可重构性和一个网络经过 SVD 扰动后可被重构的概率。

由 3.4 节分析可知，ε理论计算误差较大，故本文通过实验来获取其数值解。4.1 节主要测试了不同网络中的ε以及它与网络参数之间的关系。4.2 节则针对本文提出的非精确重构方法，进行了测试与分析。同时，针对基于奇异值分解的双重扰动方法的安全性，4.3 节测试了在 SSVD 扰动下网络的可重构性和非精确可重构性。为了便于分析，设本节测试的所有网络的权值均为服从$[w_{\min},w_{\max}]$上的均匀分布，那么对含整数权重网络的重构采用式(9)所描述的方法。

本文所有的实验是在 Windows10 专业版上的 Matlab 2015B 中进行的，电脑的配置为 Intel(R) Core™ i5-7500 CPU @3.40 GHz。

4.1　不同网络模型的可重构系数

4.1.1　ER 网络

ER（Erdős-Rényi）网络[20]用随机图的方式来描述网络的拓扑结构，具有易于描述和分析的优点，是复杂网络研究的基本理论之一。本文生成的 ER 无向加权网络采用 $G(N,p)$模型，记为 $G_p(N,W)$，其中，N 为网络的规模。在此类网络中所有节点对之间以概率 p 相连接，且被随机赋予一个权重 w。

首先，本文对 ER 网络中ε的分布情况进行了实验分析。实验所选取的网络规模为 N=200，权值分布参数均为 $w_{\min}$=1、$w_{\max}$=5，p 分别取 0.1、0.3、0.5、0.7 以及 0.9。本文对每种类型的网络取 2 000 个进行重构系数的计算然后求ε的平均值。图 2 显示了 p 取不同值时ε的分布情况，其中离散点是真实数据，曲线是用高斯分布拟合的结果，由此可以得到结论 2。

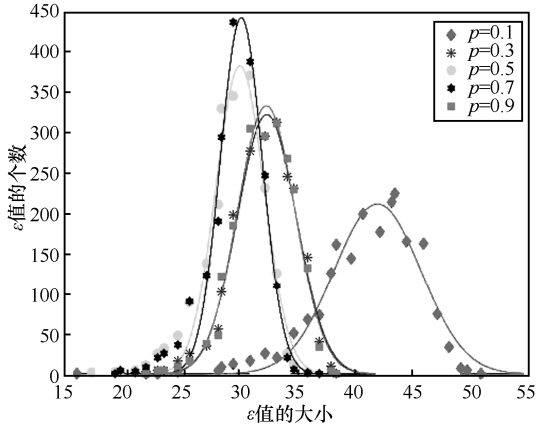

图 2　ER 网络中可重构系数的分布情况

结论 2　ER 网络中ε服从高斯分布。

由结论 2 可知，对于相同类型的网络，其ε值存在差异，那么用其均值ε_m代表一种网络类型的重构系数是否合适，本文将对其进一步分析和测试。图 3 为 $G_{0.5}(N,W)$中ε均值与方差的关系，其中网络参数为 $w_{\min}$=1、$w_{\max}$=5，N 从 50 到 1 000，每隔 50 测试 1 000 组数据计算均值与方差。从图 3 可以看出，ε的方差相较其均值ε_m 很小，且在其他模型网络中也存在同样的规律，由此可以得到结论 3。

结论 3　不能用单次实验结果代表网络的可重构系数，可以用多次实验的均值ε_m 表示。

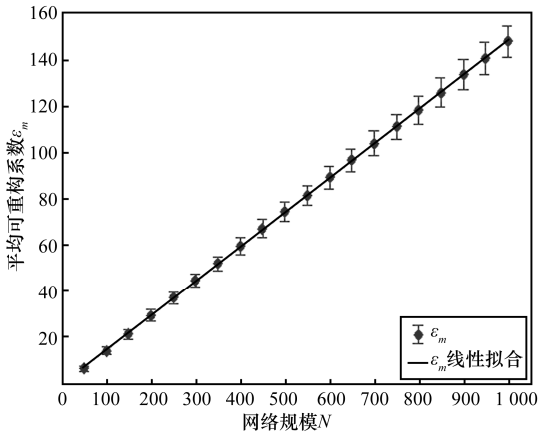

图 3　可重构系数均值与方差的分布

接下来，本文测试了在 ER 网络中网络参数对ε_m的影响。首先，利用 2 种类型的网络 $G_{0.2}(N,W)$与 $G_{0.5}(N,W)$对ε_m与 N 的关系进行了实验分析，其中权值服从[1,3]与[1,5]上的均匀分布。对于每种类型的网络，N 取 50 到 1 000，其中 N 每增加 50 取 200 个网络进行测试。图 4 显示了ε_m随 N 变化的情况，其中实线是拟合的结果。显然，ε_m 与 N 呈线性关系，即 $\frac{\varepsilon}{N}$-容忍性与网络规模无关。同时，网络的稀疏程度对 $\frac{\varepsilon}{N}$ 有一定的影响，但相对较小。而网络中权值的分布却对其有较明显的影响。表 1 中展示了图 4 中 $\frac{\varepsilon}{N}$ 的拟合结果：ε可高达 25%N，说明舍弃 25%的奇异值时，网络依然可能被重构。本文继续分析了权值分布参数（这里指 $w_{\max}$）对 $\frac{\varepsilon}{N}$ 的影响。图 5 是在 $G_{0.5}(200,W)$中测试的结果，其中 $w_{\min}$=1，$w_{\max}$ 取 1 到 20，每隔 1 取 1 000 个网络进行测试。图 5 中离散点是真实数据，曲线是用幂律函数拟合的结果。由图 5 可知，$w_{\max}$ 与 $\frac{\varepsilon}{N}$ 之间存在幂律分布，且 $w_{\max}$ 越小，$\frac{\varepsilon}{N}$ 越大。

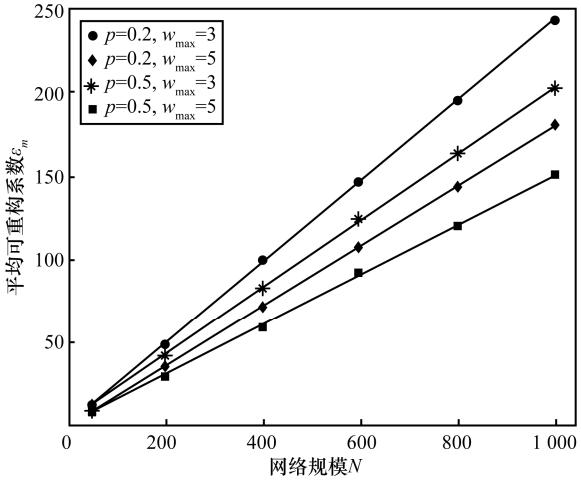

图 4　ER 网络中可重构系数与网络规模的关系

表 1　　**ER 网络的可重构比例系数**

$G_p(N,W)$	$\frac{\varepsilon}{N}$
p=0.2; w=[1,3]	0.257 1
p=0.2; w=[1,5]	0.187 8
p=0.5; w=[1,3]	0.211 2
p=0.5; w=[1,5]	0.153 9

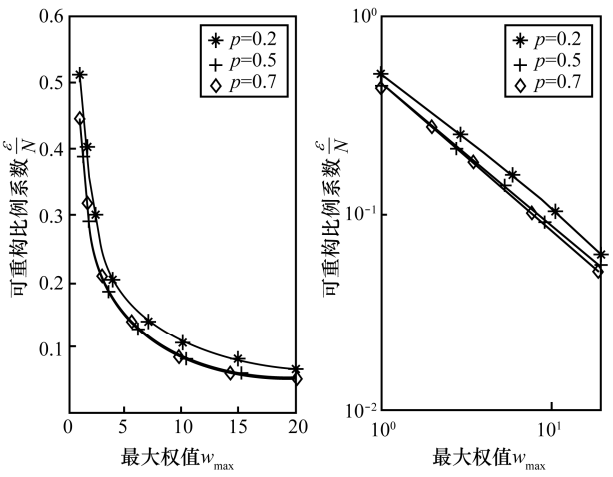

图 5　ER 网络中 $\frac{\varepsilon}{N}$ 与权值分布的关系

图 5 中，当 $w_{\max}$ 为 1 时，网络可视为无权网络，此时，$\frac{\varepsilon}{N}$ 最大，为 45%左右，这与文献[10]在无权网络中的结论基本保持一致，从而证明了本文所测数据的可用性。同时也表明了在无权网络中 SVD 扰动方法的 $\frac{\varepsilon}{N}$-容忍性大，被重构的风险大。

4.1.2　BA 网络

BA（Barabasi-Albert）网络[21]是一种随机无标度网络，它的度分布服从幂律分布。本文用于分析的无标度网络采用 BA 网络模型生成，记为 $G_{(m_0,m)}(N,W)$，其中，m_0 代表初始节点的数量，m 表示每增加一个节点所增加的边数。

本文首先利用 3 种 BA 网络：$G_{(4,4)}(N,W)$、$G_{(8,8)}(N,W)$、$G_{(12,12)}(N,W)$，测试了 ε_m 与 N 的关系，其中，网络的权值分布参数为 $w_{\min}$=1、$w_{\max}$=5，N 从 600 到 1 500，每隔 100 取 200 个网络进行测试。图 6 为测试结果，图 6 中 $\frac{\varepsilon}{N}$ 的拟合结果如表 2 所示。由实验结果可知，ε_m 与 N 之间存在线性关系，而网络参数 m_0、m 对 $\frac{\varepsilon}{N}$ 影响较小。

表 2　　BA 网络的可重构比例系数

网络参数	$\frac{\varepsilon}{N}$
m_0=m=4	0.209 9
m_0=m=8	0.215 7
m_0=m=12	0.220 2

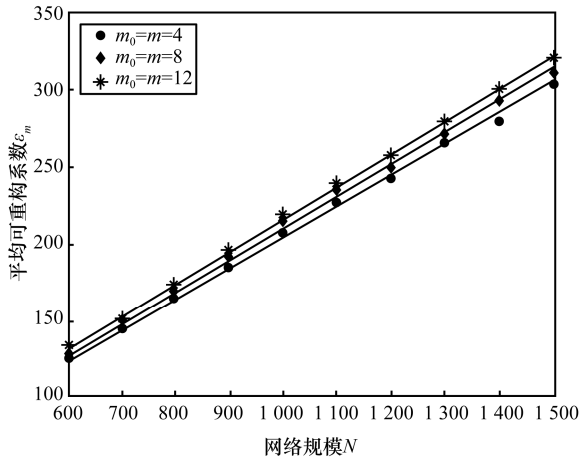

图 6　BA 网络中可重构系数与网络规模的关系

同时，利用这 3 种网络测试了权值分布对可重构系数的影响。图 7 给出了 $w_{\max}$ 与 $\frac{\varepsilon}{N}$ 的关系，其中，N=600，$w_{\min}$ 固定为 1，$w_{\max}$ 取 1 到 20，每隔 1 取 200 个网络进行测试。由图 7 可知，$w_{\max}$ 与 $\frac{\varepsilon}{N}$ 之间同样存在幂律关系，且 $w_{\max}$ 越小，$\frac{\varepsilon}{N}$ 越大。当 $w_{\max}$=1，邻接矩阵 $\boldsymbol{A}$ 可视为无权网络，$\frac{\varepsilon}{N}$ 最大为 50%，网络被重构的风险最大。

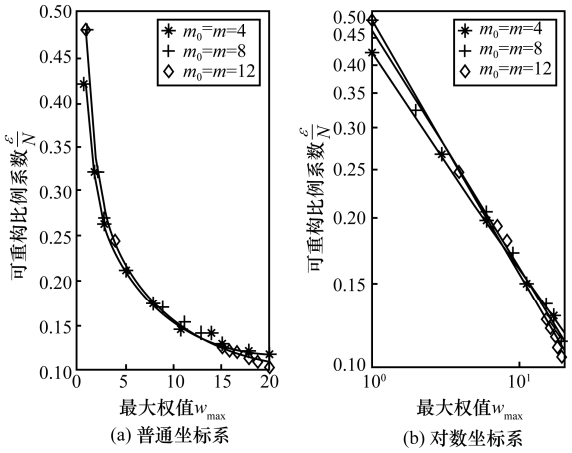

图 7　BA 网络中 $\frac{\varepsilon}{N}$ 与权值分布的关系

4.1.3　小世界网络

小世界网络[22]是介于随机网络和规则网络之间的网络。WS（Watts and Strogatz）模型首先构造一个度为 2×k 的环形规则网络，然后再以概率 p 将边打乱重连。将 WS 模型生成的网络记为 $G_{(k,p)}(N,W)$。

首先，对 ε_m 与 N 之间关系进行了测试。图 8 显示了当 k 为 2、5，p 为 0.2、0.4、0.6 时 WS 网络中 ε_m 与 N 的关系。其中，网络的权值分布参数均为 $w_{\min}$=1、$w_{\max}$=5，N 取 50 到 500，每隔 50 取 1 000 个网络进行测试。表 3 给出了图 8 中 $\frac{\varepsilon}{N}$ 的拟合值。实验发现，在不同的网络参数下，ε_m 与 N 均存在线性关系，并且 $\frac{\varepsilon}{N}$ 的取值取决于 k 和 p 的取值。

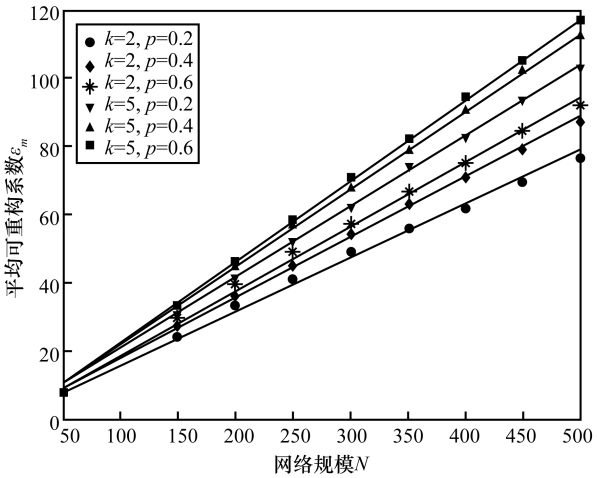

图 8　WS 网络中可重构系数与网络规模的关系

表 3　　WS 网络的可重构比例系数

网络参数		$\frac{\varepsilon}{N}$
k	p	
2	0.2	0.169
2	0.4	0.186
2	0.6	0.199
5	0.2	0.213
5	0.4	0.235
5	0.6	0.243

对于网络中权值分布参数 $w_{\max}$ 对 ε_m 的影响，本文在 $w_{\min}$=1，N=200，k=2、5，p=0.2、0.4 时进行了测试分析，图 9 为实验结果。从图 9 可以看出，$\frac{\varepsilon}{N}$ 与 $w_{\max}$ 之间依然存在幂律关系，且 $w_{\max}$ 越小，$\frac{\varepsilon}{N}$ 就越大。

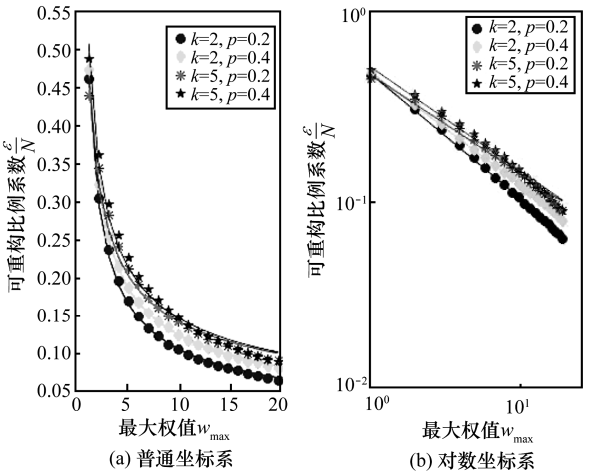

图 9　WS 网络中 $\frac{\varepsilon}{N}$ 与权值分布的关系

4.1.4　实际网络

为了测试实际网络的 $\frac{\varepsilon}{N}$-容忍性，本文对 4 个真实数据集进行了实验，4 个数据集分别为 Facebook[23]、US Airlines[23]、Jazz[24]、Metabolic[23]。它们均为无向网络，包含的节点数分别为 4 039、332、198、453。为了测试权值不同时网络的 $\frac{\varepsilon}{N}$-容忍性，本文为网络中的每条边随机赋予一个服从均匀分布的正整数权值，并测试了 $w_{\min}$ 固定为 1 时 $w_{\max}$ 与 ε_m 的关系，图 10 为实验结果。从图 10 可以

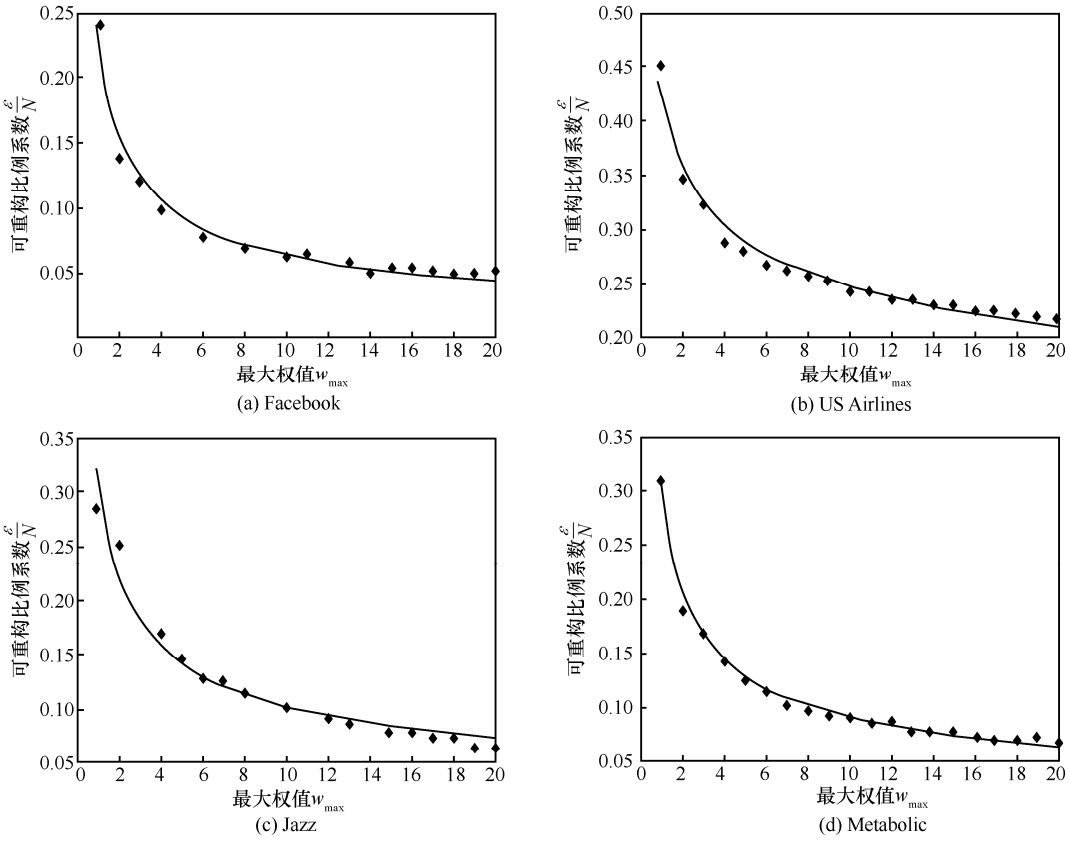

图 10　实际网络中 $\frac{\varepsilon}{N}$ 与权值分布的关系

看出，实际网络对舍弃的维度信息具有不同程度的容忍度，在某些情况下舍弃 45%的维度信息仍不能保证数据的安全性。

本节主要测试了不同网络的$\frac{\varepsilon}{N}$-容忍性。通过分析实验结果，可以得到以下结论。

结论 4　BA、ER 及 WS 网络中，可重构系数与网络规模存在线性关系，可重构比例系数与网络规模无关，即$\frac{\varepsilon}{N}$-容忍性与网络规模无关。

结论 5　BA、ER 及 WS 网络中，当网络中权值为整数且服从[1,$w_{\max}$]上的均匀分布时，$w_{\max}$与$\frac{\varepsilon}{N}$之间存在幂律分布，且 $w_{\max}$越小，$\frac{\varepsilon}{N}$越大。

通过对各种网络$\frac{\varepsilon}{N}$-容忍性的测试发现，SVD 扰动对网络维度的舍弃具有一定的容忍性，甚至在 BA 网络中$\frac{\varepsilon}{N}$可高达 50%，由此可得结论 6。

结论 6　SVD 扰动方法用于含整数权重的社交网络的数据隐私保护并不总是安全的，存在被重构的危险。

4.2　非精确重构中的可重构系数

为测试含任意权重网络的$\frac{\varepsilon}{N}$-容忍性，本文利用 ER 网络对 2 种非精确重构方法中的可重构系数进行了测试。首先测试了ε_m与网络规模 N 的关系，图 11 为实验结果。

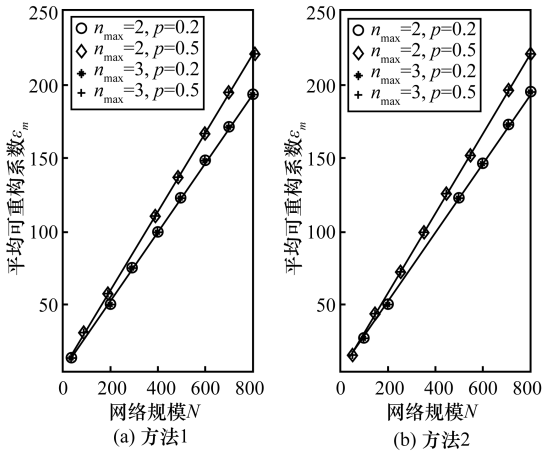

图 11　非精确重构中可重构系数与网络规模的关系

由图 11 可以看出，ε_m与 N 同样存在线性关系，同时也可以看出 $n_{\max}$对重构系数基本没有影响。表 4 给出了可重构比例系数 a 的拟合值，对比方法 1 与方法 2 可以发现，它们的结果是保持一致的。

表 4　　非精确可重构比例系数

网络参数		$\frac{\varepsilon}{N}$	
$n_{\max}$	p	方法 1	方法 2
2	0.2	0.24	0.24
2	0.5	0.26	0.26
3	0.2	0.24	0.24
3	0.5	0.26	0.26

图 12 为利用 ER 网络 $G_p(200,W)$测试$\frac{\varepsilon}{N}$与 $w_{\max}$关系的结果，可以发现在非精确重构中权值的分布对重构系数不再有明显的影响。这是由于非精确重构的定义相对于原网络的值而言并不是绝对的，在ε取值相同的条件下，网络中权值越大，重构噪声矩阵中的值就越大。最后利用 ER 网络 $G_p(200,W)$测试了 θ 与重构比例系数$\frac{\varepsilon}{N}$的关系。

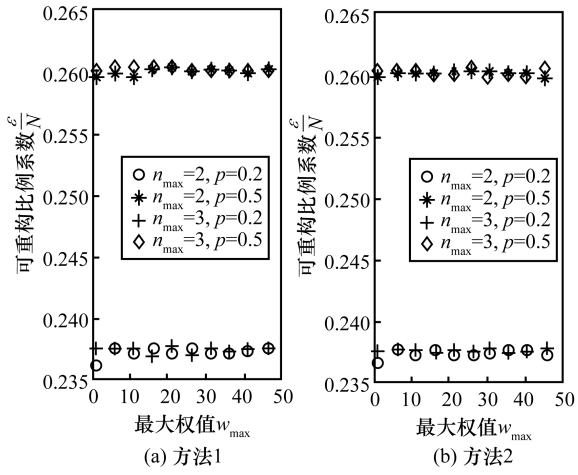

图 12　2 种非精确重构方法下可重构比例系数与权值分布的关系

图 13 是 2 种非精确重构方法下可重构比例系数与 θ 的关系，其中曲线是用幂律分布拟合的结果。由此可看出 θ 与$\frac{\varepsilon}{N}$服从幂律分布。

由 4.1 节的实验可知，在各模型网络及实际网络中，ε与 N 以及$\frac{\varepsilon}{N}$与 $w_{\max}$之间具有相同的规律。本节选取 ER 网络为代表，由实验结果可得以下结论。

结论 7　在非精确重构中，BA、ER 及 WS 网络

中可重构系数与网络规模之间存在线性关系；可重构比例系数不受网络权值分布的影响；非精确重构阈值与可重构比例之间存在幂律关系，且阈值越大，可重构比例越大。

结论 8　SVD 扰动用于任意权值的社交网络隐私保护时，存在被非精确重构的危险。

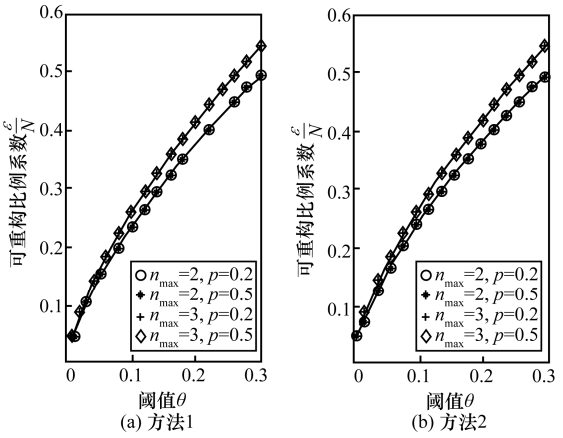

图 13　2 种非精确重构方法下可重构比例系数与 θ 的关系

4.3　双重扰动时网络的可重构系数

对邻接矩阵 $\boldsymbol{A}$ 进行稀疏化奇异值扰动后，在 $j\leqslant\varepsilon$的条件下依然可以利用上面提到的方法重构或非精确重构 $\boldsymbol{A}$，并且ε_m与 N 依然存在线性关系。由 4.1 节可知，在各模型网络中ε与 N 之间存在相同的关系，所以本文利用具有代表性的 ER 网络 $G_{0.5}(N,W)$ 测试了在不同 β 值下ε_m与 N 的关系，其中，N 取 50 到 800，每隔 50 取 100 个网络进行测试。图 14 是在权重全部为整数的情况下对 $\boldsymbol{A}$ 重构的结果，其中网络的权值服从[1,5]上的均匀分布。图 15 是在权重为任意值时对 $\boldsymbol{A}$ 非精确重构的结果，其中，$n_{\max}$=2，且网络的权值均服从(0,5)上的均匀分布，非精确重构阈值 θ=0.1。从图 14 和图 15 可以看出，当 β 高达 16%时，$\boldsymbol{A}_j$ 将不再能重构邻接矩阵 $\boldsymbol{A}$；当 β 高达 30%时，$\overline{\boldsymbol{A}_j}$ 将不再能非精确重构邻接矩阵 $\boldsymbol{A}$。而当 β=4%左右时，$\boldsymbol{A}_j$ 重构邻接矩阵 $\boldsymbol{A}$ 基本不受影响；当 β=6%左右时，$\overline{\boldsymbol{A}_j}$ 非精确重构邻接矩阵 $\boldsymbol{A}$ 基本不受影响，由此可得结论 9。

结论 9　SSVD 用于加权社交网络数据的隐私保护时，存在被重构和被非精确重构的危险。

本节主要在 SVD 扰动下测试了不同含权网络的可重构系数，发现了可重构系数、可重构比例系数与网络参数之间存在的关系。针对含任意权重网络的可重构系数，在 ER 网络中采用非精确重构方法进行测试，发现了可重构系数、可重构比例系数与网络参数之间存在的关系。最后，对基于奇异值分解的双重扰动策略，分别采用重构和非精确重构测试了其可重构系数。实验结果表明，即使进行了双重扰动，网络依然会表现出不同程度的$\frac{\varepsilon}{N}$-容忍性，从而说明了基于奇异值分解的扰动策略用于加权社交网络的隐私保护时具有被重构的危险。

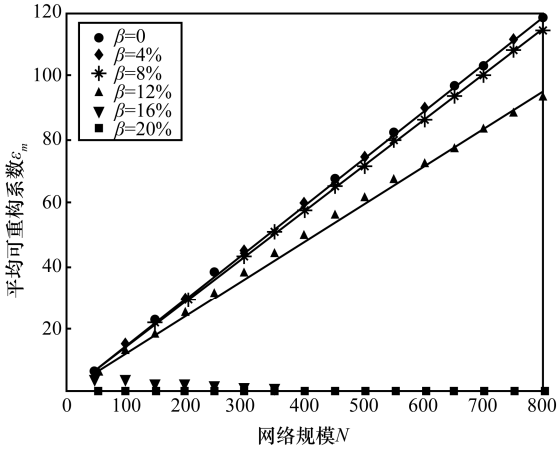

图 14　SSVD 中 β 对网络重构的影响

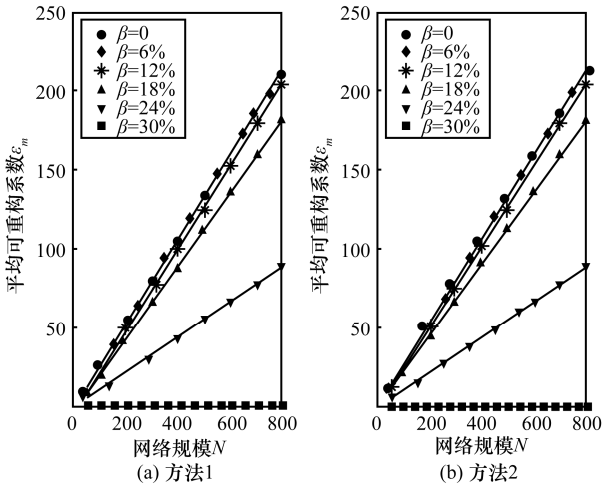

图 15　SSVD 中 β 对非精确重构的影响

5　结束语

本文测试并分析 SVD 扰动方法用于加权社交网络的安全性，提出了网络的$\frac{\varepsilon}{N}$-容忍性，ε越大，网络被重构的概率越大，越不安全，并分别用理论和实验对网络的$\frac{\varepsilon}{N}$-容忍性进行了分析，同时实验

测试了网络非精确重构时的$\frac{\varepsilon}{N}$-容忍性和在 SSVD 扰动下的$\frac{\varepsilon}{N}$-容忍性。实验发现，BA、ER 及 WS 网络中，$\frac{\varepsilon}{N}$-容忍性与网络规模无关，同时表明了 SVD 扰动对含权社交网络信息的丢失具有较大的容忍性，即使在双重扰动中舍弃少于 5%~50%的维度，依然存在被重构和被非精确重构的危险。

本文的分析与测试建立在权值服从均匀分布的基础之上，下一步的工作将考虑非均匀分布权值下可重构系数、可重构比例系数与网络结构之间的关系。同时，在重构方法中考虑攻击者背景知识的影响也值得进一步研究。

参考文献：

[1] 冯登国, 张敏, 李昊. 大数据安全与隐私保护[J]. 计算机学报, 2014, 37(1): 246-258.
FENG D G, ZHANG M, LI H, et al. Big data security and privacy protection[J]. Chinese Journal of Computers, 2014, 37(1): 246-258.

[2] WONG R C W, LI J, FU A W C, et al. (*α*, *k*)-anonymity: an enhanced *k*-anonymity model for privacy preserving data publishing[C]//The 12th ACM SIGKDD international conference on Knowledge discovery and data mining. 2006: 754-759.

[3] CORMODE G, SRIVASTAVA D, YU T, et al. Anonymizing bipartite graph data using safe groupings[J]. The VLDB Journal—The International Journal on Very Large Data Bases, 2010, 19(1): 115-139.

[4] CHENG J, FU A W, LIU J. *K*-isomorphism: privacy preserving network publication against structural attacks[C]//The 2010 ACM SIGMOD International Conference on Management of Data. 2010: 459-470.

[5] ABAWAJY J H, NINGGAL M I H, HERAWAN T. Privacy preserving social network data publication[J]. IEEE Communications Surveys & Tutorials, 2016, 18(3): 1974-1997.

[6] LIU L, WANG J, LIU J, et al. Privacy preserving in social networks against sensitive edge disclosure[R]. Technical Report Technical Report CMIDA-HiPSCCS 006-08, 2008.

[7] DAS S, EĞECIOĞLU Ö, EL ABBADI A. Anonymizing weighted social network graphs[C]//2010 IEEE 26th International Conference on Data Engineering (ICDE). 2010: 904-907.

[8] LI Y, LI Y, YAN Q, et al. Privacy leakage analysis in online social networks[J]. Computers & Security, 2015, 49: 239-254.

[9] 苏洁, 刘帅, 罗智勇, 等. 基于信息损失量估计的匿名图构造方法[J]. 通信学报, 2016, 37(6): 56-64.
SU J, LIU S, LUO Z Y, et al. Method of constructing an anonymous graph based on information loss estimation[J]. Journal on Communications, 2016, 37(6): 56-64.

[10] LIU D, WANG H, VAN M P. Spectral perturbation and reconstructability of complex networks[J]. Physical Review E, 2010, 81(1): 016101.

[11] XU S, ZHANG J, HAN D, et al. Singular value decomposition based data distortion strategy for privacy preserving[J]. Knowledge and Information Systems, 2006, 10(3): 383-397.

[12] WU L, YING X, WU X. Reconstruction from randomized graph via low rank approximation[C]//The 2010 SIAM International Conference on Data Mining, Society for Industrial and Applied Mathematics. 2010: 60-71.

[13] PECH R, HAO D, PAN L, et al. Link prediction via matrix completion[J]. EPL (Europhysics Letters), 2017, 117(3): 38002.

[14] XU X, LIU B, WU J, et al. Link prediction in complex networks via matrix perturbation and decomposition[J]. Scientific Reports, 2017, 7(1): 14724.

[15] SARKAR S. Spectral (re) construction of urban street networks: generative design using global information from structure[M]//Design Computing and Cognition'14. Springer International Publishing, 2015: 41-55.

[16] 张贤达. 矩阵分析与应用(第 2 版)[M]. 北京: 清华大学出版社, 2013.
ZHANG X D. Matrix analysis and applications.(second edi-tion)[M]. Beijing: Tsinghua University Press, 2013.

[17] 戴华. 矩阵论[M]. 北京: 科学出版社, 2001.
DAI H. Matrix theo-ry[M]. Beijing: Science Press, 2001.

[18] HORN R A, JOHNSON C R. 矩阵分析[M]. 天津: 天津大学出版社, 1989.
HORN R A, JOHNSON, C R, Matrix analysis[M]. Tianjin: Tianjin University Press, 2001.

[19] ZHANG X D. Matrix analysis and applications[M]. Cambridge University Press, 2017.

[20] ERDŐS P, RÉNYI A. On the evolution of random graphs[J]. Transactions of the American Mathematical Society, 1984, 286(1): 257-274.

[21] BARABÁSI A L, ALBERT R. Emergence of scaling in random networks[J]. Science, 1999, 286(5439): 509-512.

[22] WATTS D J, STROGATZ S H. Collective dynamics of “small-world” networks[J]. Nature, 1998, 393(6684): 440-442.

[23] ABUELHAIJA S, PEROZZI B, ALRFOU R. Learning edge representations via low-rank asymmetric projections[C]//The 2017 ACM Conference on Information and Knowledge Management. 2017: 1787-1796.

[24] GLEISER P M, DANON L. Community structure in jazz[J]. Advances in Complex Systems, 2003, 6(4): 565-573.

[作者简介]

曾勇（1978-），男，湖南石门人，博士，西安电子科技大学副教授、硕士生导师，主要研究方向为信息安全、无线传感器网络等。

周灵杰（1993-），女，山东滨州人，西安电子科技大学硕士生，主要研究方向为信息安全、社交网络安全等。

蒋忠元（1988-），男，陕西榆林人，博士，西安电子科技大学副教授、硕士生导师，主要研究方向为复杂网络视角下的网络安全、城市计算等。

刘志宏（1968-），男，湖南常德人，博士，西安电子科技大学副教授、硕士生导师，主要研究方向为密码学、信息安全、网络编码、复杂网络、传感器网络等。

马建峰（1963-），男，陕西西安人，博士，西安电子科技大学教授、博士生导师，主要研究方向为信息安全、密码学与无线网络安全等。