因特网流量矩阵的流形结构

钱叶魁[①②] 陈 鸣[①]
[①](解放军理工大学指挥自动化学院 南京 210007)
[②](解放军防空兵指挥学院 郑州 450052)

摘 要：当前，流量矩阵已经被广泛应用于异常检测、流量预测、流量工程等领域，但是现有研究仅仅发现流量矩阵存在线性结构。为了寻找流量矩阵中可能存在的非线性结构，构建流量矩阵模型并从实际因特网骨干网Abilene中采集流量矩阵数据集，应用经典的流形学习算法进行实测数据分析，发现这些高维(81维或121维)的流量矩阵数据集实际上是嵌入的固有维度为5维的低维流形，且其受采样密度和噪声数据等各种因素的影响呈现出不同的结构。

关键词：网络流量分析；流量矩阵；流形学习；非线性降维；流形结构
中图分类号：TP393 **文献标识码**： A **文章编号**：1009-5896(2010)12-2981-06
DOI: 10.3724/SP.J.1146.2010.00130

On the Manifold Structure of Internet Traffic Matrix

Qian Ye-kui[①②] Chen Ming[①]
[①](*Institute of Command Automation, PLA University of Science & Technology, Nanjing* 210007, *China*)
[②](*Air Defence Forces Command Academy of PLA, Zhengzhou* 450052, *China*)

Abstract: Currently, traffic matrices have been applied to anomaly detection, traffic forecasting and traffic engineering widely, but existing researches only find the linear structure of traffic matrix. In order to search the nonlinear structure of traffic matrix, a traffic matrix model is constructed and traffic matrix datasets are collected from real Internet backbone Abilene. Using classical manifold learning algorithms, based on measurement data from Abilene find that these traffic matrix datasets with high dimensionality (81 or 121 dimensions) have a intrinsic dimensionality of 5 and have all kinds of manifold structures in low-dimension embedding space, influenced by sampling density and noise data.

Key words: Network traffic analysis; Traffic matrix; Manifold learning; Nonlinear dimensionality reduction; Manifold structure

1 引言

迄今为止，网络流量分析和建模的大部分研究工作[1,2]都是孤立地研究和揭示单条链路流量具有的时间尺度特性。但是，网络研究者面临的很多问题都迫切需要同时对多条链路或多个源-目的流(OD flow)的流量进行建模和分析，例如流量工程[3]、异常检测[4–6]等。为此，人们提出了流量矩阵模型[7]，并开展了一系列有关流量矩阵时空特性[8,9]的研究。文献[8]利用主成分分析(PCA)方法分析了因特网实测的 OD 流流量矩阵数据的维度和结构，证实其具有低维特性，分析了特征流(eigenflows)的类型和构成，揭示了流量矩阵具有的空间特性。PCA 方法属于线性降维方法，它旨在寻找存在于高维输入空间中的最佳线性低维嵌入空间，然而真实的因特网流量矩阵高维数据本质上是否存在更为复杂的非线性结构值得进一步研究。

本文的目标就是试图发现流量矩阵高维数据中可能存在的流形结构。为此，我们首先描述流量矩阵模型以及 Abilene 实测的流量矩阵数据集，然后应用两种经典的流形学习算法分析 Abilene 实测数据研究流量矩阵高维数据的固有维度(intrinsic dimensionality)和存在的低维流形结构，此外还对影响试验结果的重要因素进行分析。

本文的主要贡献在于通过非线性降维分析 Abilene 实测数据获得以下发现：一是发现 OD 流流量矩阵数据的固有维度为 5 维，远低于高维输入空间的维度(81 维或 121 维)；二是发现 OD 流流量矩阵数据在低维嵌入空间中具有各种不同的流形结构；三是发现获得的低维流形结构受采样密度和噪声数据等因素影响。

2010-02-02 收到，2010-06-14 改回
国家自然科学基金重大研究计划(90304016)，国家 863 计划项目(2007AA01Z418)和江苏省自然科学基金(BK2009058)资助课题
通信作者：钱叶魁 qyk1129@hotmail.com

2 流量矩阵模型和数据集

为了方便后续讨论，首先引入流量矩阵模型的定义和符号，然后描述流量矩阵数据集被测量的网络、流量测量方法以及构造的流量矩阵数据集。

2.1 流量矩阵模型

定义 1 流量矩阵 假设某自治系统(Autonomous System，AS)有 n 个 PoP 点，以一定的时间间隔(周期)连续地被动测量任意一对 PoP 点之间的流量，然后将获得的测量值排列成一个 $T\times p$ 的矩阵 $\boldsymbol{X}$，它表示所有这些流量测量值的时间序列。其中，T 表示测量的周期数，p 表示每个周期内测量获得的流量测量值的个数，即 $p=n\times n$；第 t 行表示在第 t 个周期内流量测量值的向量，通常用 $\boldsymbol{x}_t$ 表示，第 j 列表示第 j 个 PoP 点对之间流量测量值的时间序列。矩阵 $\boldsymbol{X}$ 称为 AS 的 PoP 级流量矩阵，简称为流量矩阵。

本文采用流量特征(源 IP 地址、目的 IP 地址、源端口和目的端口)的熵作为流量测度，因此流量矩阵的任一元素 x_{tj} 表示第 t 个间隔时间内第 j 个 OD 对之间流量特征的熵的取值，其中流量特征的熵定义如下：

定义 2 流量特征的熵[10] 假定随机地观察流量特征 C，观察的样本总数为 S，不同的样本取值的个数为 N，其中流量特征 i 出现了 n_i 次，那么该流量特征的样本熵定义为

$$H(C)=-\sum_{i=1}^{N}(n_i/S)\log_2(n_i/S) \tag{1}$$

其中 $S=\sum_{i=1}^{N}n_i$。

流量矩阵可以看作高维观测数据集 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_T\}$，每个时刻的流量测量数据 $\boldsymbol{x}_t,t=1,\cdots,T$。可以排列成一个 $n\times n$ 的矩阵，本文称之为流量快照，定义如下。

2.2 流量矩阵数据集

本文使用的流量矩阵数据集来自于 Abilene 网络，具体总结见表 1。

3 流形学习算法

流形学习的主要目标就是要揭示蕴含在高维数据中潜在的几何结构，它根据有限的离散样本数据学习和发现嵌入在高维空间中的低维流形。

3.1 ISOMAP 算法

文献[11]提出的 ISOMAP 算法的基本思想就是首先计算流形上的测地线距离，然后应用 MDS 算法，发现嵌入在高维空间的低维坐标，这样 ISOMAP 就通过数据间的测地线距离，保留了数据固有的几何分布结构。下面给出标准 ISOMAP 算法的具体步骤：

第 1 步 构建输入空间 X 中流形 M 上所有数据点 $\boldsymbol{x}_i,i=1,2,\cdots,T$。$\boldsymbol{x}_i\in\mathbb{R}^p$ 的邻接图，距离定义为欧式距离 $d_x(i,j)$，邻接关系定义为 ε 球或 K 最近邻。

第 2 步 通过计算图 G 上两点间的最短路径 $d_G(i,j)$ 估计流形 M 上测地线距离 $d_M(i,j)$，得到的矩阵 $\boldsymbol{D}_G=\{d_G(i,j)\}$ 为图 G 上任意两点间的最短路径距离。

第 3 步 对 D_G 应用 MDS 算法，构建 d 维欧式空间 Y 上的嵌入。

3.2 LLE 算法[12]

局部线性嵌入(LLE)算法是针对非线性数据的一种新的降维方法，基本思想是它认为流形上的每个局部邻域内的任意点都可以描述为邻域内其它点的线性表示，各个邻域之间的连接信息也可以通过相互重叠的部分得以描述，而这个线性关系在映射时保持不变，这样可以把输入数据映射到统一的全局低维坐标系统，并保留邻接特性。

表 1 Abilene 流量矩阵数据集

序号	持续时间	间隔时间(min)	测度	矩阵形式	数据集
1	2009.07.01–07.02	5	源 IP 的熵	576×81[1)]	X_09 (SrcIP)
2	2009.07.01–07.02	5	目的 IP 的熵	576×81	X_09 (DstIP)
3	2009.07.01–07.02	10	源 IP 的熵	576×81	Y_09 (SrcIP)
4	2009.07.01–07.02	10	目的 IP 的熵	576×81	Y_09 (DstIP)
5	2003.12.15–12.21	5	源 IP 的熵	2010×121	X_03 (SrcIP)
6	2003.12.15–12.21	5	目的 IP 的熵	2010×121	X_03 (DstIP)

[1)]由于 Abilene 网络的 3 个 PoP 点在 2009.07.01-07.02 期间的测量数据有误，所以本文仅仅利用其余 9 个 PoP 点的测量数据构建流量矩阵，因此获得的矩阵为 9× 9=81 列。

LLE 算法的具体步骤如下：

第 1 步　计算出每个样本点的 k 个近邻点。把相对于所求样本点距离最近的 k 个样本点规定为所求的样本点的 k 个近邻点，k 是一个预先给定的值。

第 2 步　计算出样本点的局部重建权值矩阵。这里定义一个误差函数：

$$\min \varepsilon(W)=\sum_{i=1}^{T}\left|\boldsymbol{x}_i-\sum_{j=1}^{k}w_j^i\boldsymbol{x}_{ij}\right|^2 \tag{2}$$

其中 $\boldsymbol{x}_{ij}(j=1,2,\cdots,k)$ 为 $\boldsymbol{x}_i$ 的 k 个近邻点，w_j^i 是 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_{ij}$ 之间的权值，且满足：$\sum_{i=1}^{k}w_j^i=1$。

第 3 步　将所有的样本点映射到低维空间中。映射条件满足下式：

$$\min \varepsilon(Y)=\sum_{i=1}^{T}\left|\boldsymbol{y}_i-\sum_{j=1}^{k}w_j^i\boldsymbol{y}_{ij}\right|^2 \tag{3}$$

其中 $\varepsilon(Y)$ 为损失函数值，$\boldsymbol{y}_i$ 是 $\boldsymbol{x}_i$ 的输出向量，$y_{ij}(j=1,2,\cdots,k)$ 是 $\boldsymbol{y}_i$ 的 k 个近邻点，且要满足以下两个条件：

$$\sum_{i=1}^{T}\boldsymbol{y}_i=0,\ \frac{1}{T}\sum_{i=1}^{T}\boldsymbol{y}_i\boldsymbol{y}_i^{\mathrm{T}}=\boldsymbol{I} \tag{4}$$

其中 $\boldsymbol{I}$ 是一个 $T\times T$ 的对称矩阵。这里 $w_j^i(i=1,2,\cdots,T)$ 可以存储在 $T\times T$ 的稀疏矩阵 $\boldsymbol{W}$ 中，当 $\boldsymbol{x}_j$ 是 $\boldsymbol{x}_i$ 的近邻点时 $W_{i,j}=w_j^i$，否则 $W_{i,j}=0$。要使损失函数达到最小，则取 $\boldsymbol{Y}$ 为 M 的最小 m 个非零特征值对应的特征向量。

4　试验分析

4.1 维度分析

对流量矩阵进行维度分析是寻找流量矩阵固有维数的重要途径，对于建立有意义的低维嵌入空间，揭示流量矩阵内在的几何结构具有重要意义。

利用 3.2 节中 2 种流形学习算法对表 1 中数据集 1~2 和 5~6 进行非线性降维。图 1 给出了应用 ISOMAP 算法得到的维数和残余方差之间的关系。可以看出，当维数大于 5 时，随着维数的增加，残余方差减少变得较为缓慢。此外，以捕获总方差 85% 的主成分数作为高维数据的固有维度，可以获得类似结论。因此，可以认为这些流量矩阵的固有维度为 5 维。LLE 算法也获得同样的结论，限于篇幅，这里就不画出维数和残余方差之间的关系图。

流量矩阵高维数据为什么具有低维度特性呢？这主要是由于流量的空间相关性 (spatial correlation)。众所周之，因特网分为核心网络和边缘网络，流经因特网核心骨干网络(如 Abilene 网络)不同入口和出口的流量可能来源于相同区域的边缘网络，这使得在不同的 OD 对的流量之间存在部分共同的变化模式。

4.2 结构分析

为了揭示流量矩阵高维数据的低维流形结构，我们应用第 3.2 节中 2 种流形学习算法分别对表 1 中数据集 1~2 和 5~6 进行非线性降维，由 4.1 节可

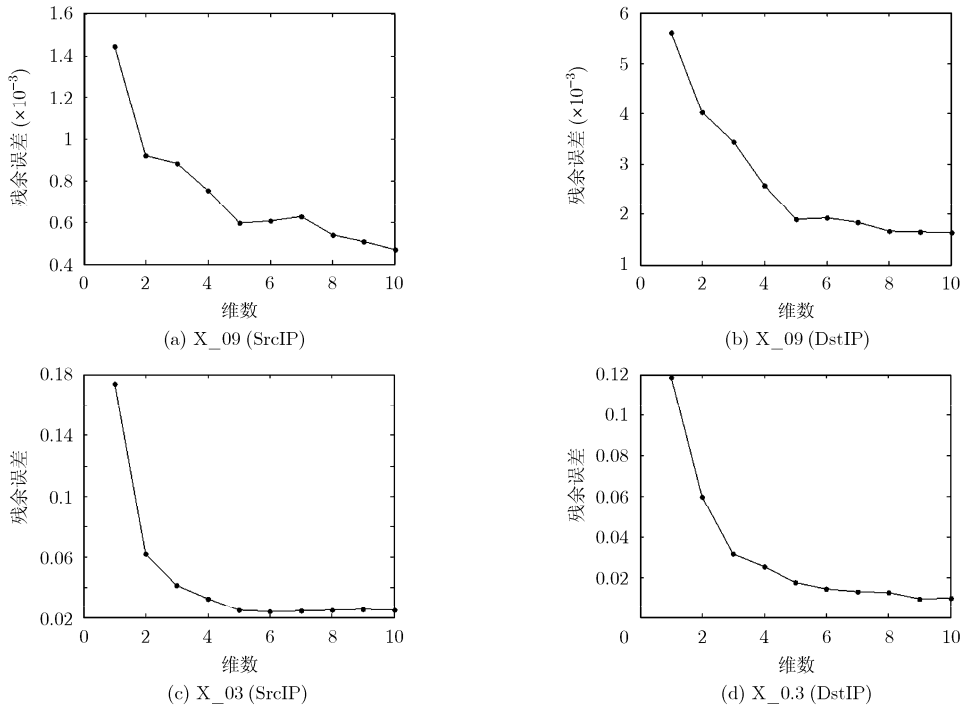


图 1 应用 ISOMAP 降维得到的维数和对应的残余方差

知，数据集的固有维度均为 5，由于无法显示 5 维空间，我们在 3 维空间中观察这些数据集在低维嵌入空间中的几何结构，如图 2~图 5 所示。可以看出，两种流形学习算法对各种不同测度的流量矩阵高维数据进行非线性降维得到的在 3 维空间中的几何结构都具有一定形式的流形结构。需要注意的是，由于大部分数据集的固有维度为 5 维，而我们仅仅在 3 维空间中显示其低维嵌入空间中的几何结构，因此在某些情况下会出现结构退化的现象(例如呈现的几何结构为一条直线或若干个点)。从以上的分析，可以认为在流量矩阵高维数据中蕴含着更为本质的非线性流形结构，而这些结构是线性降维算法无法获得的。

此外，两种流形学习算法分析同一数据集得到不同的流形结构，但是流量矩阵高维数据中蕴含的低维流形结构从本质上来说应该是唯一的，因此这些流形学习算法仅仅是从不同的角度对这一本质结构的逼近。

利用 ISOMAP 算法对表 1 中数据集 X_09 (SrcIP)进行非线性降维，观察流量矩阵高维数据在 2 维空间中的几何结构，如图 6 所示。可以看出，在图 6 中随着横坐标的递增，2 维空间中的样本点形成平滑的弧线，它们对应流量的某种平滑的变化。

类似地，文献[12]通过实验证实人脸图像数据的低维流形能够刻画人脸表情(喜、怒、哀、乐)的变化；文献[13]通过实验证实人脑图像可以用低维流形表示。

4.3 讨论

应该指出，流形学习是建立在数据点可被看做是采样于一个潜在的低维流形这一假设之上。有以下几个因素可能会对试验结果产生影响：

一是采样密度。对于观察维数较高的空间中，如果没有足够的数据，样本的分布一般都相当稀疏，下面通过试验来分析采样密度对分析结果的影响。分别对表 1 中数据集 1 和数据集 3 进行非线性降维，观察这些高维数据在 3 维空间中的几何结构，分别见图 2 和图 7。可以看出，在不同的采样密度情况下，流形学习获得的低维流形结构有很大的不同。

二是噪声数据。在实际的流量测量过程中，经常会引入各种噪声数据。下面同样通过模拟试验来分析噪声数据对分析结果的影响。我们从第 100 个时间间隔至 110 个时间间隔在数据集 Y_09 (SrcIP)的第 1~2 个 OD 流中引入噪声数据，再利用流形学习算法对生成的高维数据进行降维分析，其在 3 维空间中的几何结构如图 8 所示。对比图 7 和图 8，可以看出，噪声数据对学习到的低维流形结构产生很大影响。为此，我们下一步计划利用更具拓扑稳定性的流形学习算法[14,15]来寻找流量矩阵高维数据中蕴含的低维流形结构。

5 结束语

本文应用各种经典的流形学习算法对 OD 流流量矩阵高维数据进行非线性降维分析，试图发现蕴含在其中的非线性结构。通过因特网实测数据分析，我们发现 OD 流流量矩阵的固有维度为 5 维，远低于高维输入空间的维度(81 或 121 维)；OD 流流量矩阵在低维嵌入空间中具有各种不同的流形结构；此外，还发现低维流形结构受采样密度和噪声数据等因素影响。下一步我们将利用 OD 流流量矩阵具

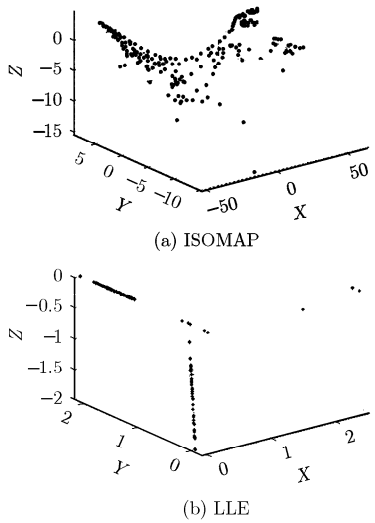


图 2 数据集 X_09(SrcIP)的流形学习 3 维可视化结果

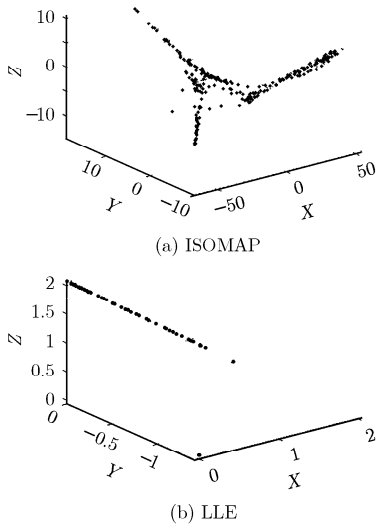


图 3 数据集 X_09(DstIP)的流形学习 3 维可视化结果

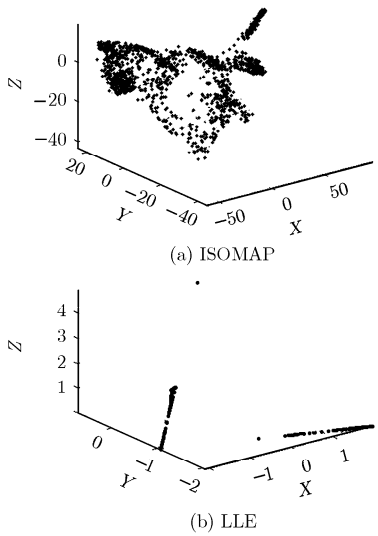
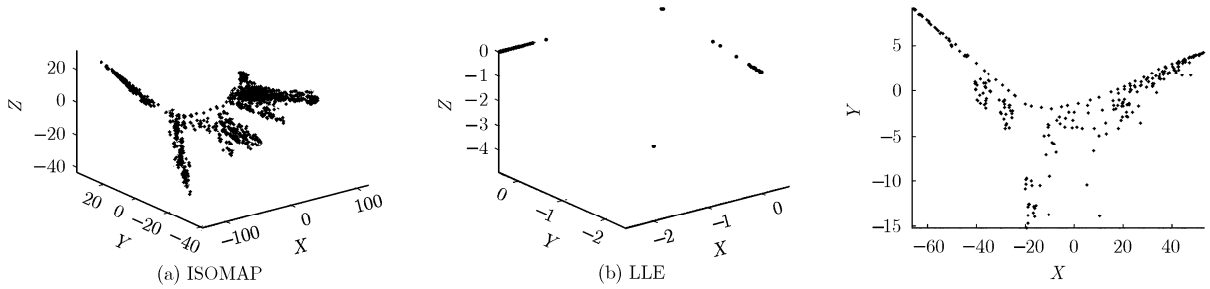


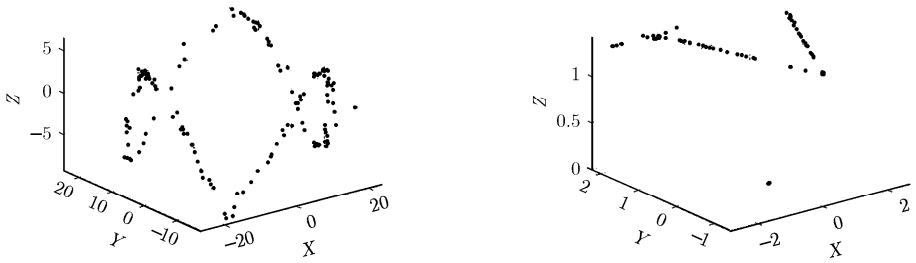
图 4 数据集 X_03(SrcIP)的流形学习 3 维可视化结果



(a) ISOMAP　(b) LLE

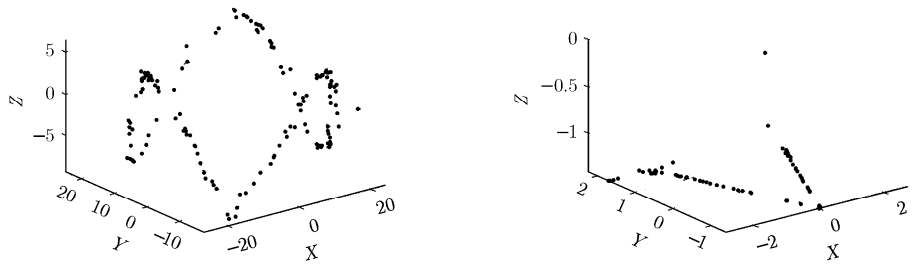
图 5 数据集 X_03 (DstIP)的流形学习 3 维可视化结果

图 6 数据集 X_09 (SrcIP)的流形学习 2 维可视化结果(ISOMAP)



(a) ISOMAP　(b) LLE

图 7 数据集 Y_09 (SrcIP)的流形学习 3 维可视化结果



(a) ISOMAP　(b) LLE

图 8 数据集 Y_09 (SrcIP)在引入噪声数据后的流形学习 3 维可视化结果

有的这种低维流形结构进行全网络异常检测。除此以外，其它网络特性(如流量大小、时延、丢包率、可用带宽等)是否具有类似的规律也值得进一步研究。

致谢　本文研究得到了南京航空航天大学陈松灿教授的指导和帮助，在此深表谢意！

参 考 文 献

[1] Leland W, Taqqu M, and Weland W, *et al*.. On the self-similar nature of ethernet traffic (Extended version). *IEEE/ACM Transactions on Networking*, 1994, 2(3): 1–15.

[2] Paxson V and Floyd S. Wide-area traffic: the failure of poisson modeling. *IEEE/ACM Transactions on Networking*, 1995, 3(2): 226–244.

[3] Uhlig S, Quoitin B, and Lepropre J, *et al*.. Providing public intradomain traffic matrices to the research community. *ACM SIGCOMM Computer Communication Review*, 2006, 36(3): 156–167.

[4] Lakhina A, Crovella M, and Diot C. Diagnosing network-wide traffic anomalies. SIGCOMM, Portland, Oregon, USA, 2004: 224–235.

[5] Rubinstein B I P, Nelson B, and Huang L. Stealthy Poisoning Attacks on PCA-based Anomaly Detectors. SIGMETRICS, 2009: 168–179.

[6] Rubinstein B I P, Nelson B, and Huang L, *et al*.. Compromising PCA-based anomaly detectors for network-wide traffic. Technical Report UCB/EECS-2008-73, 2009.

[7] Vardi V. Network tomography: estimating source-destination traffic intensities from link data. *Journal of the American Statistical Association*, 1996, 91(6): 365–377.

[8] Lakhina A, papagiannaki K, and Crovella M, *et al*.. Structural analysis of network traffic flows. SIGMETRICS, New York, NY, USA, 2004: 345–356.

[9] Zhang Y, Roughan M, and Willinger W, *et al*.. Spatio-

temporal compressive sensing and Internet traffic matrices. SIGCOMM, Barcelona, Spain, 2009: 110–121.

[10] Xu K, Zhang Z L, and Bhattacharyya S. Internet traffic behavior profiling for network security monitoring. *IEEE/ACM Transactions on Networking*, 2008, 16(4): 1241–1252.

[11] Tenenbaum J B, Silva V D, and Langford J C. A global geometric framework for nonlinear dimensionality reduction. *Science*, 2000, 290(12): 2319–2323.

[12] Roweis S T and Saul L K. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 2000, 290(12): 2323–2325.

[13] Gerber S, Tasdizen T, and Joshi S, *et al*.. On the manifold structure of the space of brain images. MICCAI, USA, 2009: 263–268.

[14] 邵超, 黄厚宽, 赵连伟. 一种更具拓扑稳定性的 ISOMAP 算法. 软件学报, 2007, 18(3): 869–877.

[15] 文贵华, 陆庭辉, 江丽君. 基于相对流形的局部线性嵌入. 软件学报, 2009, 20(6): 2376–2386.

钱叶魁： 男，1980 年生，博士生，讲师，研究领域为网络测量、网络安全.

陈 鸣： 男，1956 年生，博士，教授，博士生导师，研究方向为网络测量、网络体系结构、网络管理等.