

将每个时间步的图降至点：一个动态网络探索的可视分析方法*

徐进(11621002)

May 26, 2016

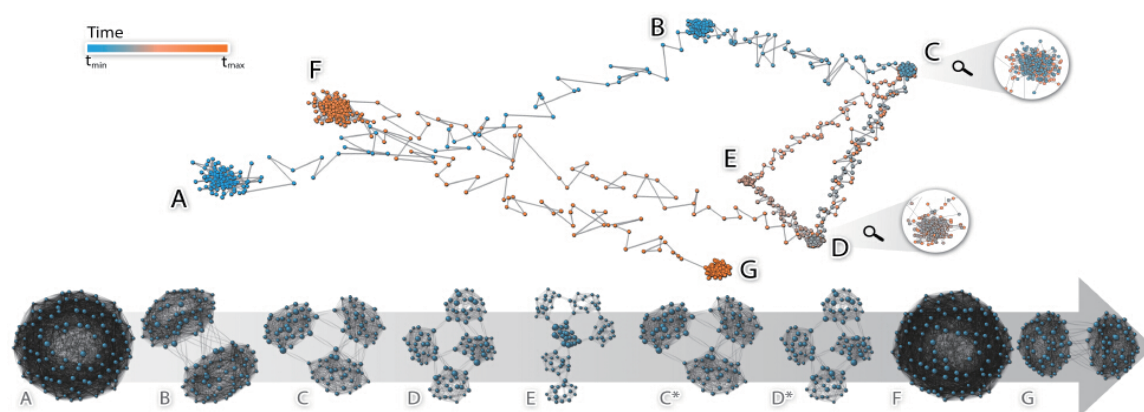


Figure 1: 通过将动态网络每个时间步的图 (snapshots) 映射成点，可以有效探索动态网络的演化。每个时间步的图的映射 (上图) 揭示这个网络包含7个稳定状态 (A-G)，其中两个重现状态 (C, D)，并且展示了这些状态之间的转化。这些时间步的图的代表被展示出来 (下图)。

1 引言

网络作为描述实体之间关系的数据结构，无处不在。这些网络通常是大规模的，并且随着时间动态变化。而对于动态网络的可视化方法，通常分为两类：Animation (time-to-time) 和 Small-multiple (time-to-space)。对于前者，用户每一时刻只能看到一帧，前后内容需要用户去记忆、理解，往往需要来回观看很多遍才能粗略理解动态网络的变化过程，给用户带来了很大的认知负担；对于后者，将不同时刻的网络进行并排罗列，由于屏幕空间有限，很多时候并不能将所有时刻的网络同时进行展示，且难以进行网络动态变化模式的发现。本文提出一种新颖的方法，来分析动态图的变化模式。他们将每个时间步的网络转换成高维向量，将这些高维向量投影到二维平面，用点表示，每个动态网络表示成一张静态的node-link图，每个node代表某个时刻的网络，link连接了相邻时刻的node，可以有效地帮助用户发现稳定状态(stable state)、重现状态(recurring state)、异常拓扑(outlier topologies)及分析网络状态间的转移过程。

2 方法概述

首先，介绍全局概念，并提供一个方法概览。接下来定义动态网络。然后是本文方法的四个步骤的描述：1) 离散化，2) 矢量化和归一化，3) 降维，以及4) 可视和交互。Fig. 2提供了方法的概览。

2.1 可视分析方法

可视分析方法包括四个步骤，以离散化为开始。在这一步中，准备动态网络数据作为分析使

*选读论文：van den Elzen, Stef, et al. "Reducing snapshots to points: A visual analytics approach to dynamic network exploration." Visualization and Computer Graphics, IEEE Transactions on 22.1 (2016): 1-10.

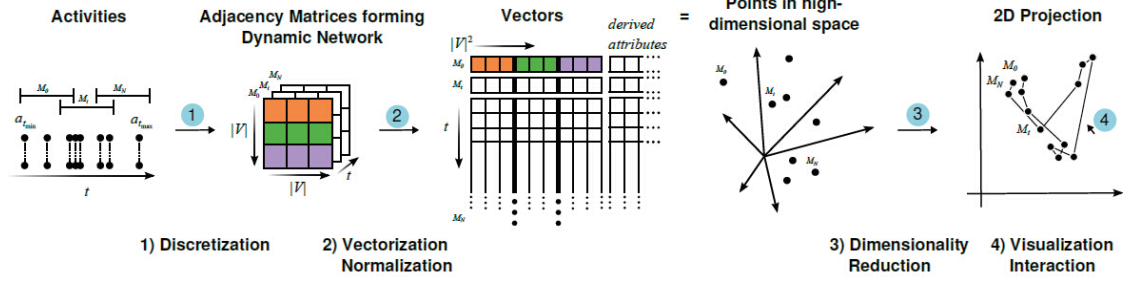


Figure 2: 探索动态网络的可视化分析方法包括四个步骤：1) 离散化，2) 矢量化和归一化，3) 降维，以及4) 可视和交互。

用。基本的思想是，基于活动记录数据，比如交易或者交流数据，创造动态网络的一系列的快照（每个时间步的图）。每个快照包含一个动态网络实例 G_i 。

接下来是比较简单的但同时对于我们的这个方法来说又是至关重要的。并且据我们所知，在之前的方法中并没有被考虑过：将网络快照矢量化并将其视为高维空间中的点。这种做法有效地将一个时间步的边表示成一个行向量；高维空间中的每个点表示不同时间步的网络。空间中快照（每个时间步的网络）的位置揭示网络的演化：如果两个时间步的网络比较相似，那么他们对应的快照的位置就是相近的并形成聚类。相似地，如果某些时间步的网络不同于那些正常的快照的位置，那么这些快照被认为是异常的。点的聚类揭示稳定或者重现的网络状态。不同聚类之间的点揭示网络如何从一个状态转化为另一个状态。

为了能够分析和探索动态网络的这一系列快照，在第三步中，我们应用降维的技术，并且将这些点映射到二维空间。最后，我们介绍两个并列的关联视图：一个快照的映射图和一个辅助的视图，展示被选中的某个快照的网络结构。

2.2 动态网络模型

我们建模一个动态网络 Γ 作为一系列 N 个快照：

$$\Gamma = (G_1, G_2, \dots, G_N), \quad (2.1)$$

其中，每个快照是一个有向图 $G_i = (V, E_i, t_i)$ ，结点（顶点）集合 V 和边（连接，交易，事件）集 $E_i \subseteq V \times V$ 带有起始结点和终止结点 (v_m, v_n) ， t_i 代表第 i 个时间步。动态网络中边的集合是所有时间步的快照的边集的并集：

$$E = \bigcup_{i=1}^N E_i. \quad (2.2)$$

更进一步，我们考虑到权重图。 w 是一个定义在边集合上的函数 $w : i \times E \rightarrow \mathbb{R}$ 分配一个实数值的权重 $w(i, e)$ 给一条边 $e \in E_i$ 。

2.3 动态网络离散化

在实际中，包含时间标签的活动记录数据集（比如，电子邮件交流数据），相对于那些提前定义好的网络快照的数据集来说，是更加普遍的。如果这些活动卷入两个实体，比如从一个地址向另一个地址发送一条信息，把这样的数据集看成是一个动态网络数据集是非常自然的。一个活动记录建模为：

$$A = (A_1, A_2, \dots, A_M), \quad (2.3)$$

其中， $A_j = (a_j, s_j) \in E \times \mathbb{R}$ 是一个活动记录，带有边 $a_j(v_m, v_n)$ 和时间戳 s_j 。为了分析，通过创建快照，这个活动记录被转化为动态网络。我们假设 $t_{j+1} - t_j = \Delta t$ 是相等的，可以表示天，小时等等。每个快照 G_i 跟宽度为 ω 的取样时间窗口 $[t_i - \omega/2, t_i + \omega/2)$ 有关。现在每个快照的边集 E_i 是：

$$E_i = a_j | s_j \in [t_i - \omega/2, t_i + \omega/2). \quad (2.4)$$

在一个快照中一条边的权重 e 简单地被定义为在边集中出现的次数。

为了创建快照，我们对于 Δt 有多种选择。在最极端的情况下，每个快照包含一个单独的活动；另一个极端的情况是一个快照包含所有的活动。显然，两个人的任意一个都是不可选的。实际中，对于 Δt 的最好选择是基于领域知识的或者是数据中逻辑的区分时间段，比如一天，一小时或者半小时。随着时间事件的可视化可以帮助决定窗口长度。

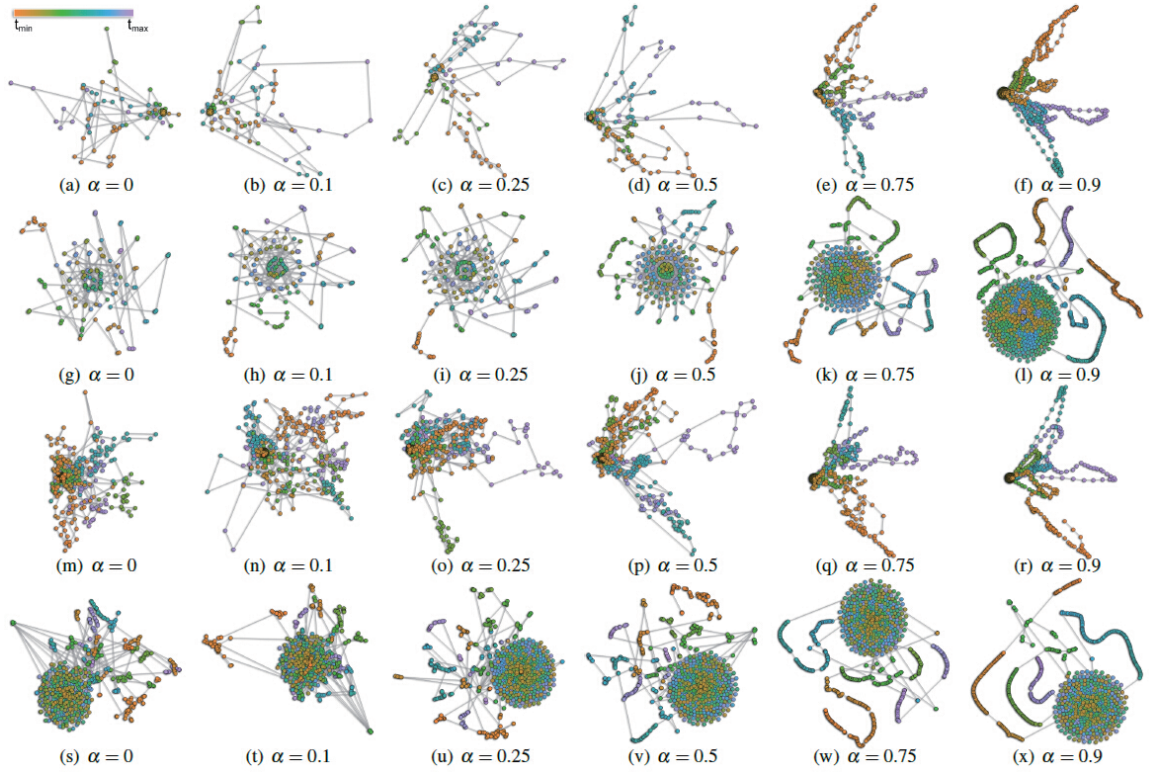


Figure 3: 在离散化过程中不同重叠值 α 对结果的影响。图片展示了线性降维的PCA的二维映射（第一行和第三行）和非线性降维的t-SNE（第二行和第四行）。窗口宽度 ω 在第一行和第二行保持常量， Δt （所以影响取样的数量）改变。a-f和g-l的快照的数量是，151,167,201,301,602,1500。在第三行和第四行，快照的数量保持常量1000，窗口的宽度 ω 变化。一般来说，重叠越多，在映射图中更多的界都被展示出来。连续的快照被用边连接起来，用来揭示随时间变化的模式。点的颜色映射为时间。

这里我们选择保持 Δt 是一个常量。如果它的取值是变化的，也是有道理的。然而，这会使得后期的解释更加麻烦。为了避免丢失任何细节，窗口宽度允许相邻时间段有重叠 α ，比如我们使用 $\omega > \Delta t$ 。这个窗口的表现就类似滑动平均值，通过增加新的边以及丢掉边，有效平滑数据。这个重叠 $\alpha \in [0, 1]$ 定义为：

$$\alpha = (\omega - \Delta t) / \omega. \quad (2.5)$$

替代一致加权不同时间段的情况，在这里使用的这种简单的做法能够满足需求，并且对于非专家来说更容易解释。Fig. 3展示了不同重叠值对结果的影响。Fig. 3, 4, 5, 6使用的是相同的数据集，在部分3会有详细描述。这篇论文建议重叠值取0.5或者更大，这是取决于数据集本身的。

每个快照包含一些边。一个快照 G_i 可以拜师为一个 $|V| \times |V|$ 的邻接矩阵 M_i

$$M_{i,jk} = w(i, (v_j, v_k)). \quad (2.6)$$

作为结果，我们现在有一系列的 N 个矩阵对应的快照，这些快照组成了一个动态的网络。概括来说，我们介绍离散化是为了创建基于活动数据集的快照，因此，我们可以更好地控制快照的数量。介绍重叠是为了平滑不同时间步的快照，可以提高最终映射的可解释性。重叠也可以减少暂时模式的缺失。动态网络的快照结果用于接下来的步骤中。

2.4 矢量化和归一化

现在我们有动态网络的快照，我们的目标是分些这些快照以及从中获取到演化信息。为了达到这个目的，我们简单地使用降维技术使每个快照降至点。

所有 N 个网络快照被表示为 $|V| \times |V|$ 的邻接矩阵，我们将这些邻接矩阵重新排列为 $1 \times |V|^2$ 的行向量，也就可以看成 $|V|^2$ 维空间中点。这些行向量组成 $N \times |V|^2$ 的矩阵（见Fig. 2，步骤2）。这个矩阵的每一列表示网络中的边，行表示不同的快照。一列就是一个特征向量，代表一条边随着时间的演化。其他属性矩阵也可以使用，比如每个快照的度等。这个方法的下一步



Figure 4: 线性降维方法PCA被应用在Figure 5: 简化状态识别。a) 很多快照包含相动态网络快照数据集里。a) 没有归一化的网络状态被布局在了一起。b) 重新定位化, b) binarization, c) min-max, 以及d) z-score之后, 减少重叠, 能够更好识别出网络中的状态。score归一化。

就是降维并且将高维空间中点映射到二维空间。这里采用线性或者非线性降维的方法, 一个在统计学和数据挖掘领域已经被充分研究的主题。在应用这些方法前, 这个矩阵应该做归一化处理以获得更好的结果。也就是说, 为归一化高维空间中的每个维度。我们采用三种归一化方法: binarization, min-max以及z-score, 见Fig. 4。对于线性降维方法来说, 比如PCA, Z-normalization是更好的选择, 因为我们对方差的最大值感兴趣。Min-max归一化导致更小的标准差, 所以抑制了异常的影响。

对于我们考虑的情况, 没有非常明显的特别好的方法, 所以这些选择似乎是依赖于数据本身的以及想要寻找的模式。归一化可能强调数据集中特殊的某种模式, 比如, z-score归一化抑制了异常, 然而min-max保留了异常。这部分工作有待于进一步研究。

2.5 降维

我们的快照是高维空间中点, 但是多维是很难理解以及可视化的。所以我们采用降维技术, 可以使高维空间中的点降低到低维空间 [1]。我们的目标是减少特征空间到二维, 并且映射这些快照以利于可视化和交互。

有很多线性降维和非线性降维的方法可以使用, 比如PCA [2, 3], MDS [4], t-SNE [5]。通过采用提高的变式, 比如Randomized PCA [6, 7] 和Barnes-Hut-SNE [8], PCA和t-SNE的计算时间可以大大减少。这可以使这些方法应用在大规模的数据集上, 也可以达到实时交互。PCA是线性降维技术, 比如结果维度是原始维度的线性组合, 导致数据的变式更好地被描述。通过使用kerneltrick [9]实现非线性可以克服它的限制 [10]。

最近的比较揭示了非线性降维在人为选定的任务上性能差异, 但是PCA仍然在现实世界中数据集有着很好的表现。所以我们选择PCA作为初始降维方法, 但是在我们的方法中也提供MDS和t-SNE方法。MDS和t-SNE的优势除了两个快照之间标准的欧几里得距离外, 更多复杂的网络距离也可以被应用。比如, 我们可以在度矩阵中提前计算距离。Fig. 6提供了更加结构化的比较。

降维之后的维度通常是很难解释的, 尤其是非线性降维之后的结果维度。然而, 在这里这并不是一个议题。对于分析和探索来说, 快照, 现在已经被降为点, 可以在二维空间中可视化出来了。

2.6 可视化和交互

为了实现动态网络的探索, 我们的系统采用两个关联的并排的视图。Fig. 7展示了图形化用户接口, 包括映射视图和网络视图。

映射视图

映射视图可视化动态网络中快照, 并将这些快照编码成点。这些点如果挨在一起就说明这些快照有相似的网络状态。相邻的快照被连在一起, 用于强调不同聚类之间的模式转化。这也使得能够识别在时间上挨得很近但是在映射图里距离特别远的快照。点的颜色映射初始编码为时间。通过这些点的颜色, 聚类中的更多结构可以被识别出来。如果一个聚类几乎有一致的颜色, 说明这个状态在网络中是稳定的, 比如, 一个网络停留在一个状态很长时间。如果一个聚类里包含多种颜色, 它揭示为这是一个重现状态。除了将颜色映射编码为时间, 也可以映射为其他信息以揭示状态。

探索过程中也支持缩放的功能。缩放可以及距离观察聚类信息。为了简化类的识别, 我们采用去混乱(de-cluter)方法减少重叠 [13]。这个结果更加强调网络中的状态 (Fig. 5)。

我们使用二维降维技术避免混乱, 保持探索简单。通过降维算法得到的轴是可以选择的。比如说, 如果使用PCA, 选择除了第一维度和第二维度之外的维度, 得到的结果也是很有成效

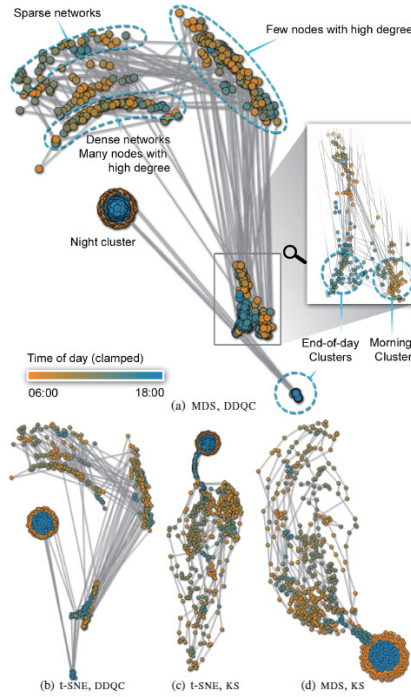


Figure 6: a,d) 采用非线性降维MDS。b,c) 采用非线性降维t-SNE根据度矩阵计算距离。c,d) Kolmogorov-Smirnov以及a,b) Degree Distribution Quantification。非线性降维也有相似的图片。这个映射揭示了一个大的夜间聚类，早晨聚类和连接夜间聚类的一天结尾的聚类。三个以上的聚类被识别出来；一个聚类有很少的节点但是有很高的度，一个聚类有很松散的网络状态，一个聚类有很紧密的网络状态。

的。此外，时间也可以被选择，作为一个维度来映射。点也可以被选择和高亮，用来展示在网络视图里相关联的快照。

网络视图

网络视图可视化的是映射视图被选中的快照的网络连接结构，用节点-连接图方式展示。每个点代表网络中的一个节点，每条线代表网络中一条边。使用用户选择的力导向图布局计算网络环境 [11]。

对于每个快照来说网络视图中的节点可以设置为可见的和稳定的，或者是对于新选择的快照来说重新排列点。重新计算布局的优点是可以生成一个清晰的网络结构的可视化，然而，这样有可能会中断人的感知图。所以，在我们的系统中，当重新计算力导向布局时，支持结点的稳定性。如果结点被重新计算位置，跟踪异常的改变变得更加容易。而且，一个稳定的布局是基于全局快照的合集的。

网络中节点的颜色可以编码为多变量属性值。类似于映射视图，缩放技术也是支持的。更进一步，高亮快照的开始和结束时间再网络视图中是被绘制的，可以提供上下文信息。简单的连接和笔刷技术也被集成进来，可以支持可视化查询。

此外，除了映射视图和网络视图，我们还集成了两个辅助的视图：时间线视图和度分布视图。时间线视图提供上下文信息，可以浏览动态网络。度分布视图展示网络中的度分布，可以在更加结构化层面比较快照。

3 实验结果

我们将这个方法应用在人造的 and 真实世界中的数据上。在这个部分将会展示结果并且讨论不同参数对结果的影响。这个可视化的方法帮助理解动态网络，揭示稳定状态，重现状态，异常状态，以及网络中的演化过程。

3.1 人造网络数据

我们创建了不同的人造网络数据来测试和评估我们提出的可视化方法。我们展示在两个数据集上的结果，第三个例子展示在Fig. 1。在我们的网络生成模型里，我们定义动态网络中结点

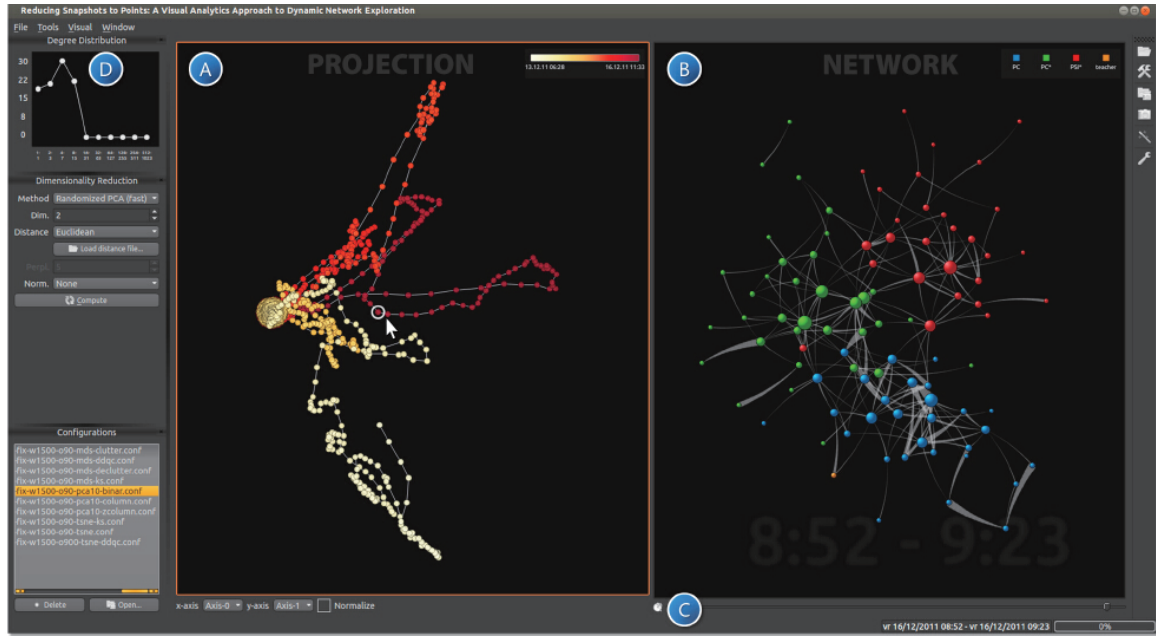


Figure 7: 图形用户接口包含两个部分：a) 映射视图，将动态网络中的快照映射为点；b) 关联的网络视图，用于展示在映射视图被选中的快照的网络结构，用结点-连接图的方式可视化，c) 关联的时间控制面板和d) 辅助的度分布视图，用于展示映射视图选中快照的度分布。

的数量和稳定状态的数量。一个稳定状态是使用Newman, Watts, 和Strogatz的小世界模型生成的 [15, 14]。接下来在不同的时间步定义不同的参数。在创建完这些稳定状态之后，网络转化的状态通过提前定义时间步的数量使用插值计算出来。

首先，我们创建一个动态网络包含100个结点，9900条边，以及650个时间步。动态网络包含4个稳定状态和3个转化状态；在一开始阶段，网络包含一个大的小世界社区；接下来这个社区分裂，两个分开的社区合并；然后其中一个更小的社区分裂成两个更小的社区；最后其中一个社区再一次加入到最大的社区里。所有状态都保持125个时间步的稳定，转化需要花费50个时间步。

如果PCA应用在网络快照时，我们能够清晰地看到快照的四个聚类以及他们中间的转化（Fig. 8）。通过高亮聚类的截图，我们在网络视图中发现网络结构。如果我们在x轴选择时间维度可以看到，最后三个状态跟第一个状态在y轴上的距离更加相近（见Fig. 9(a)）。如果我们应用z-normalization，聚类的结构更加清晰。然而聚类之间的转化就变得不清晰了（见Fig. 9(b)）。接下来我们换成t-SNE, Fig. 9(c)，聚类的稳定的状态变得更加清晰了，然而，我们再也无法看到稳定状态之间的稳定状态了。这有可能后期调到合适的参数得到解决。

第二个动态网络同样包含100个结点，9900条边，包含四个稳定状态，五个转化状态以及一

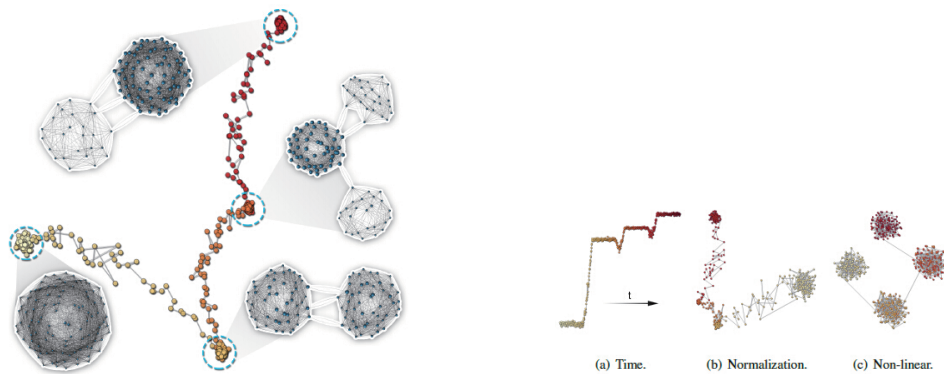


Figure 8: 使用线性降维PCA揭示网络中有四个Figure 9: 动态网络中快照的可选的映射方状态以及他们中间的转化状态。插入的灰色图式：a) 时间和第一主成分维度，b) PCA和Z-normalization, 和c) 使用t-SNE非线性降维。展示的是具有代表性的快照的网络结构。

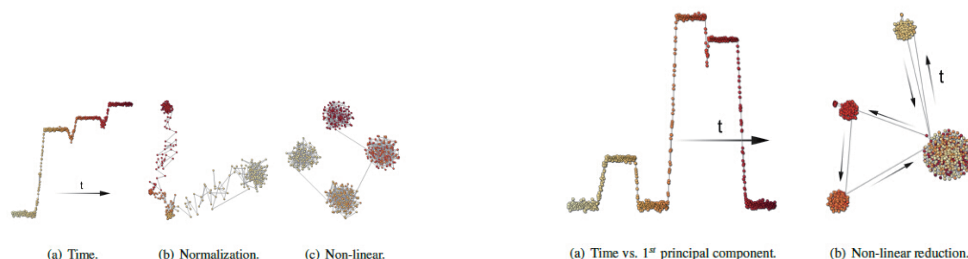


Figure 10: 线性降维PCA揭示四个状态，三个稳态和一个重现状态。最大的蓝色圆展示重个数据集。Figure 11: 可选的映射提供更多的视角观察同一状态的一个zoomed-in版本，灰色图案展示具有代表性的快照。

个重现状态。这个网络一开始有一个大的社区（重现状态），然后分裂成两个小一点的社区；接下来一个大的社区又重新组合起来；然后这个网络分裂成三个更小的社区，最后又合并成一个大的社区。这个系列包含900个时间步。

首先我们再一次应用PCA映射快照，我们可以识别出四个不同的稳定的状态。我们看见中间转化的状态并且识别出其中一个状态是重现状态，因为这个状态包含不同颜色的点，见Fig. 10。如果我们将时间作为x轴映射，我们可以更加清晰地发现稳定状态以及他们的稳定时长，见Fig. 11(a)。更进一步，我们现在能够看见重现状态何时发生。通过使用t-SNE，我们能够看见稳定的状态，见Fig. 11(b)。稳定状态之间的转化状态变得不可见，但是根据连接的相邻快照，我们仍然能够识别稳定状态的序列。而且，由于点有更好的聚类，重现状态变得更加明显。

3.2 高中交流模式

两个不同的数据集都收集于SocioPatterns倡议 [12]。通过可穿戴式的传感器，两个数据集都包含人们之间面对面交流的具有时间戳的事件，用来探索疾病在人群中是如何传播的。第一个数据集包含来自三个不同班级在四个上学日里（见Fig. 3,4,6,7），高中生面对面交流的数据。第二个数据集包含来自五个班级七个上学日（一个周一到周二，以及接下来一周的五天）里的面对面交流的数据。我们关于这两个数据集得到的结果是类似的，所以下面仅讨论第二个更加丰富的数据集。

这个动态网络包含180个结点（学生），45047个交流，以及10104条独特的边。选定窗口宽度 ω 为60分钟，重叠率 α 为0.9，创建快照。这导致相邻的两个快照间在6分钟的时间间隔内（ Δt ）有加新边或者丢掉边。生成的快照的总是数量是2015。首先我们使用PCA降维（Fig. 12右图）。跟之前比较相似，重现状态的夜间聚类被展示出来，每一天也被循环展现出来。在这个夜间聚类中，远离中心的快照有更加紧密的网络结构，而距离聚类中心越近，这些快照对应的网络越稀疏。如果点的颜色不是映射为天，而是映射为一个小时，我们能够看到紧密的网络通常发生在一天的开始，下午的快照对应的网络通常是稀疏的。

我们把降维方法改为非线性降维t-SNE。夜间聚类再一次出现，白天的聚类展示为分离的轨迹。这里我们可以得到这样的结论，每天都是不一样的，因为没有轨迹产生重叠，见Fig. 13左图。我们也能通过每个轨迹的间隙，看到每天的网络结构。这显示了一天的节奏，大概每天的中断是上课和午饭。如果我们将颜色映射为每天的一个小时，我们能够看到每个间隙发生在相似的时间点上，见Fig. 13右图。仔细检查之后，我们发现在每天的8点到9点之间有一个快照的聚类。在这个聚类里，这些快照对应的网络结构非常紧密，而且在这之前和之后网络结构都比较稀疏。既然快照的颜色映射为一天的每个小时，我们能够看到其中有一天的轨迹是不同于其他天的。周三的轨迹更短一些而且下午的快照是断开的。这可以解释为当天下午进行了一场考试，延续了整个下午并且没有中断。而且一个在周四下午的快照的聚类被识别出来。这里，这个网络的状态代表课上的交流。

4 小结与讨论

这篇文章提出了一个新颖的方法用于动态网络的可视探索。这个方法包括四个步骤：离散化，矢量化和归一化，降维，可视化和交互。最重要的思想就是把一个快照降至点用于支持动态网络的探索和可视化。

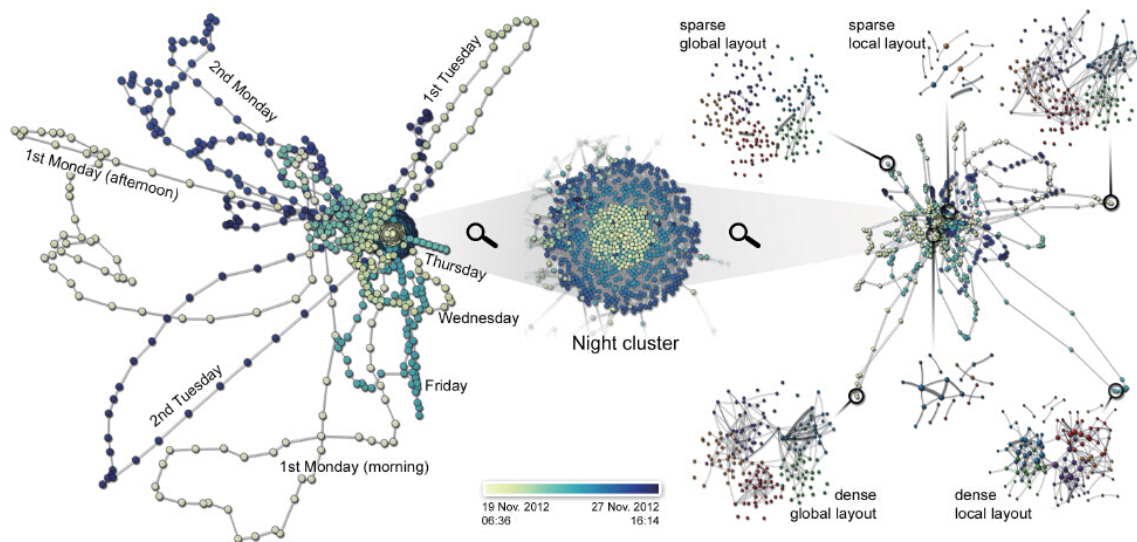


Figure 12: 线性降维PCA没有z-normalization（左）以及有z-normalization（右）。中间的聚类展示了一个zoomed-in版本，揭示这是一个重现状态，颜色编码时间，所以一天的开始和结束生成了一个循环状态。网络注释展示的是黑色圆圈圈注的快照的网络结构。

这是文章是2015 InfoVis的Best paper。文章的方法不难，文章里也指出，用到的方法都是现在研究得很成熟的方法，但就是把这些方法结合起来，做了一件开创性的事。整体思想简单，一气呵成。个人认为这篇文章的优点在于：

- 选题。动态网络分析是目前的研究热点，本文思想简单，但得到的结果效果很不错，能偶识别出稳定状态，重现状态，异常状态以及状态之间的转化。
- 工作完整性。本文的工作一共包含四步，每一步用到的方法都是已经研究成型的方法。本文对各个不同想选择都做了分析和细致的讨论。扩展性很强。
- 实验完整。在实验部分，本文用了两种数据：虚拟数据和真实数据。充分讨论了本文方法的有效性。虚拟数据的结果非常漂亮，很吸引人，理论上证明了方法的可用性；真实数据进一步证明了方法的实用性。

参考文献

- [1] L. Van der Maaten, E. Postma, and H. Van den Herik. Dimensionality Reduction: A Comparative Review. Technical Report TiCC TR 2009-005, 2009.
- [2] H. Hotelling. Analysis of a Complex of Statistical Variables into Principal Components. J. Educ. Psych., 24, 1933.
- [3] K. Pearson. LIII. On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine Series 6, 2(11):559 - 572, 1901.
- [4] J. Kruskal. Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. Psychometrika, 29(1):1 - 27, 1964.
- [5] L. van der Maaten and G. E. Hinton. Visualizing High-Dimensional Data Using t-SNE. J. Machine Learning Research, 9:2579 - 2605, 2008.
- [6] N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. ArXiv e-prints, Sep. 2009.
- [7] V. Rokhlin, A. Szlam, and M. Tygert. A Randomized Algorithm for Principal Component Analysis. SIAM J. Matrix Anal. Appl., 31(3):1100 - 1124, Aug. 2009.
- [8] L. van der Maaten. Barnes-Hut-SNE. In Int. Conf. Learning Representations, 2013.
- [9] J. Shawe-Taylor and N. Cristianini. Kernel Methods for Pattern Analysis. Cambridge University Press, New York, NY, USA, 2004.
- [10] B. Schölkopf, A. J. Smola, and K.-R. Müller. Advances in Kernel Methods. chapter Kernel Principal Component Analysis, pages 327 - 352. MIT Press, Cambridge, MA, USA, 1999.

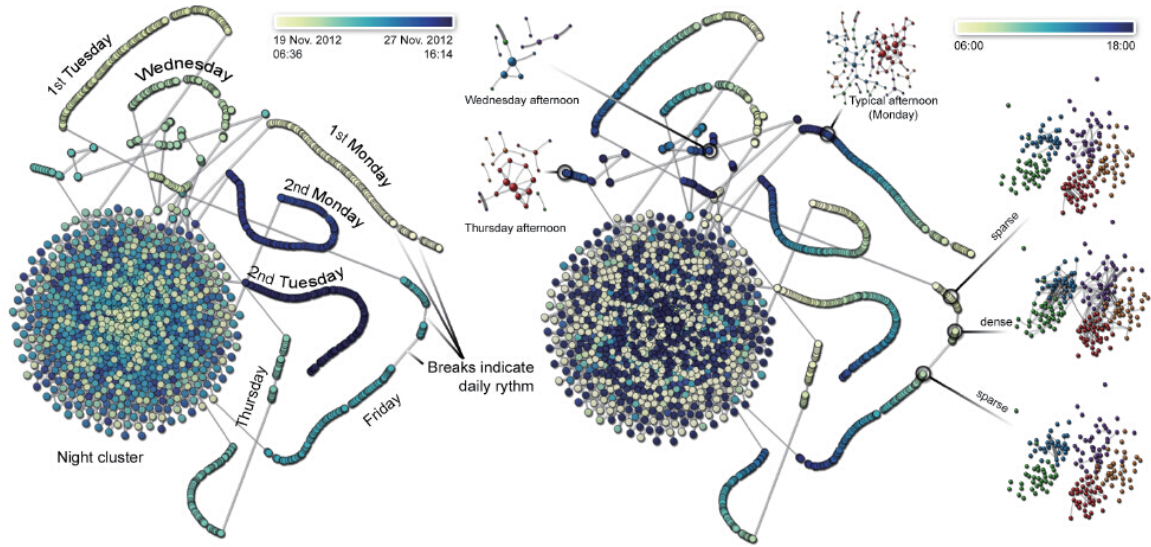


Figure 13: 非线性降维，左图颜色映射为全局时间，右图颜色映射为一天中的每个小时。标注的网络展示的是黑圈内快照的网络结构。

- [11] E. R. Gansner and S. C. North. An Open Graph Visualization System and its Applications to Software Engineering. *Software - Practice and Experience*, 30(11):1203 – 1233, 2000.
- [12] SocioPatterns. website. <http://www.sociopatterns.org/datasets/>, 2015. [Online; accessed 2015-03-18].
- [13] H. Vogel. A Better Way to Construct the Sunflower Head. *Mathematical Biosciences*, 44(3 – 4):179 – 189, 1979.
- [14] M. Newman and D. Watts. Renormalization Group Analysis of the Small-World Network Model. *Physics Letters A*, 263(46):341 – 346, 1999.
- [15] D. Watts and S. Strogatz. Collective Dynamics of ' Small-World' Networks. *Nature*, 393(6684):440 – 442, 1998.