

局部优先的动态网络重叠社团及其演变模式检测

彭 焱, 溪利亚

(武昌首义学院 信息科学与工程学院, 武汉 430064)

摘 要: 社团检测简化是重要的图挖掘问题, 动态网络上的重叠社团检测及其社团演变模式是近年来的研究热点, 但将静态网络的局部优先的社团检测算法应用到动态网络的重叠社团及其演变模式检测上的研究较少。为此, 提出一种局部优先的动态网络重叠社团演化分析方法。该方法在每个网络快照上利用标签传播算法检测局部 Ego 社团, 通过不断合并局部 Ego 社团得到全局社团结构。在此过程中引入社团相似度与关联度 2 个概念, 利用演化聚类框架进行重叠社团演化分析。实验结果表明, 该方法不仅能有效地发现动态网络中重叠社团结构, 而且还可以分析出社团随时间的演化模式。

关键词: 动态网络; 重叠社团; 局部优先; 演变模式; 时序平滑框架; Ego 社团

中文引用格式: 彭 焱, 溪利亚. 局部优先的动态网络重叠社团及其演变模式检测[J]. 计算机工程, 2016, 42(12): 188-195, 203.

英文引用格式: Peng Yan, Xi Liya. Local-first Detection of Overlapping Community and Its Evolution Pattern in Dynamic Networks[J]. Computer Engineering, 2016, 42(12): 188-195, 203.

Local-first Detection of Overlapping Community and Its Evolution Pattern in Dynamic Networks

PENG Yan, XI Liya

(College of Information Science and Engineering, Wuchang Shouyi University, Wuhan 430064, China)

[Abstract] Community detection is an important image mining problem, and the overlapping community detection and the community evolution pattern in dynamic networks is a hot topic. However, There is no work on the integration of the local-first community detection. This paper proposes a local-first analytical approach to analyze the evolution pattern of the overlapping communities on dynamic networks. This approach detects the local Ego community via label propagation on each network snapshot. Then, the global communities are generated via merging the local Ego communities successively. In this process, the two concepts of the community similarity and community association are introduced, and it analyzes the overlapping community evolution pattern via an evolutionary clustering framework. Experimental results show that this approach not only detects the overlapping communities in dynamic networks effectively, but also analyzes the evolutionary pattern.

[Key words] dynamic networks; overlapping community; local-first; evolution pattern; timing smoothing framework; Ego community

DOI: 10.3969/j.issn.1000-3428.2016.12.033

0 概述

随着互联网的快速发展, 人们周围充满了各种各样的复杂网络, 如 Internet 网络、社交网络以及论文合作网络等。这些复杂网络随时间不断变化, 例如网络流量数据网络^[1]、电话流量数据网络^[2]、论文合作网络^[3]、动态社会网络^[4]等。在动态网络上检测网络社团结构并分析其演化模式, 这对于挖掘网

络中所隐藏的知识、展现动态网络社团演化的客观规律具有重要的意义。

动态网络上社团演化过程包括出现与死亡、合并与分解、扩大与缩小等。演化聚类是一种能够自动发现动态数据集中的聚类及其演化过程的方法。文献[5]提出一种时序平滑的演化聚类框架, 该框架假设数据集中的类簇在相邻时刻内变化非常小。为满足时序平滑性, 该框架引入 2 种质量评价指

作者简介: 彭 焱 (1973—), 男, 讲师、硕士, 主研方向为动态网络、大数据分析; 溪利亚, 副教授。

收稿日期: 2015-11-30 **修回日期:** 2016-01-28 **E-mail:** yu7036077@126.com

标: 快照质量和时序质量。快照质量用于评价当前数据集上聚类结果的质量, 而时序质量用于评价当前时刻聚类结果与前一时刻聚类结果的相似性。本文通过用户自定义的参数来平衡这 2 种质量, 然后优化一个质量函数, 从而得到动态数据集中聚类的优化结果。该框架可以用于检测动态网络中演化社团结构^[6-8]。然而, 已有方法仍有一些局限性: 1) 这些方法通常从全局视角对整个网络进行社团划分, 然而很多真实网络规模巨大, 很难把握网络的全局结构; 2) 难以检测动态网络中的重叠社团结构。

本文提出一种局部优先的动态网络重叠社团演变分析方法 LEOD。该方法从网络中每个节点出发, 在该节点的邻居网络上利用标签传播算法进行局部社团检测, 得到其 Ego 社团, 然后将 Ego 社团合并到全局社团中。在合并过程中, 引入社团相似度和社团关联度 2 个概念, 利用时序平滑框架, 发现动态网络上的社团结构并进行演化分析。

1 相关工作

复杂网络中的社团检测^[9-12]是近年来的一个研究热点, 比如: 文献[9]提出一种局部优先的社团检测算法 DEMON, 该算法以网络中的每个节点为对象, 先找到该节点的局部社团, 然后合并这些局部社团得到全局社团。这种局部优先的社团检测方法不需要事先知道网络的全局结构, 具有较高的时间效率和准确率。目前, 国内也有很多团队研究复杂网络上的社团检测问题^[13-17]。已有的社团检测算法往往是从全局出发对整个网络进行社团划分^[18-19]。这种方法具有以下局限性: 1) 聚类之前必须知道整个网络的结构; 2) 对于某些社团结构不明显的网络, 社团检测能力有限。

由于很多真实网络均是动态变化的, 许多研究者开始关注动态网络上的社团检测问题^[20-22]。文献[1]基于最小描述长度 (MDL) 原理, 提出一种无需用户自定义参数的方法来发现动态网络中的社团。文献[23]提出一种随机块模型来发现动态网络中的社团以及它们的演化, 该模型通过建模网络中独立个体所在社团的发展趋势来研究社团的演化。与其他方法不同, 该方法采用贝叶斯方法处理参数评估, 从而自动找到所有未知参数的后验分布。文献[24]提出一种框架用来表示建模与挖掘时序信息网络, 该框架采用一种基于图模型的方法来表示图以及它们随时间的改变。

演化聚类作为研究动态网络的重要方法, 可以被用在动态网络社团检测中。文献[6]扩展了相似性概念, 提出一种演化谱聚类。该方法引入

了 PCQ 和 PCM2 个框架, 其中 PCQ 是指将当前检测应用于历史数据, 根据历史数据的聚类质量决定时序开销的值; 而 PCM 是直接当前聚类结果与历史聚类结果进行比较, 结果的差异性决定时序开销的值。针对多特征网络, 即一个网络中节点存在多个类型, 而每个类型节点之间可能会有联系。文献[25]提出一种基于谱聚类的演化聚类框架来检测这类网络的社团结构, 这个框架可以处理节点及节点之间边的演化等情况。文献[7]提出 FacetNet 框架, 该框架引入了社团网络和演化网络 2 个概念分别代表社团结构以及它们的演化, 利用非负矩阵分解, 将社团结构的发现与社团的演化分析结合到一个统一的过程中, 但是该框架假设相邻时刻网络中的社团数目相等。文献[8]提出一种基于微粒与密度的演化聚类方法。本文算法与上述算法的主要不同之处在于: 1) 采用局部优先的方式, 从网络中的每个节点而不是整个网络出发得到网络的社团结构; 2) 能够检测到动态网络中的重叠社团。

2 问题描述

动态网络 $G = \langle G_1, G_2, \dots, G_k \rangle$ 是由 k 个时刻的网络组成的序列。对于任意时刻 t , $G_t = (V_t, E_t)$ 代表该时刻的网络快照, 其中 V_t 与 E_t 分别表示相应的节点集合与边集合。在 t 时刻网络快照上, 通过聚类算法可以得到该时刻网络的社团结构 CR_t 。表 1 给出了相关变量及其定义。本文要解决的问题包括: 1) 检测动态网络中的重叠社团结构; 2) 挖掘动态网络相邻时刻的社团演化模式。

表 1 变量及其定义

变量	定义
G	动态网络
G_t	时刻 t 时的网络快照
V_t	G_t 中的节点集合
E_t	G_t 中的边集合
CR_t	t 时刻网络 G_t 的社团集合
$C_{n,t}$	CR_t 中的其中一个社团

图 1 展示了动态网络及其社团的演化关系。从图 1 可以看出, 随着时间的改变, 动态网络的结构是在不断发生改变的, 因此动态网络上的社团结构也会有所变化。 $t-1$ 时刻的网络被划分成了 2 个社团 (其中节点 7 属于重叠节点), 而 t 时刻的网络被划分成了 3 个社团。显然, 通过观察可以知道, 颜色相同的社团之间存在着演化关系。怎样找到动态网络上的社团结构以及对它们的演化进行分析, 是本文的工作重点。

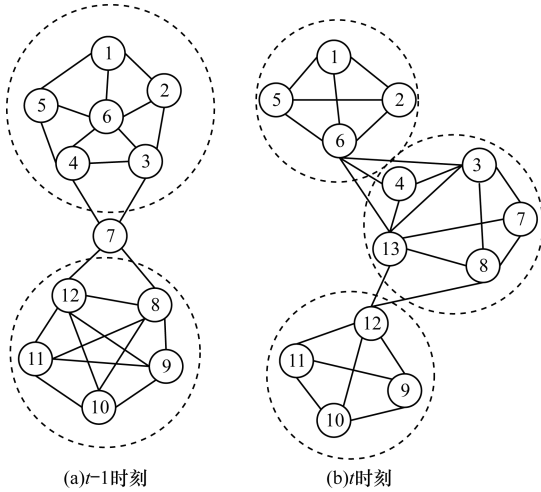


图 1 2 个相邻时刻快照网络中的社团演变示例图

3 局部 Ego 圈子检测

3.1 Ego 社团

为找到合适的粒度定义社团,同时刻画社团的演变过程,本文引入 Ego 社团的概念。

定义 1 (邻居网络) 给定 t 时刻的网络快照 $G_t = (V_t, E_t)$ 及节点 $v \in V_t$, $G_t'(V_t', E_t')$ 表示由节点 v 的邻居 $N(v)$ 构成的网络,其中:

$$V_t' = \{x_i | (v, x) \in E_t, x \in V_t\}$$

$$E_t' = \{(u, x) \in E_t | u, x \in V_t'\}$$

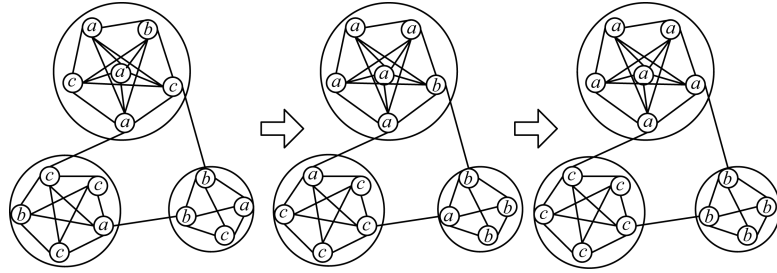


图 2 标签传播算法社团发现的过程

对于节点 v ,其局部圈子的检测算法如下:

- 1) 初始化节点 v 邻居网络中各节点 u 的标签 $l_u(0) = u$;
- 2) $t = 1$;
- 3) 将 v 的邻居网络中所有节点随机重排到集合 S 中;
- 4) 对于每一个节点 $u \in S$, 令:
$$l_u(t) = l_{\max}(l_{u,1}(t-1), \dots, l_{u,m}(t-1), l_{u,m+1}(t), \dots, l_{u,k}(t))$$

其中,函数 $l_{\max}$ 返回出现频率最高的标签;

- 5) 如果 S 中每个节点的标签均不再改变,或者迭代次数 t 已经达到某个预设的最大值,则结束算法;否则, $t = t + 1$, 返回 3) 继续执行。

4 时序平滑框架

本节首先介绍网络的全局社团检测方法,然后

定义 2 (邻居社团) 给定网络 $G_t = (V_t, E_t)$ 及节点 $v \in V_t$, v 的邻居社团 $C(v)$ 是 $N(v)$ 的一个划分,任取 $c \in C(v)$ 有 $c \subseteq N(v)$ 。

$c \in C(v)$ 是一个社团,其社团内节点间的链接要比社团外的节点间的链接更为紧密。

定义 3 (Ego 社团) 节点 v 的 Ego 社团是社团的集合:

$$Ego(v) = \{c \cup v | c \in C(v)\} \quad (1)$$

以图 1 中 t 时刻的网络为例,节点 4 的 Ego 社团只有一个社团,为 $\{3, 4, 6, 13\}$ 。而节点 6 的 Ego 社团则包含 2 个社团,分别是 $\{3, 4, 6, 13\}$ 以及 $\{1, 2, 5, 6\}$ 。

3.2 局部圈子检测方法

本文利用标签传播算法 LP^[17] 来检测局部 Ego 社团。该算法以节点的邻居网络作为输入,对其中的每一个节点赋予一个标签,其中相同标签表示一个局部社团。LP 算法对邻居网络中每个节点,统计其邻居节点的标签信息,选择其中出现频率最高的标签来更新自己的标签。随着标签的不断传播,相互连接稠密的节点组将会很快稳定在同一个标签上。当标签传播结束的时候,拥有同一个标签的节点将属于同一个局部社团。图 2 展示了标签传播算法社团发现的过程。

给出多时间戳网络的社团演化分析框架。

4.1 全局社团检测

简单地说,社团检测的任务就是找到网络中相似的节点集合,这些节点集合组成一个社团。这里所有 Ego 社团中相似的局部社团合并成一个更大的社团,因为 Ego 社团中的局部社团内部节点是相似的,所以可以将 Ego 社团中的每个局部社团看成一个整体。由此,引入全局社团的定义:

$$CR_t = \max(\bigcup_{v \in V_t} Ego(v)) \quad (2)$$

其中,给定一个集合 S , $\max(S)$ 表示 S 的极大集,即不存在包含 $\max(S)$ 的集合。由式 (2) 可知, t 时刻网络的全局社团 CR_t 是通过选取能够覆盖所有节点 Ego 社团的最大社团集合得来的。

从式 (2) 可以看出, t 时刻网络的全局社团可以通过求 Ego 社团的并集得到,而求并集的过程是通

过不断合并 Ego 社团中的局部社团来实现的。因此, 首先需要定义 Ego 社团的相似性。

定义 4 (社团相似度) 2 个社团 $C_{m,t}, C_{n,t}$ 的相似度 $\text{sim}(C_{m,t}, C_{n,t})$ 由如下公式定义:

$$\text{sim}(C_{m,t}, C_{n,t}) = \frac{|C_{m,t} \cap C_{n,t}|}{\min(|C_{m,t}|, |C_{n,t}|)} \quad (3)$$

式(3)表明, 2 个社团的相似度等于它们共享的节点数除以较小社团的节点总数。以图 1 中 t 时刻的网络为例, 社团 $\{1, 2, 5, 6\}$ 与社团 $\{3, 4, 6, 13\}$ 的相似度为 $1/4$, 即 0.25 。

4.2 时序平滑框架

由上文可知, 相似性大的 Ego 社团可以合并成一个社团。通过不断合并能够得到全局社团, 但是, 这只是针对特定的网络快照。因此, 还必须过渡到动态网络。

对于 t 时刻的 Ego 社团, 社团相似度公式可以计算出它们之间的相似度。对于演化聚类, 仅有社团相似度并不能说明这 2 个 Ego 社团可以合并到一个社团中, 因为没有考虑到它们的时序信息。利用时序平滑框架, 还需要知道这 2 个 Ego 社团在前一个时刻网络社团中的关系, 即需要知道在前一个时刻的网络中, 这 2 个 Ego 社团是否相似。

由于动态网络中 2 个节点之间的关系不断变化, 因此对于 t 时刻的 2 个社团来说, 它们在 $t-1$ 时刻的网络中可能出现多种情况: 1) 并不存在; 2) 分布在不同的社团中; 3) 属于同一个社团; 4) 只有部分节点分布在 $t-1$ 时刻的网络中。伴随着节点的加入与消失, 节点的 Ego 社团也在不断发生改变。因此, 引入社团关联度来刻画社团在 2 个相邻时刻的关系。

定义 5 (社团关联度) 2 个社团 $C_{m,t}, C_{n,t}$ 的关联度 $\text{rel}(C_{m,t}, C_{n,t})$ 定义为:

$$\text{rel}(C_{m,t}, C_{n,t}) = \frac{\max_{C_{p,t-1} \in CR_{t-1}} |(C_{p,t-1} \cap (C_{m,t} \cup C_{n,t}))|}{|C_{m,t} \cup C_{n,t}|} \quad (4)$$

其中, $C_{m,t} \cap C_{p,t-1} \neq \emptyset \wedge C_{n,t} \cap C_{p,t-1} \neq \emptyset$ 。以图 1 中 t 时刻的网络中的 2 个 Ego 社团 $\{1, 2, 5, 6\}, \{3, 4, 6, 13\}$ 为例, 可以得到它们的社团关联度为 $6/7$, 即 0.86 。

时序平滑框架最初是建立在社团检测结果 CR_t 之上的。实际上, 该框架不仅可以应用在全局社团层面上, 也可以应用在局部社团, 也就是 Ego 社团层面上。通过社团关联度, 可以得到 t 时刻的 2 个 Ego 社团在 $t-1$ 时刻的关系。所以, 综上所述, 可以改写时序平滑框架的质量函数为:

$$cq = \alpha \cdot \text{sim}(C_{m,t}, C_{n,t}) + (1 - \alpha) \cdot \text{rel}(C_{m,t}, C_{n,t}) \quad (5)$$

从式(5)中可以看出: 当 α 较大时, 更加侧重社团相似度, 即当前时刻网络的社团结构; 当 α 较小时, 则更加侧重社团关联度, 即当前时刻的社团结构与要前一时刻的社团结构相似。特别地, 当 $\alpha = 0$ 时, 等式退化为只包含社团关联度, 也就是说, 当前时刻的社团检测结果与前一时刻的社团检测结果完全类似。而当 $\alpha = 1$ 时, 等式退化为只包含社团相似度, 社团检测时则只需要考虑当前时刻的网络结构, 从而使演化聚类退化为一般的静态网络上的聚类。图 3 展示了 α 取不同值时对演化聚类结果的影响。

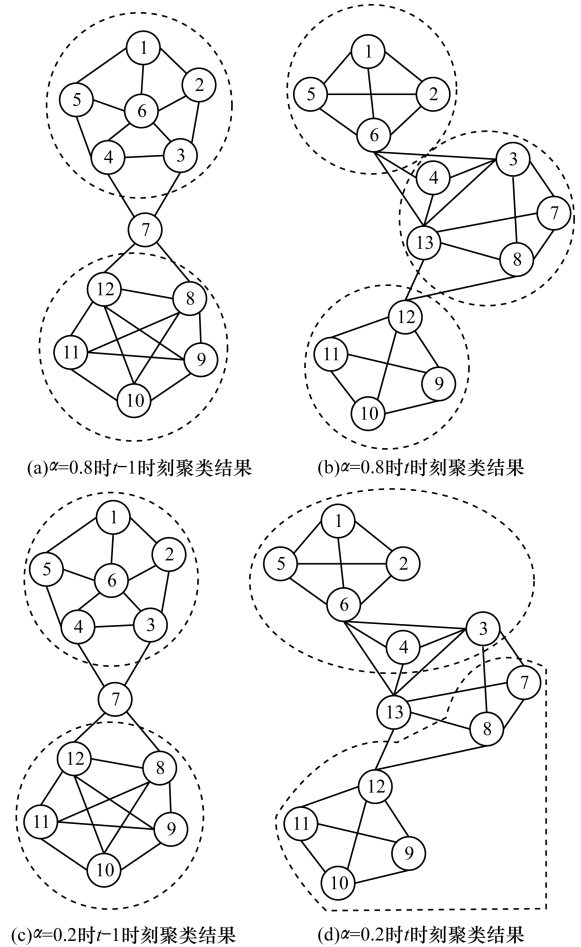


图 3 α 取不同值时对演化聚类结果的影响

从图 3(a) 可以看到, 当 α 较大时 (例如, $\alpha = 0.8$), 聚类侧重反映当前的网络结构, 所以在 t 时刻出现了一个新的社团 $\{3, 4, 7, 8, 13\}$ 。而当 α 很小时, 如图 3(b), 聚类侧重保持前一个时刻的聚类结果, 所以即使网络结构有所改变, 聚类结果却与前一时刻没有太大的差别。可以看出, α 的变化会允许社团的出现与消失。

4.3 局部优先的演化网络聚类

由式(2)可知, 全局社团 CR_t 是通过不断合并

Ego 社团得到的,在合并的过程中,利用时序平滑框架提出以下社团合并公式:

$$\alpha \cdot \text{sim}(C_{m,t}, C_{n,t}) + (1 - \alpha) \cdot \text{rel}(C_{m,t}, C_{n,t}) \geq \eta \quad (6)$$

其中, $\text{sim}()$ 表示社团相似度; $\text{rel}()$ 表示社团关联度,通过一个用户自定义参数 α 来平衡两者。同时,再引入一个用户自定义参数 η 来控制社团合并的难度。可以知道,当 $\eta = 0$ 时,即使 2 个社团没有任何共享的节点,且前一个时刻这 2 个社团没有任何关联也能够合并;而当 $\eta = 1$ 时,只有当 2 个社团完全相同时才可以合并。

这样,通过求每个节点的 Ego 社团,然后采用式(6)所示的合并策略不断合并,可以得到最终的全局社团。算法伪代码如算法 1 和算法 2 所示。

算法 1 LEOD 算法

输入 $G_t(V_t, E_t), CR_t = \emptyset, \alpha, \eta \in [0, 1]$

输出 t 时刻的社团集 CR_t

```

1: for  $v \in V_t$  do
2:    $C(v) = \text{LabelPropagation}(N(v))$ 
3:   for  $c \in C(v)$  do
4:      $c = c \cup v$ 
5:      $CR_t = \text{Merge}(CR_t, c, \alpha, \eta)$ 
6:   end for
7: end for
8: return  $CR_t$ 
```

社团 c 合并到全局社团的过程如算法 2 所示。

算法 2 Merge 算法

输入 全局社团 CR_t , 待合并的社团 $c, \alpha, \eta \in [0, \dots, 1]$

输出 合并完成的全局社团 CR_t

```

1: for  $c' \in CR_t$  do
2:   if ( $\text{sim}(c', c) = 1 \wedge |c'| \geq |c|$ )
3:     return  $CR_t$ 
4:   end if
5:   if ( $\alpha \times \text{sim}(c', c) + (1 - \alpha) \cdot \text{rel}(c', c) \geq \eta$ )
6:      $u = c' \cup c$ 
7:      $CR_t = CR_t - c'$ 
8:      $CR_t = CR_t \cup u$ 
9:     flag = true
10:  end if
11: end for
12: if (flag  $\neq$  true)
13:    $CR_t = CR_t \cup c$ 
14: end if
15: return  $CR_t$ 
```

从上面的算法可知,本文算法对网络中的每个节点只处理一次。分别扫描网络中的每个节点并获取其邻居网络,在邻居网络上利用标签传播算法^[17]检测得到邻居社团,进而形成节点的 Ego 社团。然

后,利用式(6)将 Ego 社团合并到全局社团中。最终,通过对网络中的每个节点处理一次,并合并它们的 Ego 社团到全局社团,能够得到最终的全局社团 CR_t 。

本文算法的优点在于:1)对网络中的每个节点处理一次,采用局部优先的策略,避免了聚类前必须知道整个网络结构的约束,同时能发现网络中潜在的社团信息;2)由于对于每个节点都要找到其 Ego 社团,并将 Ego 社团合并到全局社团中,因此对于某些可能属于多个社团的节点,将会被划分到多个局部社团中,经过社团合并,会产生重叠社团。

如图 3 所示,当 $\alpha = 0.8$ 时,考虑 t 时刻的聚类。若全局社团中已经加入节点 1 的 Ego 社团 $\{1, 2, 5, 6\}$,现在算法处理节点 4,可知节点 4 的 Ego 社团为 $\{3, 4, 6, 13\}$,由式(6)可得 $cq = 0.8 \times \frac{1}{4} + 0.2 \times \frac{6}{7} \approx 0.371$ 。所以,只要用户自定义的参数 $\eta < 0.371$,这 2 个社团就可以合并,从而形成一个新的社团 $\{1, 2, 3, 4, 5, 6, 13\}$;否则,它们不会合并,全局社团中这 2 个社团共同存在。当 $\alpha = 0.2$ 时, $cq = 0.2 \times \frac{1}{4} + 0.8 \times \frac{6}{7} \approx 0.736$,只有当用户自定义的参数 $\eta > 0.736$ 时,这 2 个社团才不会合并。

4.4 社团演化分析

上文讨论了局部优先的演化聚类,通过本文提出的演化聚类算法,可以得到动态网络各个时刻网络上的社团结构。下面需要对相邻时刻社团之间的演化关系进行分析。在局部优先的演化聚类算法中,通过定义社团相似度和社团关联度以及时序平滑框架的引入,使动态网络的聚类结果有了演化特性。由此,扩展社团相似度以及社团关联度的概念,对于相邻时刻社团的演化分析,引入社团演化率 $evo()$ 的定义为:

$$evo(c \in CR_{t-1}, c' \in CR_t) = \frac{|c \cap c'|_{\text{node}} + 2 \cdot |c \cap c'|_{\text{edge}}}{\min(|c|_{\text{node}} + 2 \cdot |c|_{\text{edge}}, |c'|_{\text{node}} + 2 \cdot |c'|_{\text{edge}})} \quad (7)$$

其中, $|\cdot|_{\text{node}}$ 表示节点数; $|\cdot|_{\text{edge}}$ 表示边数。由此可见,演化率就是这 2 个社团共享的节点数与共享的边数的 2 倍之和,除以这 2 个社团中节点数与边数 2 倍之和的较小者。之所以增加边的权重,是因为 2 个社团共享的边越多,越能说明社团内部结构是相似的。以图 3 中 $\alpha = 0.8$ 时的聚类结果为例, $t-1$ 时刻由节点 1~节点 6 构成的社团至 t 时刻由节点 1, 2, 5, 6 构成的社团的 $t-1$ 社

团演化率为 $\frac{4+2 \times 5}{4+2 \times 6} = 0.875$, 而 $t-1$ 时刻由节点 8 ~ 节点 12 构成的社团至 t 时刻由节点 9 ~ 节点 12 构成的社团的社团演化率为 $\frac{4+2 \times 6}{4+2 \times 6} = 1$ 。

5 实验结果与分析

本节采用人工网络和真实网络来评价 LEOD 算法的性能。将本文提出的算法与文献[8]提出的算法(称之为“KIM”)和文献[7]中提出的 FacetNet 算法进行比较。其中, KIM 是一个有效的基于密度的演化聚类算法, 而 FacetNet 算法则是基于非负矩阵分解的演化聚类算法。这 2 个算法与 LEOD 有很大的相关度, 也是演化聚类中优秀的算法。所有算法均采用 ANSI C++ 编写, 实验在 3.20 GHz, 4.0 GB 内存的计算机上完成。

5.1 数据集

5.1.1 人工数据集

利用文献[18]中所述的工具生成了节点数为 1 000, 每组均包含 15 个时刻网络的 4 组动态网络, 这些网络是带有重叠的, 且每组动态网络各自包含不同的社团演化现象^[18]。

5.1.2 真实数据集

Enron 数据集: 本文使用的是在文献[16]中使用的 Enron 数据集, 该数据集包括 150 个 Enron 公司高层管理人员的邮件数据, 并且只关注那些收发至少 5 封邮件的用户。同时, 该数据集去掉了那些在邮件中出现次数超过 19 000 次和低于 10 次的关键词。这样, 最终的数据集中包含 2 359 个收发邮件的用户以及 36 658 个主要关键词。本文只选取该数据集中从 2000 年 8 月 - 2001 年 10 月总共 15 个月的数据作为实验数据集。

DBLP 数据集: DBLP 是德国特里尔大学的 Michael Ley 负责开发和维护的计算机领域科学文献搜索服务的项目。本文从该项目中提取出 2001 年 - 2010 年在数据库、人工智能和数据挖掘 3 个领域的文献信息, 共同完成一份著作或论文的学者们被构造成完全图 (Clique), 去掉那些与其他节点没有连边的节点, 从而形成具有 10 个时刻的动态网络。所有 10 个时刻的网络都包含至少 3 000 个节点及 40 000 条边。

5.2 评价标准

采用 NMI 和社团质量 2 个指标作为评价标准。其中 NMI 主要是用来评价社团检测结果的准确度, 广泛应用于社团检测的评价标准中, 其计算公式如下:

$$NMI = \frac{H(X) + H(Y) - H(X, Y)}{(H(X) + H(Y)) / 2} \quad (8)$$

其中, $H(X)$ 与 $H(Y)$ 分别代表随机变量 X 与 Y 的熵; $H(X, Y)$ 代表两者的联合熵。最初的 NMI 只适用于非重叠社团, 文献[12]扩展了 NMI 的定义, 使得其能够应用于重叠社团。本文利用该扩展的 NMI 作为人工数据集上的实验的评价标准。

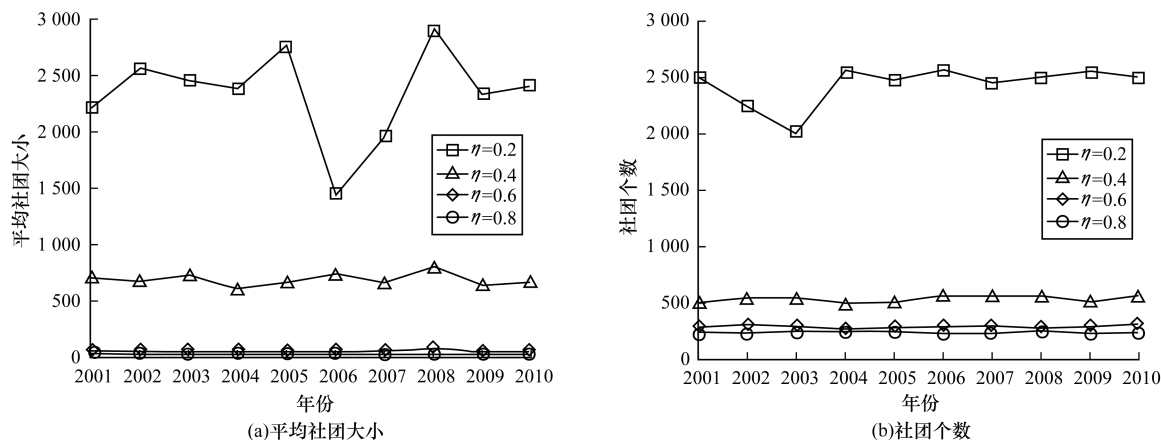
对于很多网络来说, 网络中的节点都带有一些原子属性。例如, 社交网络中每个节点表示一个人, 则这些节点就会带有诸如出生年月、出生地点、受教育情况等信息。将这些原子属性称为节点的质量属性。这些质量属性并不属于网络本身但却描述了节点, 所以可以假设“相似的”节点共享的质量属性要比“不相似的”节点共享的质量属性多。而社团质量是根据节点的质量属性来定义的, 其公式如下:

$$CQ(P_t) = \frac{\sum_{(n_1, n_2) \in P_t} |QA(n_1) \cap QA(n_2)|}{\sum_{(n_1, n_2) \in E_t} |QA(n_1) \cap QA(n_2)|} \quad (9)$$

其中, P_t 是 t 时刻网络 G_t 的一个划分; $QA(n)$ 是节点 n 的质量属性。CQ 的值越大, 说明将相似节点划分在同一个社团的程度越大, 也就是挖掘到的社团潜在的信息越大。

5.3 参数选择

本文算法含有 2 个参数 α 和 η , 其中 α 用来平衡快照质量和时序质量在时序平滑框架下的比重, η 用来指导局部社团的合并。 α 越大, 则算法更加侧重反映当前时刻网络的结构; α 越小, 则算法更加侧重匹配前一时刻社团结构。 α 越大, 后面时刻的社团结构拥有较大的快照质量, 而对前一时刻的社团结构匹配能力不足; 反之, α 越小, 则没有很好地考虑当前网络的真实结构, 但前面时刻社团结构却得到了充分的继承, 也就是说在社团演化的过程中, 社团不易死亡。而参数 η 越大, 2 个社团合并就越困难, 因而最终得到的社团的平均大小偏小; 反之, η 越小, 社团合并越简单, 最终得到的社团的平均大小偏大。实验证明, LEOD 算法对参数 η 更加敏感, 图 4 呈现了在 DBLP 数据集上, 当 α 取 0.9 时, 社团个数与社团平均大小随着 η 的变化情况。从图中可以看出随着 η 的增大, 最终得到的社团个数不断增多, 而社团平均大小不断减小。这是因为随着 η 的增大, 局部社团不易合并, 而只能加入到全局社团中去, 从而社团个数就会相应增大, 而平均社团大小则会相应变小。因为所选网络自身不规整的特点, 曲线有所波动。

图4 社团个数与平均社团大小随参数 η 的变化情况

5.4 准确率分析

为了比较算法的准确度,在人工网络上进行实验。图5给出在人工网络上3种算法准确度的对比。其中,图5(a)中的人工网络在相邻时刻涉及到社团演化过程中的合并与分裂;图5(b)中的人工网络在相邻

时刻涉及到社团演化过程中的社团的出现与死亡;图5(c)中的人工网络在相邻时刻涉及到社团演化过程中社团的暂时消失,而在其后某一时刻又重新出现的情况;图5(d)中的人工网络在相邻时刻涉及到社团演化过程中社团的增大与减小。

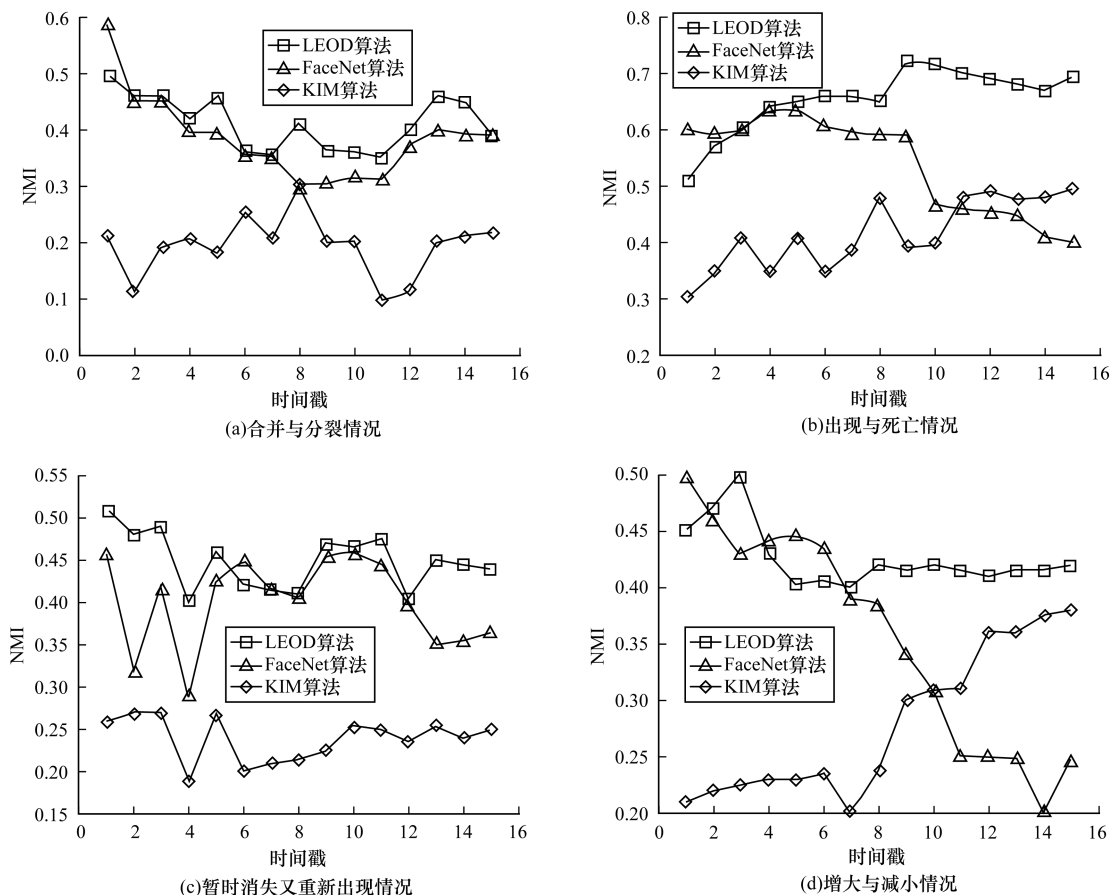


图5 3种算法在带有重叠社团的人工网络上的性能对比

从图5中可以看出,LEOD在处理这种带有重叠的动态网络时,要优于Facetnet,KIM这2种非重叠的演化聚类算法。这是因为Facetnet假定所有时刻网络的社团数目不会改变,所以在所有时刻都是

以一个固定的社团数目来对网络进行聚类。而社团演化的过程中往往伴随着社团的出现与消失,这样社团数目就不可能是一成不变的,所以FacetNet不能处理这种情况,从而影响其结果;KIM虽然脱离这

种固定数目社团的限制,但是只能处理非重叠社团。从图 5 可以看出,在第一个时刻 LEOD 相比 FacetNet 算法有一定的劣势,这是因为 LEOD 在第一个时刻没有引入前一个时刻的社团结构信息,且此时的快照质量有一定的损失,所以要稍逊于 FacetNet。但随着时间的推移,本文提出的算法逐渐表现出一定的优势,证明该算法更适合做动态网络社团的检测。

5.5 真实网络社团质量分析

选取 Enron 网络来评价 LEOD, FacetNet 及 KIM 算法所检测到的社团的质量。由于按照等式(9)计算所有社团时时间复杂度较大,因此从每个时刻的社团集中随机选择社团来进行计算。最终,3 种算法在 Enron 数据集 10 个连续时刻网络上分别得到的社团质量如表 2 所示。从表 2 中可以看出,LEOD 在社团质量这一指标上要明显优于其他 2 种算法,这是因为 KIM 算法是通过最大化全局模块度来划分整个网络,而模块度最大并不意味着能将“相似的”节点最大程度上地划分到同一个社团。而 FacetNet 计算 2 个节点之间的相似度时,以概率的方式从全局的角度来处理,相比局部视角将相似的节点划分成一个社团的能力,略有不足。同时, FacetNet 和 KIM 只能处理非重叠社团,社团质量也有一定的损失。LEOD 通过局部视角来划分社团,容易将相似的节点划分到同一个社团中,且最终得到的社团具有重叠性,从而能够得到较高的社团质量。

表 2 3 种算法在 Enron 网络下的社团质量

数据快照	LEOD 算法	FacetNet 算法	KIM 算法
1	7.398 30	1.059 070	0.195 597 0
2	12.670 30	0.914 754	1.181 440 0
3	7.377 30	1.197 220	0.063 060 0
4	15.569 30	0.702 273	2.519 490 0
5	10.155 20	0.751 334	0.598 959 0
6	6.380 67	1.230 460	0.009 964 0
7	9.283 30	0.939 550	0.043 907 5
8	12.240 70	0.635 500	0.098 141 6
9	3.080 77	0.638 924	0.016 777 9
10	6.903 08	0.838 946	0.004 105 6

6 结束语

本文提出一种新的演化聚类算法,该算法以局部视角来观察网络,从网络中的每个节点出发,首先利用标签传播算法得到该节点的 Ego 社团,然后引入社团相似度与社团关联度的概念,并利用演化聚类框架将这些 Ego 社团合并到全局社团集中。该方法能够有效地发现动态网络结构中的潜藏信息,将相似的节点尽可能地划分到同一个社团。同时,由于是对每个节点都操作一次,节点的 Ego 社团可能会有重叠,且

最终的全局社团是由 Ego 社团合并而成的,因此能够发现动态网络中的重叠社团。随后对动态网络中的社团进行了演化分析。在真实和人工数据集上的实验结果表明,本文方法对动态网络的社团发现和演化分析具有很好的效果。该方法的主要创新点如下:1)采用局部优先的社团检测方法,避免聚类前必须知道整个网络的结构,且更容易发现网络中潜藏的社团结构;2)通过对网络中每个节点处理一次,并合并这些节点的 Ego 社团,可以发现网络中的重叠社团;3)引入社团相似度和社团关联度 2 个概念,并利用时序平滑性框架,使得对动态网络的社团检测具有时序平滑性,进而对动态网络社团进行演化分析。

参考文献

- [1] Sun J, Faloutsos C, Papadimitriou S, et al. Graphscope: Parameter-free Mining of Large Time-evolving Graphs[C]//Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, USA: ACM Press, 2007: 687-696.
- [2] Telephone Traffic Data [EB/OL]. (2007-10-21). <http://reality.media.mit.edu/download.php>.
- [3] DBLP Bibliographic Data. (2009-05-06). <http://www.informatik.uni-trier.de/ley/db>.
- [4] Tantipathananandh C, Berger-Wolf T Y, Kempe D. A Framework for Community Identification in Dynamic Social Networks[C]//Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, USA: ACM Press, 2007: 717-726.
- [5] Chakrabarti D, Kumar R, Tomkins A. Evolutionary Clustering[C]//Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA: ACM Press, 2006: 554-560.
- [6] Chi Y, Song X, Zhou D, et al. Evolutionary Spectral Clustering by Incorporating Temporal Smoothness[C]//Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, USA: ACM Press, 2007: 153-162.
- [7] Lin Y R, Chi Y, Zhu S H, et al. FaceNet: A Framework for Analyzing Communities and Their Evolutions in Dynamic Networks[C]//Proceedings of International Conference on World Wide Web. Beijing, China: [s. n.], 2008: 685-694.
- [8] Kim M S, Han J. A Particle-and-density Based Evolutionary Clustering Method for Dynamic Networks[J]. Proceedings of the VLDB Endowment, 2009, 2(1): 622-633.
- [9] Coscia M, Rossetti G, Giannotti F, et al. Demon: A Local-first Discovery Method for Overlapping Communities[C]//Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China: [s. n.], 2012: 615-623.
- [10] Raghavan U N, Albert R, Kumara S. Near Linear Time Algorithm to Detect Community Structures in Large-scale Networks[J]. Physical Review E, 2007, 76(3).

(下转第 203 页)

- [8] Dev H. A User Interaction Based Community Detection Algorithm for Online Social Networks[C]//Proceedings of ACM SIGMOD International Conference on Management of Data. New York, USA:ACM Press,2014:1607-1608.
- [9] Weng W, Zhu S, Xu H. Hierarchical Community Detection Algorithm Based on Local Similarity[J]. Journal of Digital Information Management, 2014, 12(4):275-286.
- [10] Tian Y, Hankins R A, Patel J M. Efficient Aggregation for Graph Summarization [C]//Proceedings of ACM SIGMOD International Conference on Management of Data. New York, USA:ACM Press,2008:567-580.
- [11] Dang T A, Viennet E. Community Detection Based on Structural and Attribute Similarities[C]//Proceedings of International Conference on Digital Society. New York, USA:ACM Press,2012:7-12.
- [12] Clauset A, Newman M E J, Moore C. Finding Community Structure in Very Large Networks[J]. Physical Review E,2004,70(6).
- [13] Xu Z, Ke Y, Wang Y, et al. A Model-based Approach to Attributed Graph Clustering[C]//Proceedings of ACM SIGMOD International Conference on Management of Data. New York, USA:ACM Press,2012:505-516.
- [14] Zhou Y, Cheng H, Yu J X. Graph Clustering Based on Structural/Attribute Similarities[J]. Proceedings of the VLDB Endowment,2009,2(1):718-729.
- [15] Lü L, Zhou T. Link Prediction in Complex Networks: A Survey[J]. Physica A: Statistical Mechanics and Its Applications,2011,390(6):1150-1170.
- [16] 张春英. 基于属性图的社交网络建模与态势分析理论研究[D]. 秦皇岛:燕山大学,2013.
- [17] Ma N, Guan J, Zhao Y. Bringing PageRank to the Citation Analysis[J]. Information Processing & Management,2008,44(2):800-810.
- [18] Arasu A, Cho J, Garcia-Molina H, et al. Searching the Web[J]. ACM Transactions on Internet Technology, 2001,1(1):2-43.
- [19] Yan E, Ding Y. Discovering Author Impact: A Page Rank Perspective[J]. Information Processing & Management,2011,47(1):125-134.
- [20] Yu X, Yang J, Xie Z Q. A Semantic Overlapping Community Detection Algorithm Based on Field Sampling[J]. Expert Systems with Applications,2015,42(1):366-375.
- [21] Kaufman L, Rousseeuw P. Clustering by Means of Medoids[M]. [S. l.]:North-Holland Press,1987.
- [22] Adamic L A, Glance N. The Political Blogosphere and the 2004 US Election: Divided They Blog [C]//Proceedings of the 3rd International Workshop on Link Discovery. New York, USA:ACM Press,2005:36-43.

编辑 索书志

(上接第195页)

- [11] Greene D, Doyle D, Cunningham P. Tracking the Evolution of Communities in Dynamic Social Networks[C]//Proceedings of IEEE International Conference on Advances in Social Networks Analysis and Mining. Odense, USA:IEEE Press,2010:176-183.
- [12] Lancichinetti A, Fortunato S, Kertész J. Detecting the Overlapping and Hierarchical Community Structure in Complex Networks[J]. New Journal of Physics,2009,11(3).
- [13] 付立东,高 琳,马小科. 基于社团检测的复杂网络中心性方法[J]. 中国科学(信息科学),2012,42(5):550-560.
- [14] 姜雅文,贾彩燕,于 剑. 基于节点相似度的网络社团检测算法研究[J]. 计算机科学,2011,38(7):185-189.
- [15] 易三莉,苗 莹,钱 洁,等. 基于感兴趣区域的边缘结构相似度图像质量评估[J]. 计算机工程,2015,41(6):216-220.
- [16] 杨建新,周献中,葛银茂. 基于拉普拉斯图谱和 K 均值的多社团发现方法[J]. 计算机工程,2008,34(12):178-180.
- [17] 王观玉. 基于聚类的复杂网络社团发现算法[J]. 计算机工程,2011,37(10):58-60.
- [18] Fortunato S. Community Detection in Graphs [J]. Physics Reports,2010,486(3):75-174.
- [19] Huang Jianbin, Sun Heli, Han Jiawei, et al. SHRINK: A Structural Clustering Algorithm for Detecting Hierarchical Communities in Networks [C]//Proceedings of the 19th ACM International Conference on Information and Knowledge Management. Toronto, Canada:ACM Press,2010:219-228.
- [20] Kumar R, Novak J, Raghavan P, et al. On the Bursty Evolution of Blogspace [J]. World Wide Web, 2005, 8(2):159-178.
- [21] Leskovec J, Kleinberg J, Faloutsos C. Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations[C]//Proceedings of ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. Chicago, USA:ACM Press,2005:177-187.
- [22] Palla G, Barabási A L, Vicsek T. Quantifying Social Group Evolution[J]. Nature,2007,446(7136):664-667.
- [23] Yang T, Chi Y, Zhu S, et al. A Bayesian Approach Toward Finding Communities and Their Evolutions in Dynamic Social Networks [C]//Proceedings of SIAM International Conference on Data mining. Sparks, USA:SIAM,2009:990-1001.
- [24] Cuzzocrea A, Folino F. Community Evolution Detection in Time-evolving Information [C]//Proceedings of Joint EDBT/ICDT'13 Workshops. Genoa, Italy:ACM Press, 2013:93-96.
- [25] Tang L, Liu H, Zhang J, et al. Community Evolution in Dynamic Multi-mode Networks [C]//Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA:ACM Press,2008:677-685.

编辑 索书志