

基于事件框架的社区进化预测研究

张学武^{1),2)} 沈浩东¹⁾ 赵沛然¹⁾ 张卓¹⁾ 李敏¹⁾ 许海燕¹⁾

¹⁾(河海大学物联网工程学院 江苏 常州 213022)

²⁾(江苏省“世界水谷”与水世界生态文明协同创新中心 南京 211100)

摘 要 有效预测社交网络中的社区演化规律和趋势,在广告精准投放、网络舆论监测与引导方面具有广泛的应用前景.近年来,基于事件框架的社区进化预测建模在反映社区演化规律和趋势方面一直是热点,研究的难点在于提高预测模型的预测准确度.为解决这一难点,文中首先提出一种改进的事件框架,以新型事件框架为基础,针对不同事件分别构建有效的预测模型.其中预测模型的输入指标包括网络结构特征、社区的结构特征和社区进化特征.最后为了充分验证所建预测模型的有效性并保证预测模型的实用价值,在实验中分别用人造动态网络数据集、DBLP 动态网络数据集和 Facebook 数据集作为实验数据集,以确保实验数据集的多样性.实验结果表明文中的预测模型针对形成、消失、保持、合并和分裂等事件的预测具有较高的准确性,结果也表明文中提出的特征指标和对应事件的预测模型在实际事件预测中将具有较高的实用价值.

关键词 复杂网络;社区进化;特征指标;预测模型;社交网络

中图法分类号 TP393 DOI号 10.11897/SP.J.1016.2017.00729

Research on Community Evolution Prediction Based on Event-Based Frameworks

ZHANG Xue-Wu^{1),2)} SHEN Hao-Dong¹⁾ ZHAO Pei-Ran¹⁾ ZHANG Zhuo¹⁾ LI Min¹⁾ XU Hai-Yan¹⁾

¹⁾(College of Internet of Things Engineering, Hohai University, Changzhou, Jiangsu 213022)

²⁾(Jiangsu Province Collaborative Innovation Center of World Water Valley and Water Ecological Civilization, Nanjing 211100)

Abstract Research on community detection is one of the fundamental investigations in social networks, and event-based frameworks for characterizing the community evolution and revealing the tendency of communities in social networks can be lately seen in some international workshops, conferences and seminars. To improve the predictive accuracy of predictive models based on event frameworks is a difficult part for present research, therefore, in this paper we proposed an improved event-based framework. According to the improved definition of all kinds of events, we presented some effective predictive models which have many important applications, such as recommending advertisements and guiding public opinions. To find appropriate predictive models which are able to predict community events with higher accuracy, we presented several features involved social network structure, community structure and community evolution features. In our experiments, we use real world datasets including artificial dataset, DBLP datasets and Facebook datasets which are classical and often used datasets to validate the efficiency of our predictive models. The experimental results confirmed that the predictive models with features presented by us can predict the evolution of communities including birth, death, remain, merge and split accurately, and the results also show that the predictive models would have high practical value.

Keywords complex networks; community evolution; features; predictive model; social networks

收稿日期:2015-11-18;在线出版日期:2016-06-17. 本课题得到国家自然科学基金(61671202,61573128,61273170)、国家“九七三”重点基础研究发展规划项目基金(2016YFC0401606)和中央高校基本科研业务费专项资金(2015B25214)资助. 张学武,男,1973年生,博士,教授,主要研究领域为智能感知与计算、大数据分析和可视化. E-mail: lab_112@126.com. 沈浩东,男,1991年生,硕士研究生,主要研究方向为社交网络分析、社区发现. 赵沛然,男,1994年生,硕士研究生,主要研究方向为社区发现. 张卓,女,1978年生,博士,主要研究方向为机器学习、社交网络理论和大数据分析研究. 李敏,女,1982年生,博士,主要研究方向为机器学习和数据挖掘. 许海燕,女,1978年生,博士,主要研究方向为信息处理和人工智能.

1 引言

复杂网络分析一直以来是研究人员关注的重点,诸多领域的学者分别从不同的角度对网络进行分析,通过发现网络中的社区去分析整个网络,是大家关注的方法之一,也是领域内研究的热点和难点.所谓社区,即节点和边的集合.在社区内部,节点连接很紧密.但是,属于不同社区的节点连接则比较稀疏^[1-2].在社交网络中,节点一般表示单个个体,边则表示个体之间存在的关系,如同学关系、亲属关系和朋友关系等,而社区则代表具有共同兴趣爱好或其他共同属性的朋友圈.

早期关于社区的研究主要集中在静态网络的社区发现,并涌现出大量的算法^[3-5].其中 Clauset 和 Newman 等人^[6]提出的 CNM 算法就是一种典型的社区发现算法.该算法是基于模块度的社区发现算法,同时也是迄今为止时间复杂度最低的算法之一.然而该算法是一类非重叠社区发现算法(单个节点只能属于一个社区),实际情况下,在多数复杂网络中,尤其是社交网络,不同的个体可以分属于多个社区.2005 年 Palla 等人^[7]提出了第一个重叠社区发现算法,即派系过滤算法(Clique Percolation Method, CPM).此后,关于社区的研究才开始逐步转向重叠社区领域. CPM 算法中定义了 K-派系,它将相邻的 K-派系进行合并,而那些同时属于不同 K-派系的节点则组成了网络中社区的重叠区域. CPM 算法可以成功的检测网络中的重叠社区,然而最终的结果受 k 值影响. k 值越大,派系之间很难满足存在 $k-1$ 个共同节点,因此会导致社区规模越小;而 k 值过小,又会造成社区规模过大.此外, Zhang 等人^[8]于 2015 年对 CNM 算法进行了相应的改进使之适用于检测重叠社区.该算法不仅保留 CNM 算法时间复杂度低这一优点,而且提高了社区划分准确度.

近些年,对于社区的研究集中于动态网络的社区进化,主要研究动态网络中社区结构的变化过程^[9-10].早期人们研究网络中的社区结构时都将复杂网络认为是静态的,然而实际上任何一个复杂网络尤其是社交网络都是在变化的^[3].由于社交网络的动态性,研究其中社区进化的过程更具有实用价值.例如:在 Facebook 中随时会有新增用户,但相应的也会有用户注销账号;用户会和其他人建立或者解除好友关系,整个网络包含的社区数目可能增加或减少.因此通过对网络中社区的动态性及其进化过

程的研究,对于更好地理解网络的内部结构有极大的帮助.关于社区进化的研究主要包括以下两类:动态网络中的社区发现,典型的算法包括 Graph Scope 算法^[11]、Facet Net 算法^[12]等;动态网络中社区的变化过程,最典型的是基于事件框架的社区进化算法.

动态网络中社区变化过程的研究算法中,最重要的一类是基于事件框架的社区进化算法.所谓事件,一般表示为当前时刻网络中社区或节点的相互影响在下一时刻网络中的反映.最早的算法是由 Palla 等人^[13]于 2007 年提出的,他们通过对 CPM 算法进行改进,用于识别动态网络中与社区相关的各类事件.算法主要是研究相邻时刻网络中社区的变化,从而定义出各类事件. Asur 等人^[14]提出了一个事件框架用于描述不同时刻社区间的各种关系,并定义了 5 类社区事件: Merge、Split、Form、Dissolve 和 Continue.首先,从相邻时刻的两个网络中提取出各类事件,然后基于事件分别分析其中的社区,从而找出它们的变化过程和规律.此外, Takaffoli 等人^[15-17]也提出了类似的事件框架,并对实际网络数据集进行分析处理,而且 Takaffoli 等人^[18]还于 2014 年提出了一个模型用于预测网络中发生的各类事件.此外,基于事件框架 Greene 等人^[19]提出了一个有效地识别动态网络中社区的算法,并且设计了人工合成动态网络的方法.

本文提出一种基于事件框架的重叠社区进化算法.主要针对现有事件定义过于严格,仅考虑相邻网络的社区变化等问题进行相应的改进.为此本文重新定义了各类事件.首先将动态网络视为一系列静态网络的组合,利用 Zhang 等人^[8]提出的重叠社区发现算法检测出各个网络中的社区;然后使用新的事件定义分析社区发生的事件类型,从而发现网络中社区的演化过程 and 变化规律.本文还定义了 5 个与社区相关的指标,根据这些指标与事件之间的关系,通过不同分类器构建预测模型,实验结果表明,该模型具有较好的预测准确度,同时也证明了指标选取的合理性.

2 相关知识

目前,基于事件框架来分析社区的形成变化过程已经成为了一个动态网络的研究热点.研究人员提出了诸多关于事件的定义,然而已知的各类涉及事件的定义均存在着不少问题.例如, Asur 等人^[14]提出的事件框架中关于社区保持、合并和分裂 3 类

事件的定义过于严格,导致满足这些事件发生条件的社区数目相当少.针对 Asur 等人提出的事件框架定义的局限性,结合 Palla 等人^[13]的研究工作,首次提出使用相似度这一指标来定义社区事件,并定义了 7 类社区相关的事件.为确保事件定义的准确,通过实际实验确定定义中的参数 α 和 β ;而对于 Takaffoli 事件框架^[17]而言,可以较好地识别各类事件,但是由于其计算社区是否发现某一事件时,将综合考虑当前时刻网络与剩余时刻所有网络,这使得整个过程十分耗时,而对于节点众多的社交网络这一缺点将会被放大.为此,文中以降低事件检测的时间复杂度为目标,不考虑当前时刻网络与剩余时刻所有网路,通过实验确定只需要考虑 t 个时间间隔即可得到相近的实验结果;另外,为了更好地理解事件发生的内在规律,文中提出了 5 类涉及社区的指标,并根据指标与事件间存在的关系构建出相应的预测模型.不同于文献^[18,20-21],文中只定义了 5 类低重复度的指标,来实现高准确度的事件预测.

2.1 重叠社区

众所周知,社区可视为网络中同一类节点和边的集合,同一社区内部的节点连接较为紧密,而不同社区的节点连接相对比较稀疏.重叠社区则是一种特殊的社区,在重叠社区中节点可以同时属于多个社区^[7,22-23].在实际社交网络中可理解为某个人同时拥有多个朋友圈,而对他而言各个朋友圈都是同等重要的.图 1 展示了两个不同的重叠社区 C_1 和 C_2 ,它们都包含节点 6.

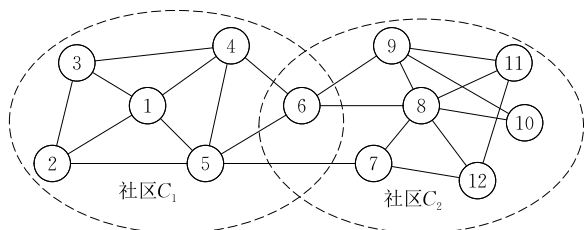


图 1 重叠社区示意图

2.2 动态网络

动态网络可以认为是一系列静态网络的集合,因此可以将它表示为集合,其中 $G_t = (V_t, E_t)$ 代表第 i 个时刻的静态网络, V_t 和 E_t 分别表示网络中节点和边的集合^[12,24].而 G_t 中检测到的 n 个重叠社区同样可以表示为集合 $C_t = \{C_t^1, C_t^2, \dots, C_t^m\}$,同理其中的某一社区 C_t^m 可表示为 (V_t^m, E_t^m) .图 2 是一个动态网络的例子,该网络包含两个时刻的静态网络.其中 $T=1$ 时刻网络由 3 个重叠社区 C_1^1, C_2^1 和 C_3^1 组

成, $T=2$ 时刻网络则由另外 3 个社区构成,社区的检测是由上文提到的重叠社区算法实现的.

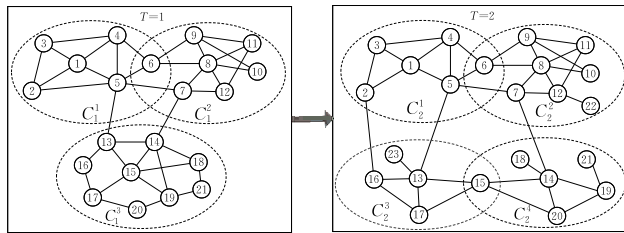


图 2 动态网络示意图

2.3 社区相似度

本文为了衡量不同时刻动态网络中两个社区的相似程度,使用了社区相似度指标,定义公式如式(1)所示.

$$Sim(C_i^a, C_j^b) = \frac{|V_i^a \cap V_j^b|}{\max(|V_i^a|, |V_j^b|)} \quad (1)$$

其中 C_i^a, C_j^b 分别表示第 i, j 时刻网络中的社区,而 $|V_i^a|$ 和 $|V_j^b|$ 代表两个社区中节点数目, $|V_i^a \cap V_j^b|$ 是它们之间的相同节点数.社区相似度的值在 0 到 1 之间,当值大于某一阈值时,可认为两个社区是相似的,阈值的确定将在实验部分进行详细说明.

3 基于事件框架的社区特征指标

事件严格的定义是指当前时刻网络中社区或 $\{G_1, G_2, \dots, G_n\}$ 节点之间的相互影响在下一时刻网络中的反映^[14].本文重新定义了 7 类涉及社区的事件,包括保持(Remain)、合并(Merge)、分裂(Split)、形成(Birth)、消失(Death)、收缩(Contract)和扩张(Expand).这些事件涵盖了动态网络中不同时刻网络的社区可能会出现的变化.

图 3 直观展示了 7 种与社区相关的事件产生过程,图中包含 6 个时刻的网络,每两个相邻时刻网络的社区间发生一种事件.从 $T=1$ 时刻到 $T=2$ 时刻之间,社区发生了保持事件,不难看出社区 1 中节点保持不变.但这是一类理想的情况,实际中,不同时刻网络中的社区通常不可能完全一致,严格的定义也导致了很难检测到基于 Asur 等人定义的事件,因此文中定义了社区相似度,当两个社区相似而非完全相同时,即可认为该社区发生可保持事件.此外,保持事件也不仅仅局限于通过判断相邻时刻的网络社区进行确定.如果在当前时刻存在一个社区,但是在其后 t 个时刻中某一时刻该社区存在,则认为该社区保持(Remain).

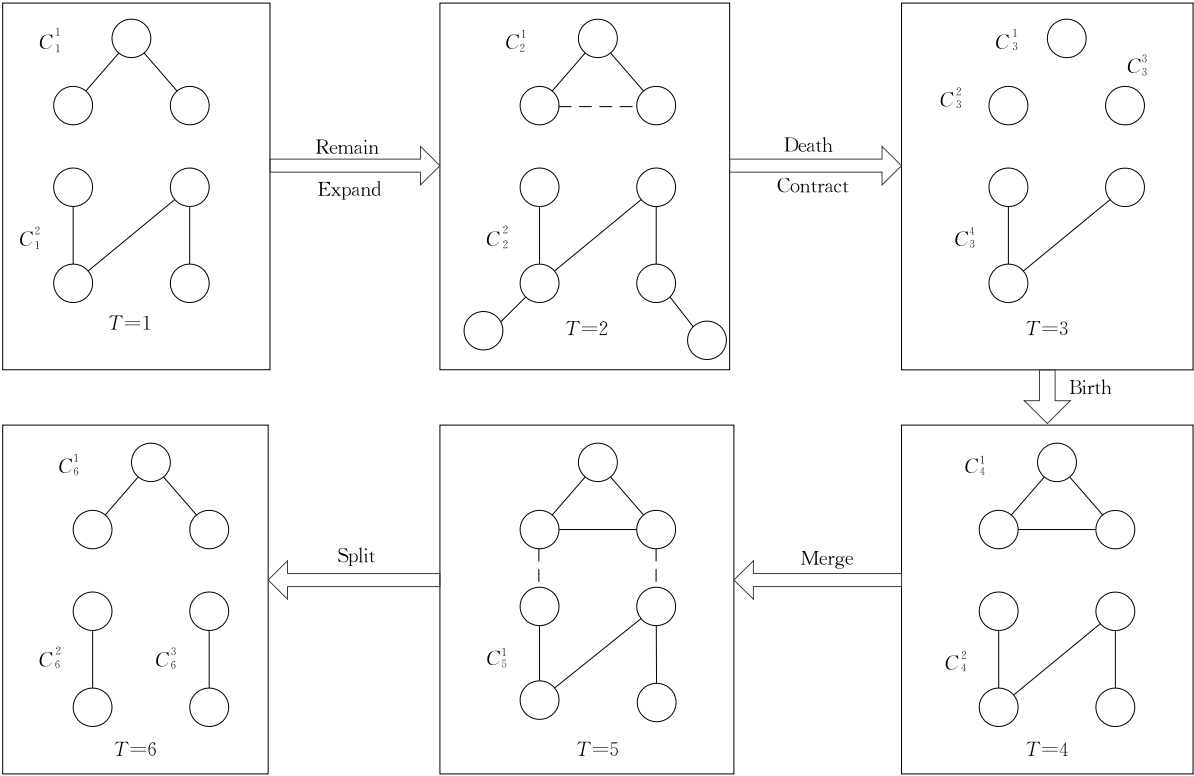


图 3 7 类社区相关的事件产生示意图

考虑这样一种情况, $T=1$ 时刻社区 1 存在, 而在 $T=2$ 时消失, 但是在 $T=3$ 时重新产生. 仅仅通过相邻时刻判断社区是否形成, 则会认为社区 1 在 $T=3$ 时发生了形成事件, 但实际上该社区在 $T=2$ 时只是暂时消失, 而且考虑到社区发现算法可能存在误差如在该时刻没有检测到社区, 社区不应该被认为发生保持事件. 因此社区是否保持应判断在其后某时刻该社区是否存在, 然而对于网络规模较大的动态网络, 这会使得计算时间成倍增长, 因此本文采用判断后 t 个时刻该社区是否存在来确定社区保持事件的发生. 同理, 如果当前时刻存在一个社区, 但是在前 t 个时刻该社区均不存在, 则认为该社区形成(Birth). 图 3 中 $T=4$ 到 $T=5$ 时刻网络间的社区发生了形成事件, 一个新的社区 C_5^1 诞生. 如果当前时刻存在一个社区, 但是在其后 t 个时刻该社区均不存在, 则认为该社区消失(Death). 图 3 中 $T=5$ 到 $T=6$ 时刻网络间的社区发生了消失事件, 社区 C_5^1 消亡, 分解为单个节点. 不同于 Asur 等人的定义, 事件保持(Remain)、形成(Birth)和消失(Death)在已有的定义基础上, 通过判断 t 个时刻网络社区的变化确定该事件是否发生. 但是对于合并(Merge)和分裂(Split)事件而言, 事件发生只存在于相邻时刻更符合实际情况. 但这 5 个事件的检测都是基于检测到的社区与原社区的相似度来确定是

否发生某一事件, 实验证明本文所提出的定义能够检测出更多的社区事件.

当某一时刻存在一个社区, 该社区是由前一时刻多个社区所组成的, 则认为该社区发生合并事件(Merge). 如图 3 中 $T=2$ 时刻中的两个社区在 $T=3$ 时刻合并为一个社区, 因此可认为这两社区发生合并事件. 当某一时刻存在一个社区, 该社区是由下一时刻多个社区所组成的, 则认为该社区发生分裂事件(Split). 图 3 中 $T=3$ 时刻的社区在 $T=4$ 时刻分解为 3 个小社区, 因此该社区发生分裂事件. 当某一时刻存在一个社区, 该社区的规模在下一时刻变大, 则认为该社区发生扩张事件(Expand). 图 3 中 $T=1$ 时刻的黄色的社区在 $T=2$ 时刻, 社区节点数目从 4 增加到 6, 因此该社区发生了扩张事件. 当某一时刻存在一个社区, 该社区的规模在下一时刻缩小, 则认为该社区发生收缩事件(Contract). 图 3 中 $T=2$ 时刻的黄色社区在 $T=3$ 时刻, 社区节点数目从 6 减少到 3, 一次该社区发生了收缩事件. 关于各类事件的详细定义及其计算公式将在下文详细展开.

3.1 事件的定义

3.1.1 形成

如果社区 C_i^a 与其前 t 个时刻网络中的任何一个社区都不相似(即两社区间相似度不大于 α), 则认为社区发生形成事件. 计算公式如下:

$Birth(C_i^a) = \text{true}$ iff $\exists \text{ no } j < i \text{ } Sim(C_i^a, C_j^b) \geq \alpha$ (2)

其中为保证时刻 j 为整数, j 的取值是 $i-1$ 和 1 之间的最大值。

3.1.2 消 失

如果社区 C_i^a 与其后 t 个时刻网络中的任何一个社区都不相似, 则认为社区发生消失事件. 计算公式如下所示:

$Death(C_i^a) = \text{true}$ iff $\exists \text{ no } j > i \text{ } Sim(C_i^a, C_j^b) \geq \alpha$ (3)

其中为保证时刻 j 有意义, j 的取值是 $i+t$ 和最大网络时刻数之间的最小值。

3.1.3 保 持

如果社区 C_i^a 与其后 t 个时刻网络中的某一个社区相似, 则可以认定社区发生保持事件. 计算公式如下:

$Remain(C_i^a) = \text{true}$ iff $\exists j > i \text{ } Sim(C_i^a, C_j^b) \geq \alpha$ (4)

其中为保证时刻 j 有意义, j 的取值是 $i+t$ 和最大网络时刻数之间的最小值。

3.1.4 合 并

如果社区 C_i^a 与其前一个时刻网络中多个社区的并集相似, 则认为社区发生合并事件. 这些社区可表示为 $\{C_i^1, C_i^2, \dots, C_i^k\}$, 其中 C_i^l 代表其中的任意一个社区, 计算公式定义如下:

$Merge(C_{i+1}^a) = \text{true}$ iff

$$\forall C_i^l \in \{C_i^1, C_i^2, \dots, C_i^k\} \frac{|V_i^l \cap V_{i+1}^a|}{|V_i^l|} \geq \alpha$$

$$\text{and } Sim\{C_{i+1}^a, (C_i^1 \cup C_i^2 \cup \dots \cup C_i^k)\} \geq \alpha \quad (5)$$

3.1.5 分 裂

如果社区 C_i^a 与其后一个时刻网络中多个社区的并集相似, 则认为社区发生分裂事件. 这些社区可表示为 $\{C_{i+1}^1, C_{i+1}^2, \dots, C_{i+1}^k\}$, 计算公式定义如下:

$Split(C_i^a) = \text{true}$ iff

$$\forall C_{i+1}^l \in \{C_{i+1}^1, C_{i+1}^2, \dots, C_{i+1}^k\} \frac{|V_{i+1}^l \cap V_i^a|}{|V_{i+1}^l|} \geq \alpha$$

$$\text{and } Sim\{C_i^a, (C_{i+1}^1 \cup C_{i+1}^2 \cup \dots \cup C_{i+1}^k)\} \geq \alpha \quad (6)$$

其中 C_{i+1}^l 代表其中的任意一个社区。

3.1.6 扩 张

如果社区 C_i^a 在下一时刻变大, 则认为社区发生扩张事件, 而其扩张的规模同样由参数 β 进行控制。

$Expand(C_i^a) = \text{true}$ iff $1 + \beta \geq \frac{|V_i^a \cap V_{i+1}^b|}{|V_i^a|} \geq 1$ (7)

3.1.7 收 缩

如果社区 C_i^a 在下一时刻缩小, 则认为社区发生收缩事件, 而其缩小的规模由参数 β 进行控制, β 值将在实验环节进行确定。

$Contract(C_i^a) = \text{true}$ iff $1 \geq \frac{|V_i^a \cap V_{i+1}^b|}{|V_i^a|} \geq 1 - \beta$ (8)

在社交网络中, 社区发生形成事件, 可理解为新的朋友圈的产生或是新的热点话题的形成. 研究网络中社区的形成事件, 有利于了解网络中人群新的关注热点. 而社区的消失, 则表明了网络中朋友圈的解散或是热点话题的消亡. 社区的保持说明了该社区本身具有很强的生命力, 可认为是朋友圈的保持或是热点话题的延续. 对于社区保持的研究有助于分析和发现网络中人们的兴趣所在以及当前的热点. 社区的分裂, 是指网络中大的社区分裂为多个小社区, 可理解为对热点话题的讨论分裂为意见相左的不同人群. 社区的合并, 可以认为是多个相近的热点话题合并为一个大的话题. 社区的扩张, 表明有更多的人加入热点话题的讨论. 社区的收缩, 说明参与话题讨论的人逐渐变少。

3.2 预测模型特征指标的定义

各类事件的详细定义已经给出, 然而对于这些事件产生的机理还不够清楚. 因此提出 5 个指标用于刻画各类事件, 分别是社区平均路径长度 Apl (Average path length)、社区聚类系数 CC (Clustering Coefficient)、社区节点边相对比 RNE (Ratio of Nodes and Edges)、社区流行度 Pop (Popularity) 和社区社交度 Soc (Sociability). 并使 Weka 统计不同分类器对特定事件进化的预测准确性, 实验结果表明预测模型有较好的准确度, 同时也证明指标的选取较为合理。

3.2.1 社区平均路径长度 (Average path length)

$$Apl(C_i^a) = \frac{2}{n(n+1)} \sum_{i=1}^n \sum_{j \geq i}^n d_{ij} \quad (9)$$

其中 d_{ij} 表示社区 C_i^a 中两个节点 i 和 j 之间的最短距离, n 是社区中节点数目, 社区中所有节点对的最短距离的平均数就是整个社区的平均路径长度. 平均最短路径一般用于分析复杂网络. 社区平均路径长度越短, 说明社区中任意两点距离越小, 整个社区节点连接紧密, 在动态网络中社区容易发生保持等事件; 平均路径长度越长, 在一定程度上表示社区节点连接较为稀疏, 因此社区可能会发生分裂、消失等事件. 该指标通常情况下与 C_i^a 社区聚类系数一同使用。

3.2.2 社区聚类系数 (Clustering coefficient)

$$CC(C_i^a) = \frac{1}{n} \sum_{i=1}^n \frac{2e_i}{k_i(k_i-1)} \quad (10)$$

其中节点 i 有 k_i 个邻接节点, e_i 表示这 k_i 个节点间实际存在的边数, 而这些节点之间最多可能存在 $k_i(k_i-1)/2$ 个节点, 两者比值就是节点 i 的聚类系

数,社区中所有节点聚类系数的平均数就是社区聚类系数. 聚类系数是一个常用的评价复杂网络的指标,可用于描述网络中节点之间集结成社区的程度,即一个节点的邻接节点之间连接的程度. 例如在社交网络中,该指标表示你的好友之间彼此认识的几率. 社区聚类系数越大,表明这个社区中节点连接相当紧密,社区更倾向于发生合并、保持、扩张等事件;聚类系数越小,表明社区间节点连接较为稀疏,社区可能发生分裂、消失和收缩等事件.

3.2.3 社区节点边相对比(Ratio of nodes and edges)

$$RNE(C_i^a) = \frac{|V_i^a|}{|E_i^a|} \quad (11)$$

其中 $|V_i^a|$ 是社区中总节点数,而 $|E_i^a|$ 是其中的总边数. 该指标同样可以反映社区内部的紧密程度. 当 RNE 值很大时,说明该社区可能人数虽然很多,但是相互之间联系很少,关系相对比较冷淡,因而可以推断出,此时社区有很大的趋势分为几个小的团体;当 RNE 值很小时,说明该社区在保持一定人数的同时,相互之间的联系还比较频繁,因而可以推断出,此时社区有很大的趋势继续生存下去或者合并其他小团体.

3.2.4 社区流行度(Popularity)

$$Pop(C_i^a) = \sum_{u=1}^m Join(u, C_i^a) - \sum_{u=1}^m Leave(u, C_i^a) \quad (12)$$

其中 $Join(u, C_i^a)$ 指节点 u 在当前时刻属于社区 C_i^a ,而在前一时刻不属于社区 C_{i-1}^a ,其中社区 C_i^a 和社区 C_{i-1}^a 是相似的. $Leave(u, C_i^a)$ 指节点 u 在当前时刻属于社区 C_i^a ,而在下一时刻不属于社区 C_{i+1}^a ,社区和社区 C_{i+1}^a 也是相似的. 社区流行度表示该社区在时间段 $[i, i+1]$ 内吸引的节点数目,即该时间段内加入社区的节点总数与离开社区节点总数的差值. 该指标主要反映社区对于节点的吸引程度. 社区流行度越高,说明该社区正处于蓬勃发展中,社区可认为是当前的热点话题,对人们有很大的吸引力,社区倾向于发生扩张、合并等事件;社区流行度越低,表示社区开始逐渐萎缩,社区可能发生收缩、分裂等事件.

3.2.5 社区社交度(Sociability)

$$Soc(C_i^a) = \sum_{u=1}^k \frac{\sum_{j=1}^i (Join(u, Cl_{j+1}^a) + Leave(u, Cl_j^a))}{|Activity(u)|} \quad (13)$$

其中 Cl_j^a 是指节点 u 在 j 时刻所属的社区,而 $|Activity(u)|$ 则是节点 u 存在的时间. 社区社交

度是社区内所有节点的社交度之和,点的社交度等于截止到当前时刻该点加入和离开事件的总和与点存在时间的比值. 这个指标可以反映出社区内节点的活跃程度,其中经常变换所属社区的节点被称为活跃节点. 当一个社区的 Soc 的值很大时,表明这个社区大部分的点都有着不断变换社区的趋势,因此社区容易发生分裂、收缩甚至消失等事件.

预测模型的构建使用了Weka^[25]软件. 通过计算各个社区的5类指标以及是否发生相应的事件,从而为指标集和事件建立了对应的联系,然后为每一个涉及社区的事件都分别建立了对应的预测模型. 研究社区进化的主要目的就是了解动态网络中社区变化的过程,而预测模型的构建主要根据现有数据分析出各指标集与事件之间的映射关系,并在给出各指标的情况下,根据预测模型,可以分析出各社区最可能发生的事件,从而了解网络的变化趋势.

4 实验结果及分析

4.1 实验设置

社区进化算法对比实验主要是为了衡量本文提出的事件框架在处理具体数据时的效果. 因此我们使用了Asur框架^[14]、Takaffoli框架^[17]与本文建立的框架进行对比实验,实验分为验证各类事件数目和预测模型的准确度两部分. 在实验中,我们针对各类事件的检测都是单独进行的,比如下一个时刻的两个或者多个社区的并集与当前时刻的某一社区相似则发生了分裂事件,此时负责统计分裂事件次数的变量加一,于此类似其他事件也是通过一个单独的变量来统计每个时刻各类事件发生的次数. 另外为了更好地描述社区的变化,我们给各个社区都作了标号. 首先对第一个时刻的数据集中检测到的社区进行标号,然后在接下来的时刻将检测到的社区都和已有标号的社区作相似度比较,相似的则标注相同的标号,否则在当前最大标号的基础上加一. 显然不同时刻的社区具有同一标号时则发生保持(Remain)事件. 而预测模型准确度是指该模型预测某一特定事件在下一时刻会发生的可能性,输入特征指标后准确度的大小可以通过Weka^[25]直接得到. 在Weka中,我们使用十折交叉校验法^[26]去得到各个模型的预测准确度. 我们把检测到的社区平均分成十份,其中9份作为训练集剩下的一份作为测试集. 其中每个社区都具有5个特征指标和对应

的响应事件. 重复试验十次取均值, 最终得到各个模型对应各事件的准确度. 为了比较各个事件框架的优劣, 在实验中采用了两个人造动态网络数据集、DBLP 动态网络数据集和 Facebook 数据集^[27]. 人造网络数据集是采用 Greene 等人^[19]提供的动态网络生成工具产生的, 表 1 展示了生成网络时设定的各类参数值. 其中 Database1 将用于调节事件中的各类参数, Database2 用于检测各个框架发现的事件数目. DBLP 动态网络数据集是 DBLP 中文数据集, 它是一个科学家合作者网络, 每个节点表示一名学者, 边代表两名科学家一同发表过文章, 原始数据集可以从 <http://www.datatang.com/> 下载. 表 2 展示该数据集中包括的节点数和总边数. 本文采用的 Facebook 数据集来自 Viswanath 等人^[27]提供的, 该数据采集 2006 年至 2009 年间 New Orleans 地区的朋友关系网络, 总共包含 60 290 名用户和 1545 686 条边. 因此我们从中抽取 2007 年 1 月至 10 月的数据集, 以月为单位构建了包含 10 个网络的 Facebook 数据集. 本文提到的算法主要使用 java 和 matlab 语言实现, 预测模型的构建采用了 Weka, 实验环境为 Intel CPU 3.0 GHz 主频和 4 GB 内存.

表 1 生成两类人造动态网络数据集的参数及其定义			
参数	定义	Dataset1	Datasets
s	时间步长	10	10
n	节点数	5000	10 000
k	平均度	30	40
$k_{\max}$	最大度	50	60
$C_{\max}$	最大社区规模	50	75
$C_{\min}$	最小社区规模	30	30
u	混合参数	0.2	0.2

表 2 DBLP 动态网络数据集中节点数和边数		
年份	节点数	边数
2005	1038	2858
2006	1283	3874
2007	1253	3876
2008	1373	3976
2009	1217	3136

4.2 参数调节

在事件的定义中用到了 3 个参数, 分别是 t , α 和 β , 它们的取值将通过实验确定. 其中由于参数 t 只与事件 Remain、Birth 和 Death 有关, 因此主要通过对这 3 个事件在不同 t 下所需的运行时间和发生的事件数目来确定 t 的值; 而 α 与事件 Remain、Birth、Death、Merge 和 Split 有关, 但是由于这些事件在不同 α 值下运行时间相差不大, 因此 α 值主要通过不同事件发生的数目进行确定. 同理, β 值也将

通过事件 Contract 和 Expand 发生的数目确定.

图 4 展示了 3 类事件在不同的参数 t 下发生的数目和运行时间. 从图 4(a)中不难看出发现这 3 类事件所需的运行时间与 t 成正比, 随着 t 值的增加, 耗时也随之变长. 图 4(b)中展示了事件 Remain 发生的次数随着 t 值的增长而增加, 但当 t 大于 3 之后发生的事件数目基本保持一致; 而事件 Birth 和 Death 发生的数目随着 t 的增长而减少, 但 t 取值大于 3 时事件发生次数也变化不大. 因此在本文中选取 t 值为 3.

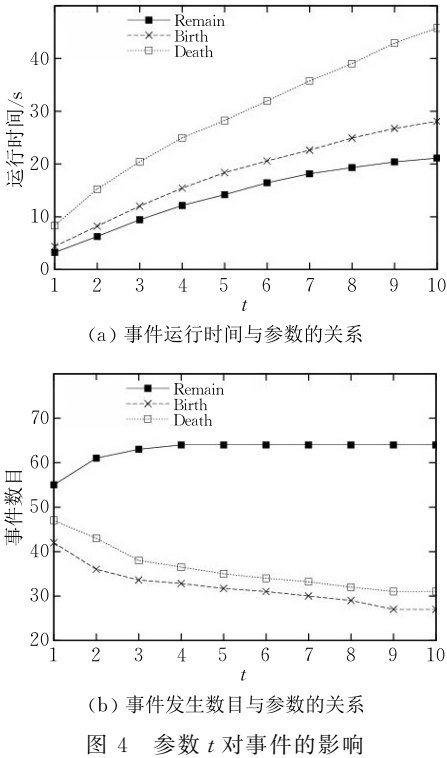
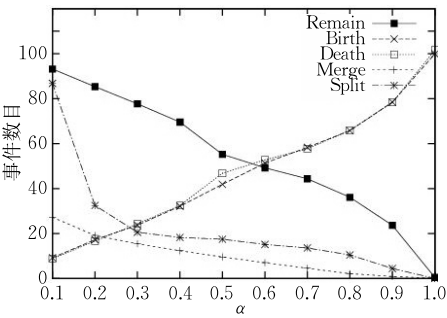
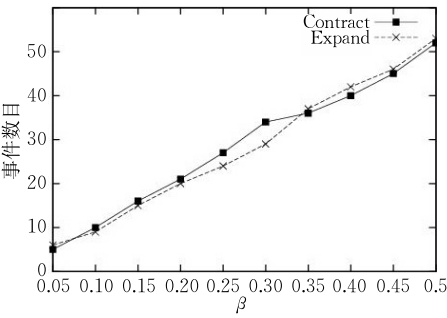


图 5 描述不同的参数 α 和 β 对事件发生数目的影响, 由于不同的参数对事件的运行时间影响不大, 因此实验主要涉及事件的发生数目. 不难发现图中事件 Birth 和 Death 发生次数是随着 α 的增加而增加, 而其他 3 个事件变化规律正好相反. 当 α 增加时, 对于社区而言它们更容易满足事件 Birth 和 Death 发生的条件, 因此这两个事件发生的次数增加; 同时另外 3 个事件发生的条件变得更加严格, 因此其他 3 个事件产生的数目减少. 从图 5(a)中不难发现当 α 等于 0.5 时, 各类事件发生的次数能达到一个平衡, 因此在本文中 α 取值为 0.5. 图 5(b)表明了事件 Contract 和 Expand 发生的次数与参数 β 成正比, β 值越大, 社区发生的这两类事件数目越多. 但是考虑到当该参数取值过大时, 这两个事件会与其他事件发生重叠, 因此在本文中该参数取值为 0.3.



(a) 参数 α 对事件的影响



(b) 参数 β 对事件的影响

图 5 参数 α 和 β 对事件发生数目的影响

4.3 事件数目比较

4.3.1 人造动态网络数据集

人造动态网络数据集是利用了 Greene 等人^[19]

提供的动态网络生成工具产生的,产生该网络
的各类参数见表 1. 本小节使用的是其中的 Dataset2,
该数据集包含 10 个时刻的网络,每个网络都包括
10 000 个节点.

表 4 和表 5 分别展示了 Asur 事件框架、Takaffoli
事件框架和本文提出的框架在人造动态网络不同
时刻检测到的各类事件发生次数. 其中由于社区在
第一个时刻都是属于刚刚出现,发生事件 Birth 的
数目等于所有社区的总数,这对于每一个事件框架
而言都是一样的,因此在 $T=1$ 时不统计社区发生
事件 Birth 的数目;同理其他事件则不计算 $T=10$ 时
刻发生的次数. 从表 3 可以看出 Asur 事件框架检
测到的事件发生次数远远小于其他两类事件框架,
这是由于该事件框架对于事件的定义较为严格,导
致满足事件发生条件的社区很少,另外该事件只计
算相邻时刻社区变化,更是使得社区发生的事件
数目相对较少. 表 4 展示的是 Takaffoli 事件框架
检测到的社区发生各类事件的次数,不难发现该
事件框架检测到的事件 Birth 和 Death 数目小于
本文事件框架(见表 5)所检测到的,这是由于
Takaffoli 事件框架检测这两个事件时统计了所有
时刻的网络,从而导致满足条件的社区较少;而
对于另外 3 个事件

表 3 Asur 事件框架在人造数据集上检测到的各类事件发生次数

事件	时间									
	$T=1$	$T=2$	$T=3$	$T=4$	$T=5$	$T=6$	$T=7$	$T=8$	$T=9$	$T=10$
Birth	—	44	54	50	54	56	54	53	57	56
Death	44	49	48	54	56	55	54	57	56	—
Remain	12	10	9	10	7	5	9	10	8	—
Merge	2	4	1	5	3	2	1	2	2	—
Split	2	3	2	4	1	3	3	5	2	—

表 4 Takaffoli 事件框架在人造数据集上检测到的各类事件发生次数

事件	时间									
	$T=1$	$T=2$	$T=3$	$T=4$	$T=5$	$T=6$	$T=7$	$T=8$	$T=9$	$T=10$
Birth	—	54	59	64	66	65	63	65	65	66
Death	54	59	58	64	66	67	65	66	67	—
Remain	130	120	100	101	99	102	112	113	110	—
Merge	33	35	32	26	26	24	25	24	24	—
Split	32	33	32	26	30	32	31	27	25	—

表 5 本文事件框架在人造数据集上检测到的各类事件发生次数

事件	时间									
	$T=1$	$T=2$	$T=3$	$T=4$	$T=5$	$T=6$	$T=7$	$T=8$	$T=9$	$T=10$
Birth	—	74	69	70	74	76	75	82	87	86
Death	74	69	70	74	76	75	82	87	86	—
Remain	120	110	92	87	87	75	79	65	81	—
Merge	22	24	16	11	13	9	11	9	12	—
Split	30	30	28	26	26	23	22	24	23	—
Contract	80	70	77	78	69	66	77	76	71	—
Expand	64	61	71	75	78	64	69	79	72	—

该框架检测到的数目比本文事件框架所检测到的多,原因同样是因为该框架计算了所有时刻的网络,但是实际上更合理的方式是只统计相邻时刻网络,而且统计所有时刻网络对于大规模节点的网络而言消耗的时间是成倍增长的.

4.3.2 DBLP 动态网络数据集

本小节使用 3 类事件框架分别处理实际的动态网络数据,即 DBLP 数据集. 该数据集包含五年的数据,每一年的数据被认为是一个网络,其中每年的网络中节点数和边数见表 2.

表 6 展示了 Asur 事件框架检测到的各类事件. 从中不难发现事件 Birth 和 Death 发生的次数远多于其他事件. 这表明整个网络稳定性很差,网络中有大量社区消失同时又有很多新社区形成. 同时可以看到检测到的事件 Remain、Merge 和 Split 都接近于 0,这一方面是因为 Asur 事件框架对于事件的定义过于严格,另一方面则是数据集本身的原因. 该数据集每个时刻网络内包含的节点 1000 个左右,然而检测到的社区有接近 300 个,这就导致了网络中存在大量的小社区,而小社区往往容易发生消失和生成等事件. 表 7 展示的是 Takaffoli 事件框架检测到的事件数目,与在人造网络数据集上的规律基本一致,该事件框架检测到的事件 Birth 和 Death 比本文检测(见表 8)所发现的事件数目少,而检测到的另外 3 个事件则更多,但是该框架是以时间为代价实现的这个目标,而且只通过分析相邻时刻网络检测事件 Merge 和 Split 发生次数更符合实际情况.

表 6 Asur 事件框架在 DBLP 动态网络数据集上检测到的各类事件发生次数

事件	年份				
	2005	2006	2007	2008	2009
Birth	—	281	318	334	325
Death	241	286	331	338	—
Remain	4	2	3	2	—
Merge	0	0	0	0	—
Split	0	0	0	0	—

表 7 Takaffoli 事件框架在 DBLP 动态网络数据集上检测到的各类事件发生次数

事件	年份				
	2005	2006	2007	2008	2009
Birth	—	288	328	337	335
Death	251	292	331	338	—
Remain	20	19	20	21	—
Merge	4	0	2	1	—
Split	4	2	0	2	—

表 8 本文事件框架在 DBLP 动态网络数据集上检测到的各类事件发生次数

事件	年份				
	2005	2006	2007	2008	2009
Birth	—	305	335	346	342
Death	264	314	338	343	—
Remain	18	17	16	19	—
Merge	2	0	1	0	—
Split	2	1	0	1	—
Contract	12	9	9	7	—
Expand	10	9	7	13	—

4.3.3 Facebook 数据集

本文采用的 Facebook 数据集是从 Viswanath 等人^[27]收集的数据集中选取 2007 年 1 月至 10 月数据所构成的,每一月的数据组成一个网络,因此总共包含 10 个网络数据.

表 9 展示了 Asur 事件框架检测到的各类事件. 同样也是事件 Birth 和 Death 发生的次数多于其他事件. 其他事件产生的次数少主要是因为 Asur 框架对于这些事件定义过于严格,所以导致满足条件的社区过少. 表 10 和表 11 分别描述了 Takaffoli 框架和本文框架检测到的事件数目,该事件框架检

表 9 Asur 事件框架在 Facebook 数据集上检测到的各类事件发生次数

事件	年份				
	2005	2006	2007	2008	2009
Birth	—	321	327	344	336
Death	321	314	345	348	—
Remain	40	42	43	42	—
Merge	22	20	23	25	—
Split	21	19	22	23	—

表 10 Takaffoli 事件框架在 Facebook 数据集上检测到的各类事件发生次数

事件	年份				
	2005	2006	2007	2008	2009
Birth	—	338	345	357	355
Death	331	372	352	354	—
Remain	320	319	321	324	—
Merge	184	182	183	181	—
Split	183	182	181	182	—

表 11 本文事件框架在 Facebook 数据集上检测到的各类事件发生次数

事件	年份				
	2005	2006	2007	2008	2009
Birth	—	356	355	368	362
Death	384	414	403	394	—
Remain	308	291	301	309	—
Merge	172	170	174	175	—
Split	171	171	170	172	—
Contract	264	258	260	251	—
Expand	239	235	248	257	—

测到的事件 Birth 和 Death 比本文所发现的事件数目少,而其他 3 类事件则正好相反,但是该框架在处理数据时所需时间远大于其他两类框架.

4.4 预测精度分析

本文根据 5 个与社区相关的指标和对应事件之间的映射关系构建出相应的预测模型. 首先使用历史数据,分析出各指标集与事件之间的映射关系,利用 Weka 十折交叉验证的方法选择多种分类器,然后在给出各指标的情况下,通过预测模型分析出各社区可能发生的事件,从而了解网络的变化过程.

首先使用 matlab 计算出网络中各个社区的指标以及发生的事件;然后将 5 个指标的集合 $\{Apl, CC, Rne, Pop, Soc\}$ 作为样本特征,而事件作为样本标签,取值为 1 时表示发生该事件,反之不发生;最

后选出某一时刻网络数据作为训练数据构建预测模型,并将其他时刻网络数据作为测试数据,最后利用预测模型对测试数据分别进行预测,将最后预测准确度取平均值则得到最终的预测精度. 考虑到各类事件相互之间存在联系,因此将分别为每类事件构建一个预测模型. 另外由于社区过小研究意义不大,因此限定社区内包含节点数需大于 3. 下面将介绍基于 3 类事件框架构建的预测模型在人造数据集、DBLP 数据集和 Facebook 数据集上的表现,其中表格中显示的准确度是预测模型在处理不同时刻网络得到的准确度均值.

4.4.1 人造动态网络数据集

人造数据集上各类事件的预测准确度如图 6 所示.

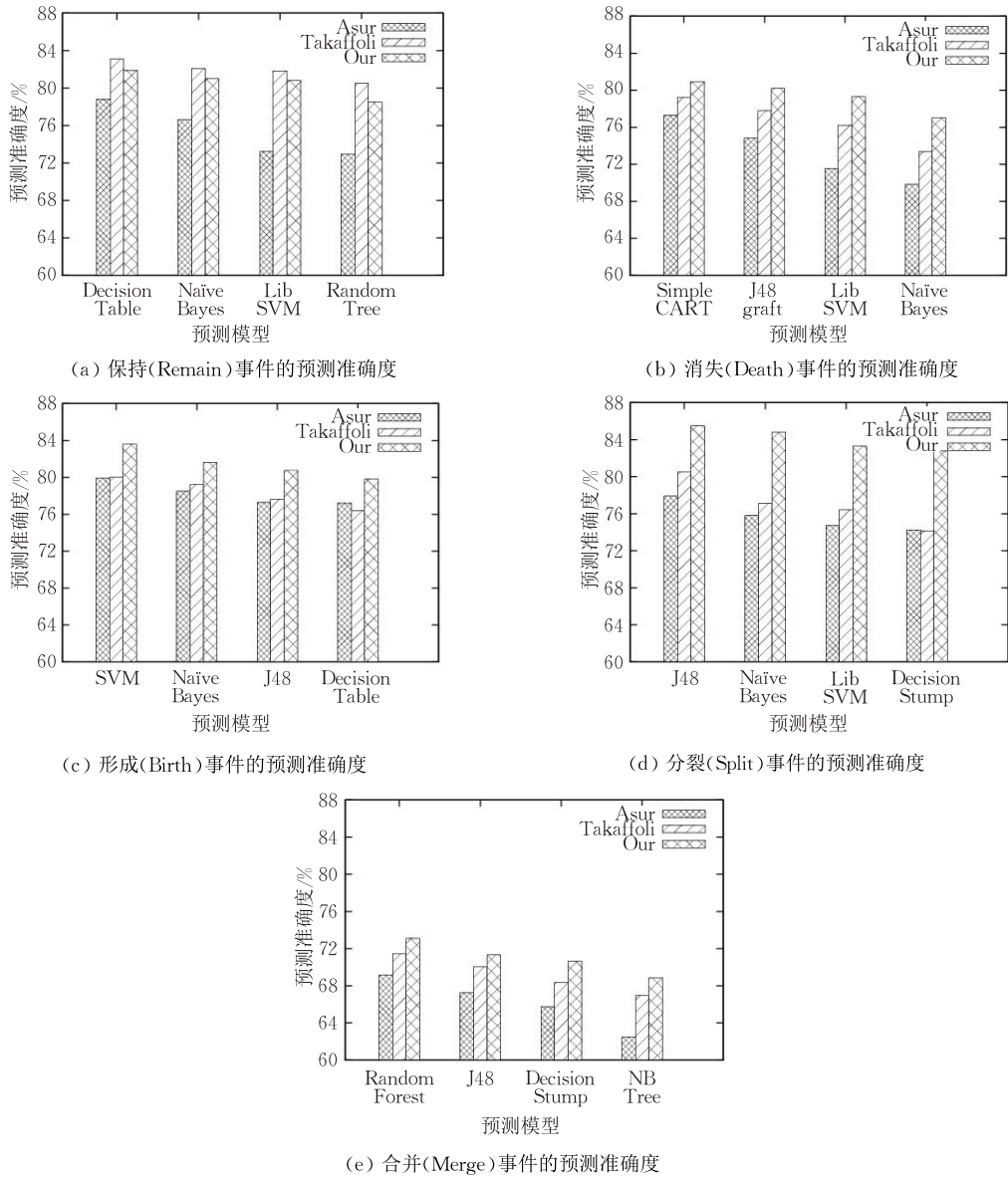
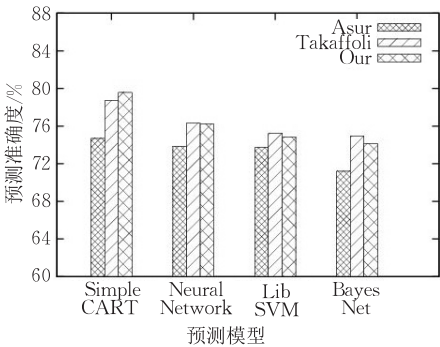


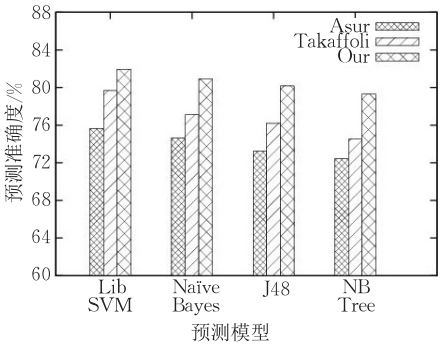
图 6 人造数据集上各类事件的预测准确度

4.4.2 DBLP 动态网络数据集

图 7 展示了基于各类事件框架构建的预测模型在 DBLP 数据集上的表现. 由于实验的 3 种事件框架在处理 DBLP 数据集时检测到的事件 Merge、Split 和 Remain 均太少, 所以对于这 3 个事件进行预测不存在实际意义, 因此在表中未给出这 3 个事件预测的准确度. 与在人造网络数据集上的表现基本一致, 基于 Asur 框架构建的预测模型准确度最低, 而基于本文框架构建的预测模型在处理其 Birth 和 Death 事件时准确度都比 Takaffoli 框架高. 因此, 本文框架更适合用于研究动态网络社区变化的过程.



(a) 消失(Death)事件的预测准确度



(b) 形成(Birth)事件的预测准确度

图 7 DBLP 数据集上各类事件的预测准确度

4.4.3 Facebook 数据集

图 8 表示基于各类事件框架构建的预测模型在 Facebook 数据集上的准确度. 从表中不难发现基于 Asur 框架构建的预测模型精度同样是最低的. 而基于本文框架构建的预测模型只是在事件 Remain 上精度略低于 Takaffoli 框架, 然而在其他事件上的预测精度均高于其他事件框架. 这也验证了本文框架较之其他两类框架表现得更为优异, 更适用于进行动态网络分析.

4.4.4 整体分析

根据人造动态网络数据集、DBLP 动态网络数据集和 Facebook 动态网络数据集的实验结果, 相比

较 Asur 等人提出的框架, 我们改进后的框架在各类事件的预测准确度上都有显著的提高. 和 Takaffoli 等人提出的框架进行比较, 我们发现只有在预测保持事件(Remain)时, Takaffoli 等人提出的框架略优于本文框架. 但是考虑到使用 Takaffoli 框架检测保持事件时会花费更多的时间去检测到更多的事件, 尤其是当动态网络数据集的时间跨度比较长的时候, 所以在处理数据量较大、时间跨度较长的数据时, 本文所定义的框架能够有效准确地预测各类事件. 根据实验结果, 可以看出对于各类不同的事件, 多个预测模型表现出不同的预测准确度, 但是每类事件预测准确度最好的模型却是可以统一的. 所以在进行消失事件进行预测时, 我们应该使用 Simple CART 模型, 预测精度超过 81%; 预测形成事件时, 我们使用 Lib SVM 模型, 可以达到很好的预测效果. 类似的, 通过实验我们发现在对分裂事件和合并事件进行预测时, 我们优先选用 J48 模型和 Random Forest 模型. 在预测保持事件时, 我们发现有人造动态网络数据集中表现最好的是 Decision Table, 但是我们发现 Facebook 数据集在进行保持事件预测最准确的是 Naïve Bayes, 且 Naïve Bayes 模型在有人造动态网络数据集上的预测也具有非常高的准确度, 因此在对保持事件进行预测时, 我们统一使用 Naïve Bayes 模型.

5 结 论

本文主要研究了动态网络的进化特征, 根据网络的社区结构我们提出了更加准确的社区事件框架定义方法. 利用新提出的事件定义方法对包括有人造动态网络数据集、DBLP 动态网络数据集和 Facebook 数据集在内的多种数据集进行社区事件检测实验, 实验表明我们的事件框架比 Asur 事件框架和 Takaffoli 事件框架能检测出更多的事件, 特别是和 Asur 框架进行比较时优势明显. 对于不同的事件, 我们构建了相应的预测模型, 每个模型都选取 5 个指标和对应的社区事件作为输入输出. 指标包括社区平均路径长度、社区聚类系数、社区节点边对比度、社区流行度和社区社交度, 特别是社区流行度和社区社交度这两个指标的提出极大地提高了预测模型的准确性. 例如对于 Split 事件我们就选择 J48 进行事件预测, Death 事件使用 Simple CART 预测模型. 在实验中以有人造动态网络数据集、DBLP 动态网络数据集和 Facebook 数据集作为实验数据集,

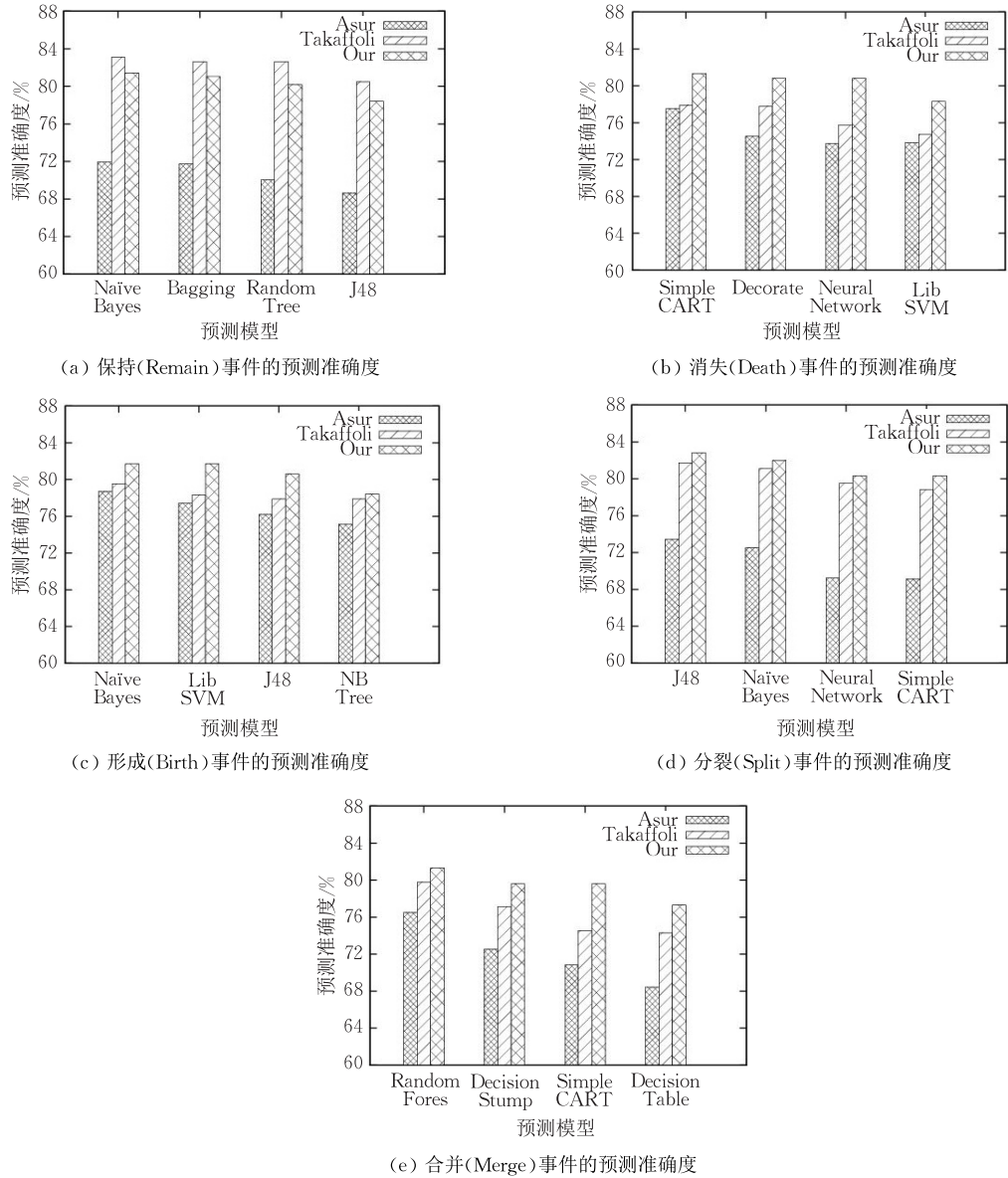


图 8 Facebook 数据集上各类事件的预测准确度

J48、Naive Bayes、Random Forest 和 Lib SVM 等作为模型输入,发现对每个事件的预测准确度都可以达到或超过 80%,这说明用我们提出的社区指标和对应事件的事件模型来预测未来该事件是否会发生是可行的.在实际网络中社区可能发生的社区事件不仅仅只有 Remain、Death、Birth、Merge、Split、Contract 和 Expand,如何在利用性能更好的重叠社区发现算法(如 Nepusz 和 Yu 等人^[28]提出的 ClusterONE 算法)的基础上,以忠实揭示复杂网络中社区的演化过程和变化规律为目标,拓展和优化本文所提出的事件框架及其指标体系,构建动态自适应的社区事件以及节点事件的预测模型,将是未来研究的重点.

参 考 文 献

[1] Girvan M, Newman M E J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 2002, 99(12): 7821-7826

[2] Newman M E. The structure and function of complex networks. *SIAM Review*, 2003, 45(2): 167-256

[3] Fortunato S. Community detection in graphs. *Physics Reports*, 2010, 486(3): 75-174

[4] Xie J, Kelley S, Szymanski B K. Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM Computing Surveys*, 2013, 45(4): 115-123

[5] Lancichinetti A, Fortunato S. Community detection algorithms: A comparative analysis. *Physical Review E Statistical Nonlinear*

8. Soft Matter Physics, 2009, 80(5): 2142-2152

[6] Clauset A, Newman M E J, Moore C. Finding community structure in very large networks. Physical Review E, Statistical, Nonlinear, and Soft Matter Physics, 2004, 70(6 Pt 2): 264-277

[7] Palla G, Derenyi I, Farkas I, et al. Uncovering the overlapping community structure of complex networks in nature and society. Nature, 2005, 435(7043): 814-818

[8] Zhang X, You H, Zhu W, et al. Overlapping community identification approach in online social networks. Physica A: Statistical Mechanics and Its Applications, 2015, 421: 233-248

[9] Backstrom L, Huttenlocher D, Kleinberg J, et al. Group formation in large social networks: Membership, growth, and evolution//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA, 2006: 44-54

[10] Falkowski T, Bartelheimer J, Spiliopoulou M. Mining and visualizing the evolution of subgroups in social networks//Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence. Hong Kong, China, 2006: 52-58

[11] Sun J, Faloutsos C, Papadimitriou S, et al. GraphScope: Parameter-free mining of large time-evolving graphs//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, USA, 2007: 687-696

[12] Lin Y R, Chi Y, Zhu S, et al. Analyzing communities and their evolutions in dynamic social networks. ACM Transactions on Knowledge Discovery from Data, 2009, 3(2): 8

[13] Palla G, Barabási A L, Vicsek T. Quantifying social group evolution. Nature, 2007, 446(7136): 664-667

[14] Asur S, Parthasarathy S, Ucar D. An event-based framework for characterizing the evolutionary behavior of interaction graphs. ACM Transactions on Knowledge Discovery from Data, 2009, 3(4): 913-921

[15] Tantipathananandh C, Berger-Wolf T, Kempe D. A framework for community identification in dynamic social networks//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, USA, 2007: 717-726

[16] Mansoureh T, Farzad S, Justin F, Osmar R Z. A framework for analyzing dynamic social networks//Proceedings of the 7th Conference on Applications of Social Network Analysis (ASNA). Zurich, Switzerland, 2010: 1-14

[17] Takaffoli M, Sangi F, Fagnan J, et al. Community evolution mining in dynamic social networks. Procedia-Social and Behavioral Sciences, 2011, 22(22): 49-58

[18] Takaffoli M, Rabbany R, Zaiane O R. Community evolution prediction in dynamic social networks//Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Istanbul, Turkey, 2014: 9-16

[19] Greene D, Doyle D, Cunningham P. Tracking the evolution of communities in dynamic social networks//Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining. Odense, Denmark, 2010: 176-183

[20] Bródka P, Saganowski S, Kazienko P. GED: The method for group evolution discovery in social networks. Social Network Analysis and Mining, 2013, 3(1): 1-14

[21] Saganowski S, Gliwa B, Bródka P, et al. Predicting Community Evolution in Social Networks//Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Paris, France, 2015: 3053-3096

[22] Gregory S. An algorithm to find overlapping community structure in networks//Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases. Warsaw, Poland, 2007: 91-102

[23] Zhang S, Wang R S, Zhang X S. Identification of overlapping community structure in complex networks using fuzzy c-means clustering. Physica A: Statistical Mechanics and its Applications, 2007, 374(1): 483-490

[24] Bródka P, Kazienko P, Kołoszyk B. Predicting group evolution in the social network. Lecture Notes in Computer Science, 2012, 7710: 54-67

[25] Hall M, Frank E, Holmes G, et al. The WEKA data mining software: An update. ACM SIGKDD Explorations Newsletter, 2010, 11(1): 10-18

[26] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection//Proceedings of the 14th International Joint Conference on Artificial Intelligence. Adelaide, Australia, 2001: 1137-1143

[27] Viswanath B, Mislove A, Cha M, et al. On the evolution of user interactions in facebook//Proceedings of the 2009 ACM Workshop on Online Social Networks. Nuremberg, Germany, 2009: 37-42

[28] Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks. Nature Methods, 2012, 9(5): 471-472



ZHANG Xue-Wu, born in 1973, Ph.D. , professor. His current research interests include intelligent sensing and calculating, big data analysis and visualization.

SHEN Hao-Dong, born in 1991, M. S. candidate. His current research interests include social network analysis and community mining.

ZHAO Pei-Ran, born in 1994, M. S. candidate. His current research interests focus on community mining.

ZHANG Zhuo, born in 1978, Ph.D. Her current research interests include machine learning, social network theory and related methods for big data analysis.

LI Min, born in 1982, Ph.D. Her current research interests include machine learning and data mining.

XU Hai-Yan, born in 1978, Ph.D. Her current research interests include information processing and artificial intelligence.

Background

Complex networks rose at the end of the twentieth century provide a new perspective for studying complex systems. It includes discovering the statistical properties of network structure, detecting the communities in networks and studying the dynamic processes of complex networks to help people improve the performance of existing networks. Social networks are a medium for disseminating information, such as news, views and advertisements, etc. To study social networks based on social network theory has the potential value of theory and application. Traditional research on social networks is limited to static networks, while more meaningful research fields are dynamic social networks. Currently, more and

more studies focus on the evolution of communities in social networks. We mainly research evolution based on improved event framework, present several features involved social network structure, community structure and community evolution features and take the features as the input to find appropriate predictive models.

This research is financially supported by the National Natural Science Foundation of China under Grant Nos. 61671202, 61573128, and 61273170, the National Basic Research Program (973 Program) of China under Grant No. 2016YFC0401606, and the Fundamental Research Funds for the Central Universities under Grant No. 2015B25214.