软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2019,30(3):537−551 [doi: 10.13328/j.cnki.jos.005684]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于角色发现的动态信息网络结构演化分析*

冯冰清[1,2], 胡绍林[1], 郭 栋[1], 钟晓歌[1], 李佩钰[1]

[1](航天器在轨故障诊断与维修重点实验室,陕西 西安 710043)

[2](四川大学 计算机学院,四川 成都 610065)

通讯作者: 郭栋, E-mail: 747818416@qq.com

摘 要: 动态信息网络是当前复杂网络领域中极具挑战的新问题之一,对其动态的演化过程进行研究,有助于分析网络结构、理解网络特性、发现网络中潜在的信息及演化规律,具有重要的理论意义与应用价值.基于网络结构本身量化表示的复杂性以及网络演化时序、复杂、多变的挑战,使用角色来量化动态网络的结构,并对模型进行分析,给出了两种角色解释的方法;在角色发现的基础上,将动态网络结构预测问题转换为可以表示结构特征的角色预测问题,通过向量自回归的方法,以历史网络角色分布矩阵作为训练数据构建模型,预测未来时刻网络可能的角色分布情况,提出了基于潜在角色的动态网络结构预测方法 LR-DNSP(latent role based dynamic network structure prediction).该方法克服了已有基于转移矩阵方法忽略历史信息的不足,并且考虑了多个预测目标之间可能存在的相互关系.实验结果表明,提出的 LR-DNSP 方法具有更准确的预测效果.

关键词: 动态信息网络;角色发现;结构演化;结构预测

中图法分类号: TP311

中文引用格式: 冯冰清,胡绍林,郭栋,钟晓歌,李佩钰.基于角色发现的动态信息网络结构演化分析.软件学报,2019,30(3):537−551. http://www.jos.org.cn/1000-9825/5684.htm

英文引用格式: Feng BQ, Hu SL, Guo D, Zhong XG, Li PY. Structure evolution analysis based on role discovery in dynamic information networks. Ruan Jian Xue Bao/Journal of Software, 2019,30(3):537−551 (in Chinese). http://www.jos.org.cn/1000-9825/5684.htm

Structure Evolution Analysis Based on Role Discovery in Dynamic Information Networks

FENG Bing-Qing[1,2], HU Shao-Lin[1], GUO Dong[1], ZHONG Xiao-Ge[1], LI Pei-Yu[1]

[1](Key Laboratory for Fault Diagnosis & Maintenance of Spacecraft in Orbit, Xi'an 710043, China)

[2](College of Computer Science, Sichuan University, Chengdu 610065, China)

Abstract: Dynamic information network is a new challenging problem in the field of current complex networks. Research on network evolution contributes to analyzing the network structure, understanding the characteristics of the network, and finding hidden network evolution rules, which has important theoretical significance and application value. The study of the network structure evolution is of great importance in getting a comprehensive understanding of the behavior trend of complex systems. However, the network structure is difficult to represent and quantify. And the evolution of dynamic networks is temporal, complex, and changeable, which increases the difficulty in analysis. This study introduces "role" to quantify the structure of dynamic networks and proposes a role-based model, which provides a new idea for the evolution analysis and prediction of network structure. As for the model, two methods to explain the role are given. To predict the role distributions of dynamic network nodes in future time, this study transforms the problem of dynamic network structure prediction into role prediction, which can represent the structural feature. The model extracts properties from historical snapshots of sub-network as the training data and predicts the future role's distributions of dynamic network by using the vector autoregressive

* 基金项目: 国家自然科学基金(61473222, 91646108)
Foundation item: National Natural Science Foundation of China (61473222, 91646108)
本文由智能数据管理与分析技术专刊特约编辑樊文飞教授、王国仁教授、王朝坤副教授推荐.
收稿时间: 2018-07-17; 修改时间: 2018-09-20; 采用时间: 2018-11-01

method. This study also proposes the method of dynamic network structure prediction based on latent roles (LR-DNSP). This method not only overcomes the drawback of existing methods based on transfer matrix while ignoring the time factor, but also takes into account of possible dependencies between multiple forecast targets. Experimental results show that the LR-DNSP outperforms existing methods in prediction accuracy.

Key words: dynamic information network; role discovery; structural evolution; structural prediction

随着信息技术的迅猛发展,网络数据的种类和数量巨幅增长,从社交网络到科研合作者网络,从电力网络到城市交通网络,从生物体中的大脑到各种新陈代谢网络,人们已经生活在一个充满着各种各样的复杂网络世界中[1].对复杂网络的研究通常需要对其进行建模和简化,传统复杂网络研究多将复杂系统建模为静态网络,而现实中几乎所有的复杂系统都是随时间不断变化的.以社交网站 Facebook[2]为例,从 2004 年上线到现在,网站每月的活跃用户数超过 20 亿,已经发展成为全球最大的社交网站之一,类似的还有 google+[3],Twitter[4].事实上,不仅仅是社交网站,还有科学家合作网络、城市交通网络、公司邮件网络、通信网络等,这些复杂网络的显著特征是网络的结构随时间不断地变化,而这些时序动态特征对理解系统或网络中的行为至关重要.这些不断变化的网络就是动态复杂网络,简称动态网络[5].对动态网络时序模式的深入理解,有助于分析网络结构、理解网络特性、发现网络中潜在的信息及演化规律,因此,对动态网络建模及分析得到了广泛关注.

信息网络(information network)[6]是对现实空间中海量、多维、复杂结构和问题更具一般性的抽象[7],可以有效抽象出复杂系统中有价值的特征与潜在规律,从而为系统化地分析现实中的复杂网络提供高效的研究和探索的方法.信息网络一般都有动态的演化过程,新节点会持续地加入网络,一些节点也会在中途消失,节点间连接的强度也在不断变化,信息网络中网络结构随之处于不断演化的过程中[8],这里统称为动态信息网络.动态信息网络的演化过程具有时序、复杂、多变的特点,蕴含着丰富的潜在信息和商业价值.动态网络的结构预测是网络演化中十分重要的问题,它旨在利用历史网络信息预测未来时刻节点的拓扑结构,帮助人们提前进行预警和决策.

动态信息网络是当前复杂网络研究领域中极具挑战的新方向,由于网络结构本身比较复杂,难以表示和量化,动态网络时序、多变的演化过程更增加了分析的难度.基于动态信息网络的广泛应用前景及角色(role)发现在动态网络中有限的研究现状,本文致力于研究动态信息网络中基于角色发现的结构预测问题,主要包括针对网络结构表示的复杂性以及网络演化时序多变的挑战,将静态网络中用角色来量化网络结构的方法扩展至动态网络,以角色发现为基础,对动态网络结构预测进行探索性研究.主要贡献如下.

(1) 提出动态网络角色发现模型

使用角色来表示动态网络的结构,将静态网络中基于递归地提取特征的角色发现方法扩展至动态网络,按照时间序列对每个网络快照进行特征提取,然后为每个快照学习节点的行为角色,提出动态网络的角色发现模型.同时给出两种角色解释的方法,并在不同规模的真实网络数据集上进行实验,验证本文模型的有效性和可解释性.

(2) 提出基于潜在角色的动态网络结构预测方法 LR-DNSP

网络结构的动态预测是动态网络演化分析的一个重要任务.将动态网络结构预测转换为可以表示结构特征的角色预测问题,以历史网络角色分布矩阵作为训练数据构建模型,通过向量自回归的方法预测未来时刻网络可能的角色分布情况,提出了基于潜在角色的动态网络结构预测方法 LR-DNSP(latent role based dynamic network structure prediction).该方法克服了已有基于转移矩阵方法未能充分利用历史信息的不足,并且考虑了多个预测目标之间可能存在的相互关系.实验结果表明,本文提出的 LR-DNSP 方法具有更准确的预测效果.

1 相关工作

在动态网络的相关研究工作中,节点中心性分析、节点影响力分析、链接预测、异常发现以及社团发现、社团演化等近年来得到了较多的关注,而相对而言对节点结构行为分析关注较少.相比之下,本文更关注如何发现网络中节点的行为模式,通过网络随时间的变化来捕获节点行为的模型,并建模这些模式随时间的变化.针对

网络结构的表示问题,Henderson 等人[9]在 KDD2012 上首次提出用潜在的角色来刻画节点的结构行为.角色代表网络结构的某种类型,结构类型相似的节点属于同一种角色,如中心节点、桥梁节点、边缘节点等.角色发现是静态网络结构的一种有效量化方法,可以将复杂的网络结构表示为相对简单的角色.节点角色分析试图将在网络中有着相同地位或发挥相同角色作用或有着相同功能的节点归为一类.它与社团发现有着本质的区别:社团发现是根据节点连接的紧密程度进行聚类,而角色发现主要依赖于网络中节点的拓扑结构特征.

Henderson 等人在角色概念的基础上提出了基于特征的角色发现方法 RolX(role extraction),通过无监督学习,从网络中自动提取结构角色,进一步实现网络挖掘任务[9].在已有的无监督角色发现的研究基础上,Sean Gilpin 等人[10]提出了基于交替最小二乘的有监督角色发现方法,主要解决了数据集稀疏性、多样性和角色交替性问题等.Rossi 等人以角色为基础进行了一系列动态网络演化分析相关的研究[11−13],也是本文研究工作的基础.文献[11]提出一种基于自学习方法挖掘动态网络角色的混合模型,该模型以基于特征的角色发现方法为基础,将动态网络视为多个静态网络的序列,在离散的时间点上进行角色发现,比较不同时刻的角色分布情况来分析网络角色的动态变化趋势;文献[13]进一步将上述模型进行扩展,应用模型进行未来时刻网络角色的预测,其思想是将动态网络的多种角色视为网络的多个状态,通过计算网络在相邻时刻的状态转移矩阵来进行角色演化的分析与预测,但是由于该转移矩阵模型只通过相邻两个时刻的网络数据得到,未能充分利用历史时刻的网络数据,因此预测效果有待改进,该方法将作为本文角色预测问题的对比方法之一.

近年来,角色发现正在被其他领域广泛探索,如在线社交网络[14]、科技网络[15]、生物网络[16,17]、网络图[18]等.角色发现在网络挖掘的探索分析中,从传统的节点分类、异常检测、预测问题到结构相似性度量、图相似性研究、网络可视化、迁移学习等,逐步发挥着重要作用.McDowell 等人[19]使用角色作为特征进行分类;Rossi 等人[15]对给定的两个图,通过提取各自的特征和角色进行图相似研究;Henderson 等人[9]通过对给定网络的角色学习,使用已有的知识在另一网络学习同样的角色集合,以提高分类的准确性.角色发现也可以被推广到更多实际的应用中,例如,角色可用于检测 IP 网络中的异常,可以基于用户在网络中的角色来定制广告推送.在网络挖掘和实际应用中,角色正在成为一种重要的潜在分析视角.但相比于社团发现、社团演化等,角色发现仅受到了有限的关注.

2 动态网络的角色发现与角色解释

拓扑结构是复杂网络的研究基础,静态网络中已有的拓扑指标包括度、距离、直径、密度、聚集系数、介数中心性、参与三角形数量、模块性等,涉及网络不同层面的度量,为进一步分析动态网络的结构特征提供了理论依据,也是捕获角色的基础[20].静态网络中的角色发现是一种对网络结构的有效量化方法,本文旨在从节点角色的角度描述其在网络中的结构特征,图 1 中用不同颜色区分了网络中的不同角色.使用角色来量化节点的结构,可以有效化简网络结构分析的难度.

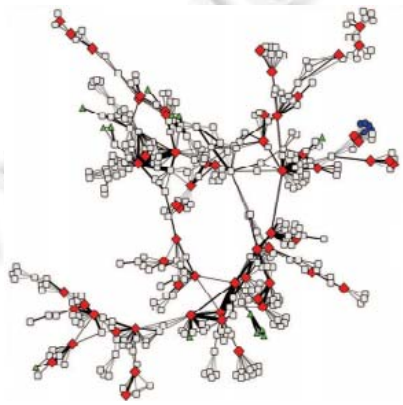


Fig.1 Role discovery[9]

图 1 角色发现[9]

网络中的角色目前还没有统一的定义,可以概括地认为角色是节点在网络中所表现出的结构行为,来刻画节点的某种重要程度[21].而角色发现则是通过量化节点结构,判定节点的结构角色.为解决动态网络结构演化分析的问题,本文首先对动态网络进行角色发现.

2.1 动态网络的角色发现

进行动态网络的角色发现之前,首先要构建动态网络.动态网络的定义如下.

定义 1(动态网络). $D=\langle N,E\rangle$表示一个动态网络,$N=\langle N_1,N_2,\ldots,N_T\rangle$为节点集合,$E=\langle E_1,E_2,\ldots,E_T\rangle$为边集合.将 D 看做一个时间有序的子图序列 $D=\langle S_1,S_2,\ldots,S_T\rangle$,其中,$S_T=\langle N_T,E_T\rangle$是动态网络 D 在 t 时刻的子图快照,N_t 为 S_t 的节点集合,E_t 为 S_t 的边集合,T 为动态网络长度.本文研究网络的结构演化,故只考虑无向网络.

研究动态网络的角色发现,首先要对网络进行特征提取,得到高维特征矩阵,然后对特征矩阵通过非负矩阵分解进行角色发现,在角色发现的过程中要确定分解的最优 r 值,最后对得到的角色模型进行解释.将动态网络表示为有序的子图序列后,对每个时刻的网络快照分别进行角色发现,即将每个子网络都转化为节点的角色信息.本文使用 KDD2012 的 RloX 方法[9]进行角色发现.相比其他传统方法,RolX 更适合大规模网络的角色发现,它不仅能发现网络中的角色,还可以得到节点在各角色上的概率取值.该方法通过以下两步过程完成.

(1) 特征提取

特征提取过程采用 ReFex[22]的迭代特征产生方法,为每个节点提取基本特征和递归特征,基本特征指节点局部结构的特征,即在与一阶邻居所形成的自网络中所表现出的特征,如节点的度、加权度、自网络包含的边数等.得到节点的基本特征后,使用聚集函数递归地对其邻居节点的基本特征进行聚集计算得到递归特征[22].以图 2 所示网络为例,虚线内表示 n_1 的自网络,选取 3 个基本特征:度、自网络包含的边数、参与三角形的个数,使用求和以及求平均两种聚集函数来产生递归特征.得到 n_1 的基本特征的向量为$\langle f_1,f_2,f_3\rangle=\langle 6,11,5\rangle$.接着计算递归特征,直到没有新特征产生终止,便可将节点 n_1 表示为一个特征取值向量 $f=\langle f_1,f_2,f_3,\ldots\rangle$.

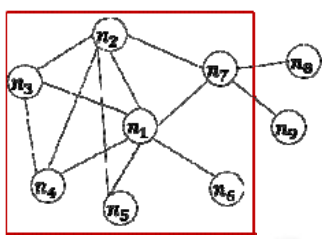


Fig.2 An example of role discovery

图 2 角色发现的示例网络

对每个节点分别进行特征提取,可以得到一个节点的特征取值向量.因此,可以将网络快照 S_t 转化为一个特征空间,记为节点的特征 $V_t\in\Re^{N\times f_t}$.

定义 2(节点-特征矩阵序列 $V=\langle V_1,V_2,\ldots,V_T\rangle$). 给定动态网络子图序列 $D=\langle S_1,S_2,\ldots,S_T\rangle$,对 S_t 进行特征提取,得到节点的特征矩阵 $V_t\in\Re^{N\times f_t}$,其中,N 为网络的节点个数,f_t 表示 t 时刻得到的特征个数.对每个网络快照 $D=\langle S_1,S_2,\ldots,S_T\rangle$分别进行特征提取,得到节点-特征矩阵序列 $V=\langle V_1,V_2,\ldots,V_T\rangle$.

(2) 角色发现

通过对节点的特征矩阵降维分解,进一步进行角色发现,降维后得到对节点特征的概括就是潜在的角色.基于非负矩阵分解实现上的简便性、分解形式和分解结果上的可解释性,本文使用非负矩阵分解方法对提取到的特征矩阵进行降维.对特征矩阵 $V_t\in\Re^{N\times f_t}$,给定一个正整数 $r<\min(N,f_t)$,NMF 可以寻找非负矩阵 $G_t\in\Re^{N\times r}$ 以及 $F\in\Re^{r\times f_t}$,满足 $G_tF\approx V_t$,通过求解以下问题的最优解:

$$\arg\min_{G_t,F}=\frac{1}{2}\|V_t-G_tF\|_F^2 \tag{1}$$

其中,$\|\cdot\|_F^2$ 表示矩阵 F 范数的平方,分解目标 G_t 便是节点的角色矩阵.

角色-特征矩阵 $F\in\mathbb{R}^{r\times f_t}$ 表示了每个角色在提取到的特征值上的贡献,学习到 F 之后,进一步对整个动态网络进行角色发现,对每个子图快照 $D=\langle S_1,S_2,\ldots,S_T\rangle$在特征提取后,根据得到节点的特征序列 $V=\langle V_1,V_2,\ldots,V_T\rangle$和 F,分别进行 NMF 过程得到全部节点的角色序列 $G=\langle G_1,G_2,\ldots,G_T\rangle$.

根据上面非负矩阵分解得到的结果,G_t是一个 N 行 r 列的矩阵,每列对应一种角色,G_t 的每个元素 $g_t(i,j)$表示节点 i 属于角色 j 的概率.进行角色发现涉及到需要确定角色的个数,本文使用最小描述长度(minimum description length)准则[23]选定角色个数 r,使得节点特征矩阵可以得到最佳的压缩.

算法 1. 角色个数选取算法.

输入:原始节点-特征矩阵 $V\in\mathbb{R}^{n\times f}$;

输出:角色个数:r.

1. *Mincost*=∞, *failed*=0, max=*m*;
2. **for** *r*←1 to min(*n*,*f*) **do** //*r* 为角色个数,即矩阵分解得到的 *G* 的列数、*F* 的行数
3. (G_r,F_r)←*NMF*(*V*); //*NMF*()为非负矩阵分解函数
4. Compute cost of model ℓ via MDL criterion;
5. **if** *cost*<*mincost* **then**
6. *mincost*=*cost*;
7. *failed*=0;
8. **else** *failed*=*failed*+1;
9. **if** *failed*≥max **then**
10. **break**;
11. **end for**
12. **return** *r*(角色个数)

用上述方法对动态网络进行角色建模,得到角色序列 $G=\langle G_1,G_2,\ldots,G_T\rangle$,$G_t$ 的每一行表示 t 时刻该节点在 r 个角色上的取值分布.但与此同时,也暴露出对得到的结果难以直观理解的问题,由矩阵分解方法得到的是特征空间中 r 个潜在的角色,虽然能得到网络中角色的个数,但这些角色并不直观,无法获知每种角色具体表示何种结构.

2.2 角色解释

为了对得到的角色有直观的认识,下面介绍如何对模型得到的角色进行解释.在得到角色序列 $G=\langle G_1,G_2,\ldots,G_T\rangle$后,节点结构被表示为在角色上的概率取值,可否利用传统的度量(如度、介数、离心率等)和邻接节点的角色分布对角色进行量化和解释?为了得到对角色的感官认识,本文给出两种角色解释方法:一种是基于节点自身度量属性的方法 NodeSense,另一种是基于邻居节点分布的方法 NeighborSense.

基于以上思路,对给定动态网络的子图快照 S_t 与角色矩阵 G_t,为每个节点计算一系列度量属性,本文选取了 8 种传统的度量属性:度、加权度、介数、特征向量中心度、紧密中心度、聚集系数、PageRank、离心率,计算可得到 t 时刻节点的度量矩阵 $M_t\in\mathbb{R}^{N\times m}$,其中,$m$=8.为得到角色与度量之间的量化关系,NodeSense 计算一个新的非负矩阵 P_t,使得 $G_tP_t\approx M_t$,其中,$P_t\in\mathbb{R}^{r\times m}$.$P_t$ 的行对应 r 个角色,列对应 m 个度量,P_t 的每一行则表示该角色在各个度量上的概率取值,即角色在各个度量属性的贡献.

NeighborSense 方法的思路类似于 NodeSense:首先,对 t 时刻子图快照 S_t 计算每个节点邻居的角色分布矩阵 $N_t\in\mathbb{R}^{n\times r}$,$N_t$ 的行表示节点,列表示角色,$N_t(i,j)$表示 t 时刻节点 i 的所有邻居节点在角色 j 上的分布统计;接着,NeighborSense 计算一个非负矩阵 Q_t,使得 $G_tQ_t\approx N_t$,其中,$Q_t\in\mathbb{R}^{r\times r}$ 表示角色与角色之间的关系.比如,有些节点更倾向与自己相同角色的节点相连,则可以认为这类角色是同质性的;而另一部分节点可能与自己相异的角色相连,这类角色可认为是异质性的.具体分析结果将在实验部分呈现.

3 动态网络角色预测

动态网络的结构预测是网络演化分析的重要问题,但由于动态网络演化过程本身复杂多变,加上网络规模的急剧增长,该问题还未得到很好的解决.网络结构本身难以量化,直接预测网络结构比较困难.本文的动态网络角色模型为网络结构预测问题提供了一个新的思路,即用角色来建模动态网络结构,动态网络结构表示为节点的角色分布序列,将网络结构预测问题转化为角色分布的预测.这样,很大程度上降低了问题的求解难度(如图 3 所示).

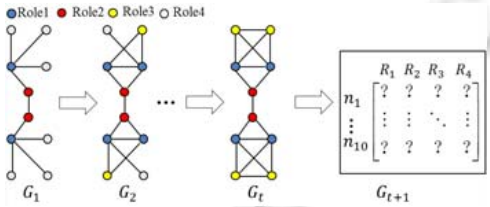


Fig.3 Role prediction of dynamic networks

图 3 动态网络角色预测

3.1 问题定义

本文将动态网络结构预测转换为可以表示结构特征的角色预测问题,即:根据历史时刻网络的角色分布,预测未来时刻网络可能的角色分布情况.形式化表示为:给定动态网络 $D=\langle S_1,S_2,\ldots,S_T\rangle$,得到动态网络角色模型 $G=\langle G_1,G_2,\ldots,G_T\rangle$,$G_i$ 为 i 时刻节点角色矩阵,角色预测就是要得到 t+1 时刻网络的节点角色矩阵 G'_{t+1}.

对整个网络,角色预测就是要得到 t+1 时刻网络的节点角色矩阵,对每个节点,就是要预测节点 n 在 t+1 时刻的角色分布向量 g_{t+1}(G'_{t+1} 的第 n 行).将网络结构的预测问题划分成子问题,对每个节点而言,预测目标是一个向量,且节点的在角色向量的分布上并非独立不相关的.

向量自回归(VAR)[24]基于数据的统计性质建立模型,适合处理多个变量分析与预测.本文借鉴 VAR 方法的思路,提出了基于潜在角色的动态网络结构预测方法 LR-DNSP(latent role based dynamic network structure prediction).LR-DNSP 模型可以充分考虑前后向量序列之间的关系和角色之间的相互影响,通过向量自回归的方法,由历史时刻网络数据得到训练数据构建潜在角色预测模型,以下一时刻网络角色的分布情况作为预测目标.LR-DNSP 不仅利用多个历史时刻的属性信息还考虑了预测目标之间的相关性.

3.2 预测模型

本文预测模型的框架如图 4 所示.

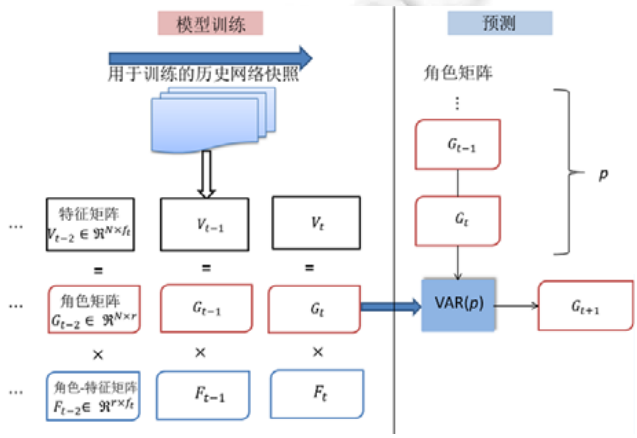


Fig.4 Model framework of LR-DNSP

图 4 LR-DNSP 模型框架

对所有的网络快照计算其特征矩阵,并进行分解得到角色矩阵,用一部分历史数据来训练模型,剩下的一部分数据用来预测.

本文预测目标为节点在下一时刻的角色分布向量,记为 $y=[y_1,\ldots,y_m]$.在本文后续实验部分的 3 个数据集通过计算验证,当角色个数 r 取 4 时模型代价最小,因此此处 m=4.为了后续实验比对,本文先将预测变量视为多个独立的单变量,为每个 y_i 通过自回归方法分别建立一个 LR-DNSP(AR)模型,如图 5 虚线所选每列所示.对网络中每个节点的角色分布进行预测,将所有预测结果合并起来作为最终角色矩阵.

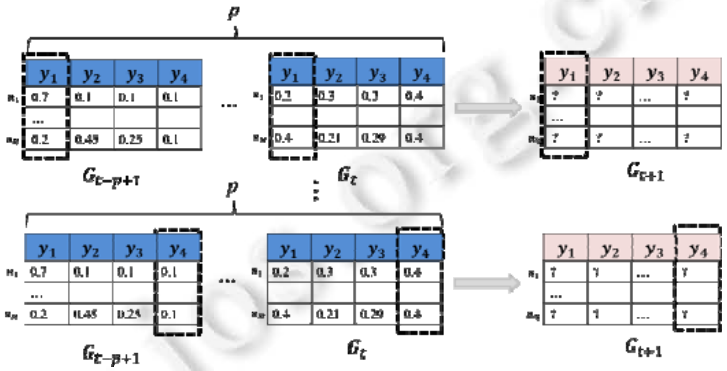


Fig.5 Prediction model of LR-DNSP (AR)

图 5 LR-DNSP(AR)预测模型

为了实验对比中的公正性,以上 LR-DNSP(AR)模型分别选取预测结果最佳的阶数 p.

考虑到多个预测目标之间存在的相互影响,在上述采用自回归方法模型的基础上,本文将预测目标视为向量,提出解决动态网络角色预测问题的向量自回归模型,对每个节点直接预测 t+1 时刻的角色分布向量(如图 6 所示).

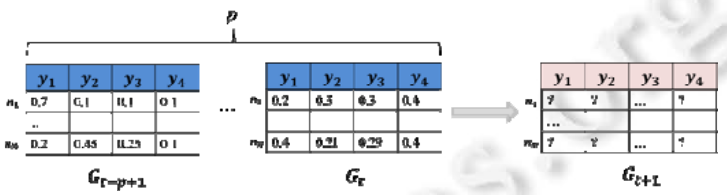


Fig.6 Prediction model of LR-DNSP

图 6 LR-DNSP 预测模型

对节点 n 用最近的 p 个历史时刻的角色向量预测模型如下:

$$\hat{G}_{t+1}^{(n)}=\sum_{j=1}^{p}\Phi_j G_{t-j+1}^{(n)}+\varepsilon_t+\alpha \tag{2}$$

其中,p 是向量自回归模型的阶数;$\hat{G}_{t+1}^{(n)}$ 为角色矩阵第 n 行在 t+1 时刻的预测值;α为(m×1)常数项向量,本文模型中,m=4;Φ_j 是自回归系数的一个矩阵,j=1,2,…,p;ε_t 为角色在时间 t+1 的分布误差,服从均值为零的高斯分布.

模型参数可以采用最小二乘估计,也可采用最大似然估计.本文模型中的参数学习通过解决以下问题的最优解:

$$\arg\min_{\Phi_j,\Phi_j,\ldots,\Phi_j}\frac{1}{2}\frac{1}{t-p}\sum_{t'=p+1}^{t}\|G_{t'}-\hat{G}_{t'}\|_F^2 \tag{3}$$

其中,$\hat{G}_{t'}=\sum_{j=1}^{p}\Phi_j G_{t-j+1}^{(n)}+\varepsilon_t+\alpha$ 是 t'时刻的预测值.模型回归的阶数 p 可以通过最小化式(4)中的最终预测误差

(final prediction error,简称 FPE)来确定:

$$FPE(p)=\frac{1}{t-p}\sum_{t'=p+1}^{t}\|G_{t'}-\hat{G}_{t'}\|_F \tag{4}$$

LR-DNSP 模型的训练过程算法可抽象如下.

算法 2. LR-DNSP 模型训练过程.

输入:网络快照 $D=\langle S_1,S_2,\ldots,S_T\rangle$和对应的节点角色序列 $G=\langle G_1,G_2,\ldots,G_T\rangle$;

输出:向量自回归模型的参数$\{\phi_1,\phi_2,\ldots,\phi_p\}$和$\alpha$;向量自回归模型的介数:$p$.

1. **for** i=1 **to** t
2. $N(i)=\{G_1^{(i)},G_2^{(i)},...,G_t^{(i)}\}$; //从角色矩阵中提取节点的角色序列
3. **end for**
4. Determine p by minimizing the final prediction error defined by Eq.(4);
5. Apply gradient descent method to learn $\{\phi_1,\phi_2,\ldots,\phi_p\}$ and α of the autoregressive model defined by Eq.(2) by minimizing the objective function defined by Eq.(3);

算法的第 1 行~第 3 行是从各个时刻的角色矩阵中提取节点的角色序列,第 4 行根据最小化最终预测误差来确定模型回归的阶数 p,算法的第 5 行使用梯度下降的方法通过求解公式(4).

算法 3. 角色分布预测算法.

输入:最近 p 个时刻节点角色序列:$\{G_{t-p+1},\ldots,G_{t-1},G_t\}$;

输出:t+1 时刻节点的角色分布矩阵: $\hat{G}_{t+1}$.

1. **for** i=t–p+1 to t
2. $N(i)=\{G_{t-p+1}^{(i)},G_{t-p}^{(i)},...,G_t^{(i)}\}$; //从角色矩阵中提取节点的角色序列
3. **end for**
4. **for** j=1 to N
5. Estimate $\hat{G}_{t+1}^{(j)}$ by the autoregressive model defined by Eq.(2);
6. $\hat{G}_{t+1}=\{\hat{G}_{t+1}^{(j)},\hat{G}_{t+1}^{(j)},...,\hat{G}_{t+1}^{(j)}\}$
7. **end for**

4 实验及分析

4.1 数据集

本文选取 3 个具有代表意义的动态网络数据集:Enron(http://konect.uni-koblenz.de/networks/enron)、Facebook(http://konect.uni-koblenz.de/networks/facebook-wosn-wall)、DBLP(http://dblp.uni-trier.de/xml/),每个数据集的规模各不相同,均来自公开网站.表 1 列出了 3 个数据集的详细信息,“*”表示平均值.

Table 1 Datasets details

表 1 数据集详细信息

数据集	节点数	*边数	*特征数	角色数	快照数	快照长度	时间区间
Enron	2 114	16 413	70	4	24	1 month	2000.4~2002.3
Facebook	5 111	14 438	144	4	24	1 month	2007.1~2008.12
DBLP	29 747	96 874	159	4	16	1 year	1996~2011

为简化起见,本文假设网络的节点数目保持不变,因而只考虑在研究时间段内一直出现的节点.对以上 3 个网络均建模为无向加权网络,采用权值衰减的方法计算边的权重.节点 a 和节点 b 之间的边在 t 时刻的权值为

$$w_{a,b}(t)=\sum_i w_i e^{-\lambda(t-t_i)} \tag{5}$$

其中:w_i 是 t_i 时刻节点 a 和节点 b 之间事件的权重(如节点间发送电子邮件数、合作发表论文数等),相应的邻接矩阵序列为 $A=\langle A_1,A_2,\ldots,A_r\rangle$;本文实验中,取 $\lambda=1$.

4.2 对比方法

本文使用 4 种方法作为对比.

(1) PRE(baseline):使用前一时刻网络的角色分布作为要预测的下一时刻的网络角色矩阵,即用时刻 t 的网络角色分布矩阵作为时刻 t+1 时所求得表示网络结构的角色矩阵,即: $G'_{t+1}=G_t$;

(2) TM(transition model):此模型是相关工作中所介绍 WSDM2014 的转移矩阵方法[13],主旨思想是计算 t–1 和 t 时刻的角色矩阵 G_{t-1} 和 G_t,使用非负矩阵分解得到角色转移矩阵 T:$G_{s(t-1)}T\approx G_{s(t)}$,由 G_t 和 T 相乘得到 t+1 时刻的目标角色矩阵 G'_{t+1}:$G'_{t+1}=G_tT$;

(3) AR:将角色分布向量($y=[y_1,\ldots,y_m]$)视为多个独立的单变量,为每个 y_i 分别建立一个自回归(AR)模型,将每个模型最后得到的预测结果合并起来作为最终的预测目标矩阵;

(4) MTR(multiple target regression):将多目标回归问题转化为多个单目标回归,假设预测目标之间相互独立.利用历史时刻的网络快照计算节点的一部分度量属性(包括度、加权度、介数、PageRank 值、离心率、聚集系数),用来建立一般的广义线性回归模型来预测角色的分布.

4.3 评估方法

本文采用两种策略对预测模型进行评价.

(a) 计算预测的角色矩阵 $\hat{G}_{t+1}$ 和真实的角色矩阵 G_{t+1} 的差异,使用均方根误差(root mean square error,简称 RMSE)和平均绝对误差(mean absolute error,简称 MAE)两种度量指标:

$$RMSE=\sqrt{\frac{\|G_{t+1}-\hat{G}_{t+1}\|_F^2}{N}} \tag{6}$$

$$MAE=\frac{\sum_{i=1}^{N}\|G_{t+1}^{(i)}-\hat{G}_{t+1}^{(i)}\|_1}{N} \tag{7}$$

其中,G_{t+1} 为 t+1 时刻网络的真实角色矩阵,$\hat{G}_{t+1}$ 为预测得到的 t+1 时刻网络角色矩阵,$G_{t+1}^{(i)}$ 和 $\hat{G}_{t+1}^{(i)}$ 分别为 G_{t+1} 和 $\hat{G}_{t+1}$ 的第 i 行向量,$\|\cdot\|_F$ 表示矩阵的 F 范,$\|\cdot\|_1$ 表示 1-范数,N 为网络中节点的数目.以上两个指标可以从不同方面评估预测值与实际值之间的差异度.误差指标值越小,表示预测越精确.

(b) 用 $\hat{G}_{t+1}$ 来预测 t+1 时刻节点的角色,评价节点角色分类的准确性.对每个节点来说,可以将角色预测视为多类标分类问题,因此,本文通过 $\hat{G}_{t+1}$ 来预测节点的角色属性.节点 i 的角色属性的真实类标是 G_{t+1} 的第 i 行,对应节点 i 的预测类标签是预测矩阵 $\hat{G}_{t+1}$ 的第 i 行的节点角色.

4.4 实验效果

4.4.1 角色解释

根据第 3.2 节所介绍的角色解释方法,用传统的度量和邻居节点的分布来刻画角色是一种有效的解释方法.本文以 Facebook 数据集为例对角色做出解释(Enron 与 DBLP 数据集类似),首先选取前面介绍的 8 种常见的度量属性(包括度、加权度、介数、特征向量中心性、接近中心度、聚集系数、PageRank 值以及离心率等)计算节点的度量矩阵,结果如图 7 所示.图 8 为根据邻居节点的角色分布统计,得到角色之间的关系.

图 7 中由 Facebook 网络得到的 4 种角色都具有明显的特征:角色 R_3 在度、加权度、介数(表示网络中包含节点 i 的所有最短路的条数占所有最短路条数的百分比,反映节点对网络资源控制的程度,类似于 gatekeeping 的角色)特征向量中心性、PageRank 等度量上都有表现,且取值都较大,但在离心率的取值上最小;角色 R_4 在聚集系数和离心率两种度量上取值较大,尤其是在离心率的取值明显高于其他角色,而在其他度量上取值均较小;R_1 在度和特征向量中心性取值均比较明显;而 R_2 在各度量的取值均不突出.从图 8 可看出:R_3 更倾向与其他角色的节点相连;而 R_4 仅在自己角色的上的分布比较显著,所表现出的同质性比较强;角色 R_3 的节点与角色为 R_1 的

节点相连更紧密,并且这种紧密性是双向的.通过以上度量属性和邻居节点角色分布的解释,可推断 R_3 表示位于重要位置的节点(例如中心节点);R_1 为具有重要邻居的节点;而 R_2 代表网络中普遍存在的较一般的节点;R_4 可能是边缘节点甚至是非激活的节点.

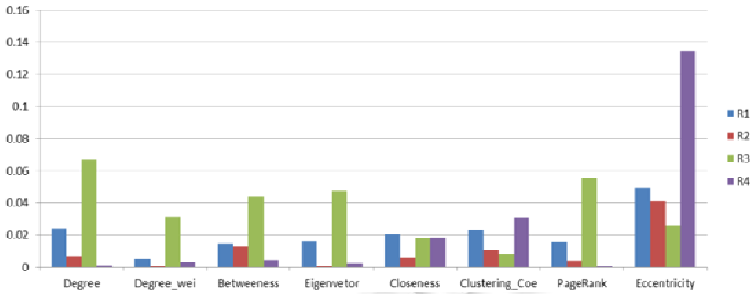


Fig.7 Role explaination—NodeSense

图 7 角色解释——NodeSense

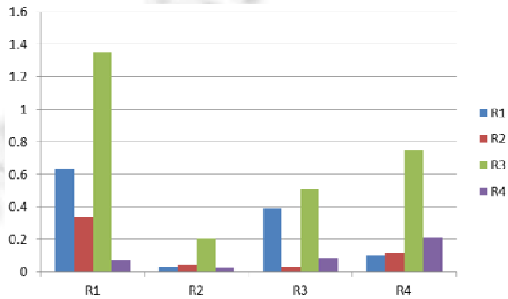


Fig.8 Role explaination—NeighborSense

图 8 角色解释——NeighborSense

4.4.2 角色预测

首先验证角色序列的平稳性,在 Enron,Facebook 以及 DBLP 这 3 个数据集中随机选取 300 个节点,计算每个节点的前 12 个快照序列的自相关函数.对一个序列$\{s_1,s_2,\ldots,s_t\}$,自相关函数(autocorrelation)计算如下:

$$r_q=\frac{\sum_{i=q+1}^{t}(s_i-\bar{s})(s_{i-q}-\bar{s})}{\sum_{i=1}^{t}(s_i-\bar{s})^2} \tag{8}$$

其中,q 为滞后阶数,$\bar{s}$ 为序列的均值.图 9 为 3 个数据值中节点的平均自相关曲线图.从图中可以看出,相关函数随滞后阶数 q 的增加而快速下降并趋向于 0,表明角色序列是平稳的[24].

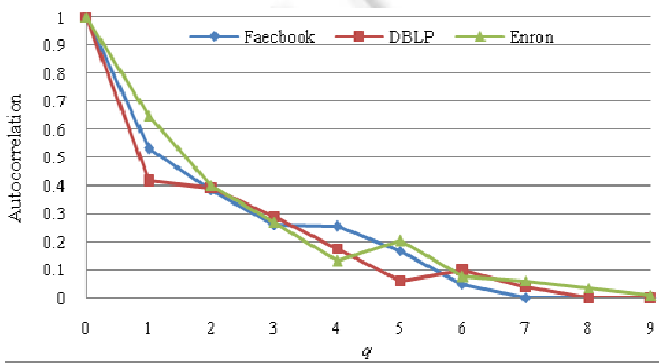


Fig.9 Autocorrelation function varies with the lag order q

图 9 自相关函数随滞后阶数 q 的变化

在本文角色预测模型中,向量自回归的阶数直接会影响模型预测的准确性,因此接下来验证阶数 p 对预测结果的影响.

图 10 是在 3 个数据集上,p 的取值为 1~5,分别计算预测矩阵 $\hat{G}_{t+1}$ 和真实矩阵 G_{t+1} 的均方根误差($RMAE \propto FPE$)的结果.

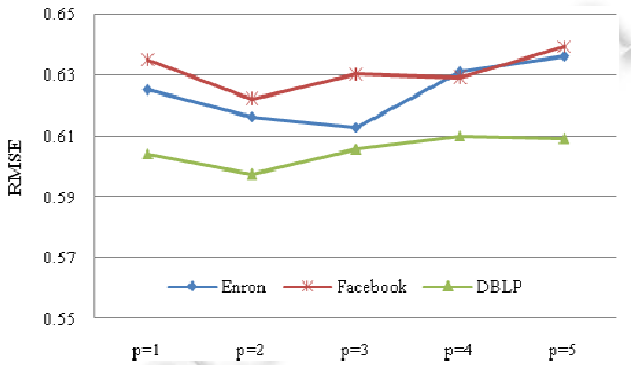
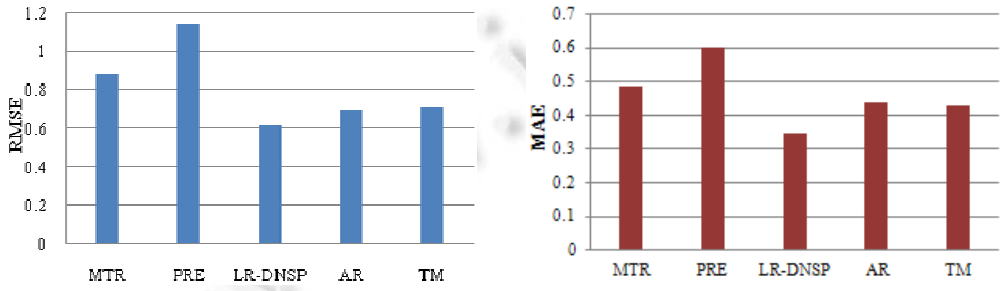


Fig.10 Influence of p to the prediction result

图 10 阶数 p 对预测结果的影响

从图 10 的结果可以看出:Enron 数据集在 p=3 时均方根误差最小,Facebok 和 DBLP 数据集在 p=2 时均方根误差最小.由图 10 可以看出:p 的值从 1 变化到 5,预测效果并没有随着阶数的增大而更准确,这说明预测时模型里包含的历史时刻属性个数并不是越多结果的准确性越高.事实上这也是合理的,p 的取值决定有多少历史快照将影响当前时刻网络的角色分布:当 p 的设定值过大时,模型需要预测的参数增多,较早时刻历史快照也会对预测结果带来干扰;另一方面,当 p 的设定值太小时,模型将会欠拟合,导致残差增加.综合考虑,最优的 p 取决于数据集,并通过公式(4)中的最终预测误差来确定,结果与分析相吻合,表明当前时刻网络的结构分布受更近时刻的历史数据的影响更大.

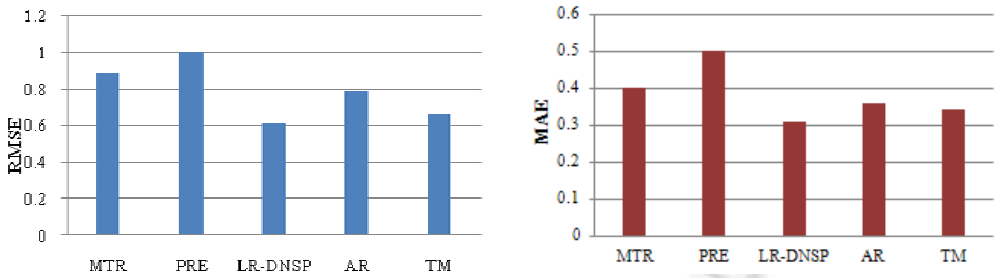
下面将分别在 Enron,Facebook 以及 DBLP 这 3 个数据集上验证 LR-DNSP 模型的有效性.首先验证评估方法(a),此处 Enron 数据集阶数 p 设定为 3,后两个数据集阶数 p 设定为 2.图 11~图 13 分别为 3 个数据集的预测效果,Enron,Facebook 数据集中均选取前 19 快照用来训练,预测 G_{t+1},19≤t≤23.DBLP 数据集选取前 12 快照用来训练,预测 G_{t+1},11≤t≤15.



(a) RMSE (b) MAE

Fig.11 Predicition in Enron dataset

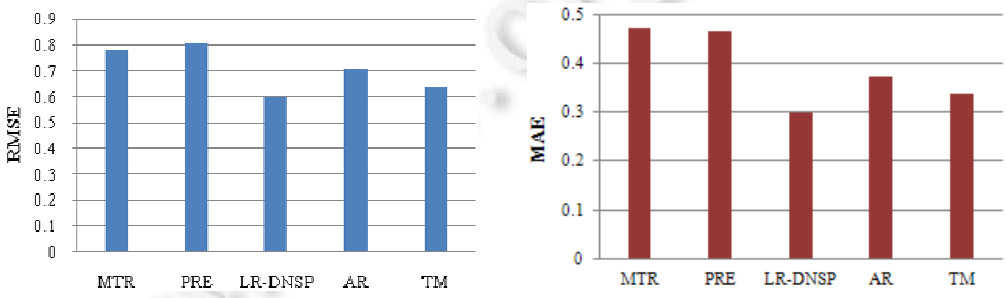
图 11 Enron 数据集预测效果



(a) RMSE (b) MAE

Fig.12 Predicition in Facebook dataset

图 12 Facebook 数据集预测效果



(a) RMSE (b) MAE

Fig.13 Predicition in DBLP dataset

图 13 DBLP 数据集预测效果

从图 11~图 13 的预测结果可以看出:本文提出的 LR-DNSP 模型在 3 个规模不同的真实数据集均取得了很好的预测效果,与对预测目标分别建立回归模型的 AR 方法相比,LR-DNSP 能得到更准确的预测值,这说明节点在 4 种角色上的取值是有一定联系的.再看与 TM 方法的对比,LR-DNSP 在 Enron 数据集上的回归阶数为 3,在后两个数据集上的回归阶数为 2,也就是说,Enron 使用了 3 个最近历史时刻的数据进行预测,DBLP 和 Facebook 使用了 2 个最近历史时刻的数据预测,而 TM 只使用了前一个时刻的数据来预测,在 3 个数据集上的预测效果均不如本文模型.MTR 模型是根据历史时刻网络计算节点的一部分度量属性(包括度、加权度、介数、PageRank 值、离心率、聚集系数),用来建立一般的广义线性回归模型来预测角色的分布,从结果可以看出,预测效果仅优于 baseline,说明节点在下一时刻的角色分布不仅取决于节点的度量值,而受节点历史时刻的角色分布影响更大一点.

接下来验证评估方法(b),即预测节点角色类标分类的准确性,图 14~图 16 分别为 3 个数据集预测的准确性.

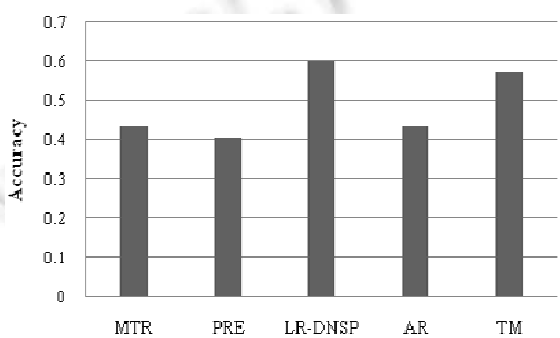


Fig.14 Classification accuracy of rolein Enorn dataset

图 14 Enron 数据集角色分类准确性

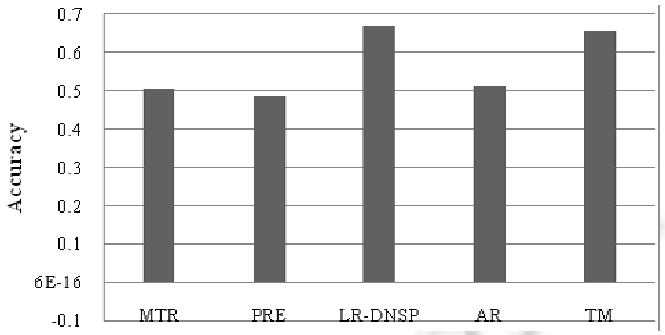


Fig.15 Classification accuracy of rolein Facebook dataset

图 15 Facebook 数据集角色分类准确性

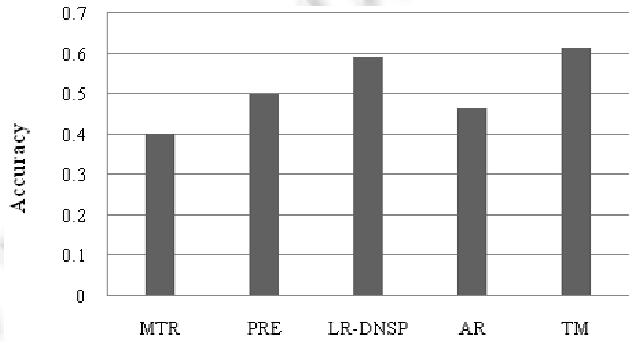


Fig.16 Classification accuracy of role in DBLP dataset

图 16 DBLP 数据集角色分类准确性

从 3 个数据集角色分类准确性的结果可以看出,本文提出的 LR-DNSP 模型和 TM 方法在准确性上优于其他 3 种方法;还可以观察到:相比 Enron 和 Facebook 网络,在 DBLP 数据集角色分类的准确性上,PRE 的预测效果与本文模型以及 TM 方法更接近.PRE 直接使用当前时刻节点的角色作为下一时刻的预测值,这是基于相邻时刻网络结构不会发生剧烈变化的假设;继续分析可知:在 Enron 和 Facebook 网络中,PRE 表现很差.这说明 Enron 和 Facebook 网络结构并不稳定.事实上,2001 年 7 月~10 月间,Enron 公司发生了巨大的人事调动,年底公司破产,Enron 网络结构理应是不稳定的,而 2008 年的 Facebook 网络正处于超速发展中,6 月正式成为全球最大、增长最快的社交网络,Facebook 的网络结构也不会是稳定的.但在 DBLP 网络中,学者一般具有稳定的研究兴趣和合作学者,网络结构自然相对稳定.

以上实验结果表明:在角色分类准确性上,TM 方法和本文提出的 LR-DNSP 模型不相上下;但综合以上两种评价策略,LR-DNSP 模型在预测结果和分类准确性的整体效果优于其他方法.

4.5 时间开销

下面分别在 Facebook 和 DBLP 数据集上进行时间开销的实验分析,Facebook 数据集上分别选取 1 000,2 000,3 000,4 000 和 5 000 个节点,在 DBLP 数据集上分别选取 3 000,6 000,9 000,12 000,15 000,18 000 和 21 000 个节点进行实验.从结果可以看出:随着节点的增加,运行时间呈线性增长.验证了本文预测模型的可扩展性.

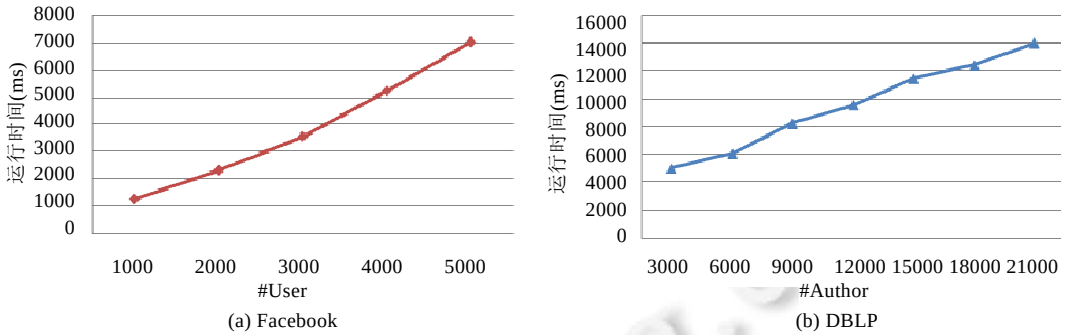


Fig.17 Influence of node's number *n* on time overhead

图 17 节点数 *n* 对时间开销的影响

5 总 结

基于动态信息网络的广泛应用前景及角色发现在动态网络中有限的研究现状,本文致力于研究动态信息网络中基于角色发现的结构演化与预测问题.在网络结构难以量化呈现和分析的基础上,本文提出了简化该问题的新思路,将基于递归地提取特征的角色发现方法引入动态信息网络的结构演化分析中,同时给出了两种角色解释的方法;进一步以角色发现的结构模型为基础,将网络结构的预测问题转化为角色预测问题,提出了基于潜在角色的动态网络结构预测方法 LR-DNSP.动态网络是目前复杂网络研究领域中极具活力的新兴研究方向,相比于静态网络的研究成果,目前动态网络的研究还处于起步阶段,本文只针对其中的演化分析和预测问题进行了研究.传统静态网络中的许多问题都需要在动态网络中得到进一步研究与扩展,未来的研究工作将继续关注动态网络的演化问题,进一步优化本文算法,以达到更好的效果.

References:

[1] Wang XF, Li X, Chen GR. The Theory and Application of Complex Networks. Beijing: Tsinghua University Press, 2006 (in Chinese).

[2] Viswanath B, Mislove A, Cha MY, Gummadi KP. On the evolution of user interaction in Facebook. In: Proc. of the Workshop on Online Social Networks. 2009. 37–42.

[3] McAuley J, Leskovec J. Learning to discover social circles in ego networks. In: Proc. of the Advances in Neural Information Processing Systems. 2012. 548–556.

[4] Bifet A, Frank E. Sentiment knowledge discovery in Twitter streaming data. In: Proc. of the Int'l Conf. on Discovery Science (DS 2010). Canberra: DBLP, 2010. 1–15.

[5] Leskovec J. Dynamics of large networks [Ph.D. Thesis]. Pittsburgh: Carnegie Mellon University, 2008.

[6] Han J, Yan X, Yu PS. Scalable OLAP and mining of information networks. In: Proc. of the 12th Int'l Conf. on Extending Database Technology: Advances in Database Technology. ACM Press, 2009. 1159–1159.

[7] Li C, Feng BQ, Li YM, Hu SL. Role-based structural evolution and prediction in dynamic networks. Ruan Jian Xue Bao/Journal of Software, 2017,28(3):663–675 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5164.htm [doi: 10.13328/j.cnki.jos.005164]

[8] Holme P, Saramäki J. Temporal networks. Physics Reports, 2012,519(3):97–125.

[9] Henderson K, Gallagher B, Eliassi-Rad T, *et al*. Rolx: Structural role extraction & mining in large graphs. In: Proc. of the 18th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2012. 1231–1239.

[10] Gilpin S, Eliassi-Rad T, Davidson I. Guided learning for role discovery (glrd): Framework, algorithms, and applications. In: Proc. of the 19th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2013. 113–121.

[11] Rossi R, Gallagher B, Neville J, *et al*. Dynamic behavioral mixed-membership model for large evolving networks. arXiv preprint arXiv:1205.2056, 2012.

[12] Rossi R, Gallagher B, Neville J, *et al*. Role-dynamics: Fast mining of large dynamic networks. In: Proc. of the 21st Int'l Conf. Companion on World Wide Web. ACM Press, 2012. 997−1006.

[13] Rossi RA, Gallagher B, Neville J, *et al*. Modeling dynamic behavior in large evolving graphs. In: Proc. of the 6th ACM Int'l Conf. on Web Search and Data Mining. ACM Press, 2014. 667−676.

[14] Scripps J, Tan PN, Esfahanian AH. Node roles and community structure in networks. In: Proc. of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis. ACM Press, 2007. 26−35.

[15] Rossi R, Fahmy S, Talukder N. A multi-level approach for evaluating Internet topology generators. In: Proc. of the IFIP Networking Conf. IEEE, 2013. 1−9.

[16] Varki A. Biological roles of oligosaccharides: All of the theories are correct. Glycobiology, 1993,3(2):97−130.

[17] Luczkovich JJ, Borgatti SP, Johnson JC, *et al*. Defining and measuring trophic role similarity in food Webs using regular equivalence. Journal of Theoretical Biology, 2003,220(3):303−321.

[18] Ma H, King I, Lyu MR. Mining Web graphs for recommendations. IEEE Trans. on Knowledge and Data Engineering, 2012,24(6): 1051−1064.

[19] McDowell LK, Gupta KM, Aha DW. Cautious collective classification. Journal of Machine Learning Research, 2009,10(Dec): 2777−2836.

[20] Everett MG, Borgatti SP. Regular equivalence: General theory. Journal of Mathematical Sociology, 1994,19(1):29−52.

[21] Rossi RA, Ahmed NK. Role discovery in networks. IEEE Trans. on Knowledge and Data Engineering, 2015,27(4):1112−1131.

[22] Henderson K, Gallagher B, Li L, *et al*. It's who you know: Graph mining using recursive structural features. In: Proc. of the 17th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2011. 663−671.

[23] Rissanen J. Modeling by shortest data description. Automatica, 1978,14(5):465−471.

[24] Cryer JD, Chan KS. Time Series Analysis with Applications in *r*. 2nd ed., 2008.

附中文参考文献:

[1] 汪小帆,李翔,陈关荣.复杂网络理论及其应用.北京:清华大学出版社,2006.

[7] 李川,冯冰清,李艳梅,胡绍林,杨宁,唐常杰.动态信息网络中基于角色的结构演化与预测.软件学报,2017,28(3):663−675. http://www.jos.org.cn/1000-9825/5164.htm [doi: 10.13328/j.cnki.jos.005164]

冯冰清(1990－),女,陕西渭南人,工程师,主要研究领域为数据库,数据挖掘,信息网络,航天器故障诊断.

胡绍林(1964－),男,博士,教授,博士生导师,主要研究领域为航天安全与大数据分析技术,过程监控与故障诊断技术,复杂系统建模与仿真.

郭栋(1986－),男,工程师,主要研究领域为软件工程,数据挖掘,航天器故障诊断.

钟晓歌(1990－),女,工程师,主要研究领域为软件工程.

李佩钰(1990－),女,工程师,主要研究领域为计算机安全.