DOI:10.7544/issn1000-1239.2015.20148134
52(11):2508-2516,2015

基于事务日志的社会网络抽取

陈 创[1] 徐 波[1] 肖仰华[1] 施 佺[2] 汪 卫[1]
[1](复旦大学计算机科学技术学院 上海 200433)
[2](南通大学计算机科学与技术学院 江苏南通 226019)
(chuangchen13@fudan.edu.cn)

Extracting Social Network from Transaction Logs

Chen Chuang[1], Xu Bo[1], Xiao Yanghua[1], Shi Quan[2], and Wang Wei[1]
[1](*School of Computer Science, Fudan University, Shanghai* 200433)
[2](*School of Computer Science & Technology, Nantong University, Nantong, Jiangsu* 226019)

Abstract Social network analysis (SNA) is a popular research topic in the field of data mining, and the quality and the scale of networks are extremely important for the research. But most previous studies are conducted on large online social networks or small real social networks. Online social networks are only the approximation of real social networks, and in general they have different properties. Some research conducts on real social networks which are constructed from the survey of quite small population. Social network study expects large real social network data. Transaction logs generated by modern software systems make it possible to construct social networks from large real social data. This paper conducts a case study of extracting student social network (SSN) from school card system transaction logs to explore how to extract social network from transaction logs. Firstly, we build a relation extracting model based on co-occurrence. Then, we define probability coefficient for edge based on the weight of edge and Jaccard coefficient, filtering noisy edges from the network. We conduct our method on real transaction logs data, and the experiments show that our method can generate social networks with high precision. The topology of the network shows that this social network has small-world network features and a scale-free degree distribution.

Key words social network analysis (SNA); transaction logs; co-occurrence; pruning; dependence measure

摘 要 社会网络分析(social network analysis, SNA)是数据挖掘领域的一个重要研究方向,社会网络数据的质量和规模对研究十分重要.在当前的社会网络分析研究中,大多数是基于社交网站生成的社会网络,社交网站生成的在线社会网络只是对真实社会网络近似模拟,其现象、结论无法代表真实社会网络;少数基于真实社会网络的研究中,由于数据采集难度较大,往往只能使用规模有限的社会网络,从而降低了分析结果的可信程度.现代软件系统产生大量的事务日志让构建基于真实环境的社会网络成为可能.以高校学生卡管理系统产生的事务日志为例,探索如何从海量事务日志中抽取社会网络.根据事务日志的特征,建立以共现(co-occurrence)特征为基础的网络抽取模型,抽取出所有可能构成这个社会网络的边;定义了一个基于边的权重和 Jaccard 相关性系数的边存在系数,识别网络中的噪音边,筛选噪音边;最后,通过同班级比率分析和网络拓扑结构分析,对抽取的网络进行验证.实验结果表明,所抽取

收稿日期:2014-10-17;**修回日期**:2014-12-16
基金项目:国家自然科学基金项目(61003001,61170006,61171132,61033010);江苏省自然科学基金项目(BK2010280)
通信作者:肖仰华(shawyh@fudan.edu.cn)

的网络具有很高的同班级比率,该抽取模型具有较好效果,同时该网络具有小世界网络(small-world)特征和满足无标度(scale-free)度分布,符合常见社会网络特征.

关键词　社会网络分析;事务日志;共现;剪枝;相关性度量

中图法分类号　TP391

社会网络分析(social network analysis, SNA)是对社会网络的关系结构及其属性加以分析的一套规范和方法[1],主要关注参与者之间的关系与网络结构及其对参与个体和整个群体的影响.社会网络分析在职业流动分析、城市化对个体幸福的影响、世界政治和经济体系分析等领域得以广泛应用,发挥了重要作用.近年来,社会网络分析日益成为了数据挖掘领域的热门研究问题,受到了来自学术界和工业界的广泛关注.在2013年知识发现与数据挖掘会议(ACM KDD 2013)的研究论文有20多篇是关于社会网络分析的.

1) 社会网络分析存在的问题

在社会网络分析相关研究中,社会网络数据的质量和规模是决定结果可信与否的2个关键因素.遗憾的是,当前社会网络分析存在如下问题:

① 当前大部分社会网络分析研究是基于在线社会网络开展,如Facebook,Twitter以及国内的新浪微博等;但是,在线社会网络只是对真实社会网络近似,其分析结果无法直接应用到真实社会网络上.在线社交网络难以完全代替真实社会网络,一方面,真实世界中并不是所有人都使用社交网站,在线社交网络难以刻画完整的社会关系;另一方面,网络世界的虚拟特性导致在线社交网络与现实社会网络存在巨大的结构差异.因此,基于在线社交网络分析观测到的现象和结论难以直接应用到真实社会网络上.

② 由于数据采集难度较大,少数基于真实社会网络的研究往往受限于小规模社会网络,难以取得统计意义下显著的分析结果.当前部分针对真实世界社会网络开展的研究工作,主要通过对人工收集的数据分析而获取社会网络.例如,早在20世纪70年代初,Zachary[2]通过观察构建了1个拥有34个节点的空手道乐部社会网络.近年来MIT Human Dynamics Lab通过采集手机位置数据、通话数据构建真实社会网络[3],由于这一方法成本较大,构造出的社会网络只有几百个节点.真实社会网络的有限规模极大地降低了分析结果的统计显著性,从而削弱了结论的可信度.

因此,当前社会网络分析研究急需高质量、具有一定规模的真实社会网络数据.本文提出一种基于事务日志的全新真实社会网络抽取方法,以满足当前研究对于真实社会网络的迫切需求.

2) 基于事务日志的社会网络抽取

随着信息化进程的推进,各类信息系统得以大量部署,产生了大量事务日志(transaction logs),这些日志为构建真实世界的社会网络成为可能.事务日志是由设备、软件、应用或者1个系统产生的记录该系统所提供的活动的日志文件.事务日志通常包括活动内容、执行时间、参与者的信息以及一些其他信息.现代应用系统如ERP、MIS、CRM、电子商务系统以及监控系统中通常会系统地产生大量此类事务日志.

通过日志信息,可以推断用户在时空上的共现(co-occurrence)(即在相同时间或/和相同地点同时出现),而时空共现往往蕴含着紧密的社会关系,这是本文基于事务日志抽取社会网络的基本思想.如例1所示,对于校园卡刷卡日志,可以根据学生好友之间倾向于同时去同一地点消费,从而推断好友关系.

例1. 从学生卡管理系统事务日志中抽取社会网络.国内某大学通过学生卡管理系统管理学生的学生卡使用情况,该系统每月产生约2 000 000条学生卡刷卡事务日志.图1为该系统产生的事务日志样例,每一条日志分别记录了刷卡流水号、用户ID、用户姓名、消费的商户ID与名称、刷卡的时间、消费金额、刷卡的POS机ID等信息,每一条交易日志记录了1个学生的1次刷卡行为.大学生好友之间通常倾向于同时进行消费,如同时去食堂吃饭、同时去超市购物以及同时去图书馆学习等,这种好友行为特征势必体现于日志数据之中.因此,根据学生是否多次同时出现在同一地点消费,可以推断2个学生是好友的可能性.利用这种方法对海量的学生刷卡日志进行分析即可获得全校学生的社会网络.

该方法具有以下优点:①数据来源广泛.该方法所依赖的事务日志数据广泛存在于各类信息系统之中.②数据全面.日志数据通常涉及用户全体,因此

容易构造出涉及全体用户的社会网络. ③构造出的社会网络更为接近真实社会网络. 事务日志较为客观地反映了人们在真实世界的活动,基于此构造出的社会网络也因而更为接近真实的社会网网络.

serialNum	userID	userName	shopID	shopName	date	time	amount	POSID
169574	79265	许	243020001	一食堂一组	2012-09-18	16:47:28.000	150.00	0007
169577	75850	吴	243020001	一食堂一组	2012-09-18	16:48:18.000	150.00	0007
169580	101116	魏	243020002	一食堂二组	2012-09-18	16:52:39.000	150.00	0015
169671	77387	徐	243020003	一食堂面包房	2012-09-18	16:55:26.000	150.00	0028
169674	45720	王	243020002	一食堂二组	2012-09-18	16:55:54.000	650.00	0022
169676	97463	李	243020022	图书馆电子阅览室	2012-09-18	16:59:18.000	201.00	1010
169583	76312	周	243020002	一食堂二组	2012-09-18	16:52:18.000	600.00	0026
169586	42587	房	243020004	一食堂三组	2012-09-18	16:52:43.000	400.00	0033
169589	79179	肖	243020006	一食堂五组	2012-09-18	16:52:48.000	200.00	0049
169592	41781	卢	243020001	一食堂一组	2012-09-18	16:52:57.000	630.00	0002
169595	75528	陆	243020001	一食堂一组	2012-09-18	16:53:30.000	300.00	0008
169598	42782	谢	243020002	一食堂二组	2012-09-18	16:52:39.000	550.00	0017

Fig. 1 Sample data of transaction logs of school card management system.

图 1 学生卡管理系统事务日志样例

3) 研究挑战

本文以学生卡管理系统事务日志为例,研究基于事务日志的社会网络抽取. 本研究有如下挑战:

① 如何准确定义共现关系模型. 利用刷卡日志推断学生的时空共现,在准确界定时空的邻近性方面仍然存在很多挑战. 首先,如何认定用户的地点共现. 不同的 POS 终端消费性质不同,地点共现认定方式也不一样. 其次,如何认定用户的时间共现. 需要设定 1 个合理的时间共现阈值作为时间共现的认定依据. 如何准确定义时空共现是本研究的挑战之一.

② 如何构建可信的社会网络. 给定共现关系模型之后,容易得到用户之间的共现关系矩阵. 但是,如何从共现关系矩阵进一步构造可信的真实网络仍然面临挑战. 一般而言,共现是好友关系的必要条件,但不一定充分. 因此,还需要结合社会网络的全局结构特征和用户个体行为特征对共现关系进行筛选.

③ 如何验证抽取出的网络真实性. 在大规模真实社会网络中,不能通过实证调查获取 1 个标准网络来验证网络抽取的正确性. 如何确定参数、使抽取的网络最接近于真实网络,以及如何验证抽取出的网络是否满足社会网络的特征,也是本文的主要挑战.

4) 主要贡献

为了解决上述挑战,本文提出了下述方案并作出了相应的贡献:

① 以学生卡事务日志数据为例,通过分析消费场所的性质来定义地点共现;通过分析不同阈值的同班级比率(将在 1.2 节定义)来选取时间共现阈值,定义了 1 个准确的共现关系模型.

② 定义了 1 个基于边的权重和 Jaccard 相关性系数的社会关系网络边存在系数,用以识别真实的社会关系,去除网络中的噪音边,从共现矩阵中抽取出可信的社会网络.

③ 提出通过计算所抽取出来的网络的同班级比率和对网络拓扑结构分析的方法,对抽取效果进行验证. 结果表明,利用我们的方法抽取出的社会网络具有很高的同班级比率,符合预期效果. 同时,网络拓扑结构分析结果表明,基于事务日志抽取出的社会网络满足无标度(scale-free)特性和小世界(small world)特性,符合常见的社会网络特征.

为保护隐私,本文将数据中涉及的真实姓名略去,以 ID 号代替,抽取出的社会网络仅供科研参考,不作其他任何用途.

1 社会网络构建

事务日志中,数据本身没有明确的指示参与者之间的好友关系. 所以,对于事务日志,只能结合系统特征分析系统用户的行为特征,将所有可能构成这个网络的边抽取出来,构建 1 个初始的社会网络,然后识别网络中的噪音边,筛选之后剩下的边即构成社会关系网络.

本节以学生卡事务日志为例,定义以共现为基础的社会网络抽取模型,将所有可能构成这个网络的边都抽取出来,构建 1 个初始的社会关系网络. 在学生中,好友之间倾向于结伴消费,如同时去食堂吃饭、同时去超市购物以及同时去图书馆等. 根据这个规律,给出如下定义:

定义 1. 共现. 在学生卡事务日志数据中,如 2 个学生 X,Y 在很小的时间间隔内在同一地点消费,则称 X,Y 共现(co-occurrence),2 个学生组成 1 个共现组.

如果 2 个学生之间多次共现,那么这 2 个学生之间可能为好友关系. 本节将分别讨论地点共现和时间共现.

1.1 消费场所分类

由于学校有不同类型的消费场所,所以,在定义 1

中的“同一地点消费”,在不同类型的消费场所的认定方法不一样.例如互为好友的2个学生 X,Y 在超市购物,大多数情况下2人都是在超市同一台POS机刷卡消费;而当 X,Y 二人同时去澡堂洗澡或者去开水房打水时,由于澡堂或开水房的每个水龙头对应着1个POS机,2人虽然同时消费,但是在不同POS机刷卡.由于这2种不同类型的消费地点有不同特征,定义网络抽取模型的方式也不一样,所以需要将所有的POS机终端分为不同类型.

通过对所有POS机对应场所进行统计发现,按场所分,该校总共有7种类型的POS机,分别为浴室、食堂、开水房、超市、图书馆、车队以及机房.按照不同场所消费性质的不同,将消费场所分为如下2类:

A——好友之间同时消费时,在同一台POS机刷卡,如超市等.

B——好友之间同时消费时,在同一商户的不同POS机刷卡,如开水房.

对于这2种不同类型的消费场所分类如表1:

Table 1　Classification of Location

表1　消费场所分类表

Location Type	Shop Type	Amount
A	Dinner Hall, Library, Supermarket, Bus, Compuer Room	447
B	Bathroom, Boiler Room	290

在类型 A 的场所中,2人在同一POS机下刷卡消费称为同一地点消费;在类型 B 的场所中,同一地点消费指2人在同一商户的同一台或者任意2台POS机刷卡消费.

将消费地点按照消费性质分类后,即可将刷卡事务日志按照消费地点分为不同的时间序列:对于类型 A 场所的日志,每个POS机对应1个刷卡日志序列,按刷卡时间排序;对于类型 B 场所的日志,每个商户对应1个刷卡日志序列.按时间排序.2类地点的刷卡序列具有同样的结构,如图1所示.

1.2　时间共现阈值

在1个刷卡日志序列中,2条日志对应2人是否共现,依据是他们的刷卡时间间隔是否小于该序列的共现阈值.共现阈值有如下2种选取方法:

1)非固定时间间隔.将1个刷卡序列中相邻刷卡的2人视为时间共现.但是,由于类型 B 的消费场所中,同时消费的2个好友不在同一POS机上刷卡,他们在刷卡序列中可能不相邻,这种方法可能会漏掉很多共现组.

2)为每个刷卡日志序列选取1个时间窗口 ΔT 作为该序列的共现阈值.如果2人刷卡时间间隔小于 ΔT,视为2人在时间上共现.ΔT 的大小选取原则是,抽取出来的共现组能尽量多地包含真实好友.

由于学生社会网络数据量大,无法通过实证调查获取1个真实网络来验证所抽取网络的正确性.实验发现,在同班同学中,互为好友的人数要远远多于随机人群.所以,本文通过验证抽取出来的共现组中对应的2个学生是否属于同一个班级,来近似确定抽取的有效性.对于1个刷卡日志序列在某一个共现阈值下的同班级比率定义如下:

定义2. 同班级比率.1个刷卡日志序列在特定的共现阈值 ΔT 下,抽取出来的共现组中,2人来自同一个班级的共现组数占所有共现组数的百分比为该序列的同班级比率.

为了确定最好的共现阈值,我们进行了如下实验:选取1个日志序列,将其中的所有刷卡间隔计算出来,并从小到大排列成1个时间间隔序列,分别取该序列中每5个百分点对应的值作为共现阈值 ΔT,计算每个阈值对应的同班级比率.

精确率(precision, P)和召回率(recall, R)以及结合两者的综合评价指标值 F 值(F-measure, F)是信息检索中最常用的衡量指标.本文中,1个刷卡日志序列的同班级比率的这3个指标定义如下:

$$P=\frac{D}{W}, \tag{1}$$

$$R=\frac{D}{Q}, \tag{2}$$

$$F=2\times\frac{P\times Q}{P+Q}. \tag{3}$$

其中,W 为总共现组数,D 为共现组中2人是同班同学组数,Q 为所有出现学生能组成的同班同学组数.

图2和图3分别为随机选取食堂、开水房、超市、图书馆的1个日志序列,在不同 ΔT 值时的同班

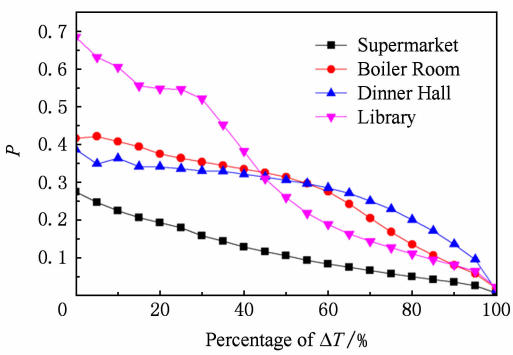


Fig. 2　Co-class precision of different ΔT.

图2　不同 ΔT 时的同班同学的精确度

级比率的 P 和 R 分布. 如图 2 和图 3 所示,当 ΔT 取值越大时,P 越小,R 越大.

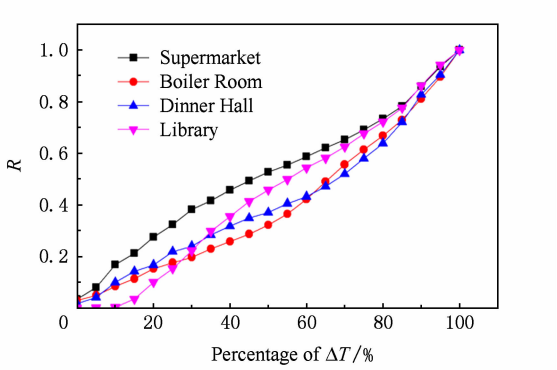


Fig. 3 Co-class recall of different ΔT.

图 3 不同 ΔT 时的同班同学的召回率

F 是综合考虑 P 和 R 两方面因素的综合评价指标. 本文中,F 值越大,对应阈值 ΔT 的效果越好. 每个日志序列的同班同学的 F 值,如图 4 所示. 可知,对于不同类型的日志序列,同班同学的 F 峰值都在中位数附近,所以,在网络抽取时,共现阈值 ΔT 分别取每组刷卡时间间隔的中位数.

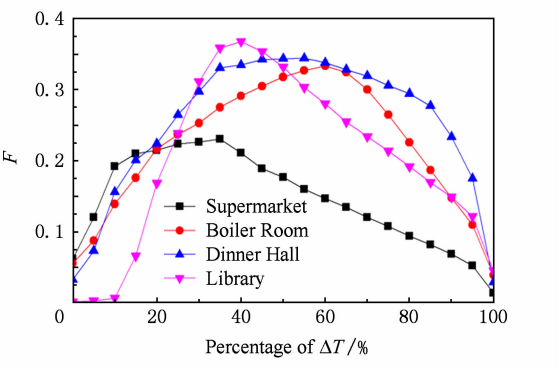


Fig. 4 Co-class F-measure of different ΔT.

图 4 不同 ΔT 时的同班同学的 F 值

1.3 网络抽取

如果 2 人多次共现,那么 2 人就可能为好友关系. 我们可以通过找出所有日志序列中的共现组,构建 1 个初始的学生社会网络.

首先,我们定义 1 个初始学生社会网络(primary student social network, PSSN). 在初始状态下,该网络没有节点和边,PSSN 为空. 当 2 个学生 X,Y 共现时,认为 2 人有可能为好友关系,向 PSSN 中添加 1 条边:如当前网络中没有点 X 或点 Y、或者没有该边,将没有的点添加至网络,然后在 2 点间添加 1 条边 $E(X,Y)$,该边的权重 $w_{X,Y}=1$;如边 $E(X,Y)$已经在于该网络中,则将该边的权重加 1,即 $w_{X,Y}+=1$. 具体抽取过程如例 2.

例 2. 如图 5 所示,Stu A 和 Stu B 前后相隔 $\Delta T=2$ s(2 s 小于阈值)在计算机实验室刷卡,这 2 个学生之间可能为好友关系,将他们对应的 ID 分别作为 2 个节点添加到 PSSN 中;然后给这 2 个节点之间添加 1 条边,边的权重为 1,如果这条边已经在 PSSN 中,则将他们的权重加 1.

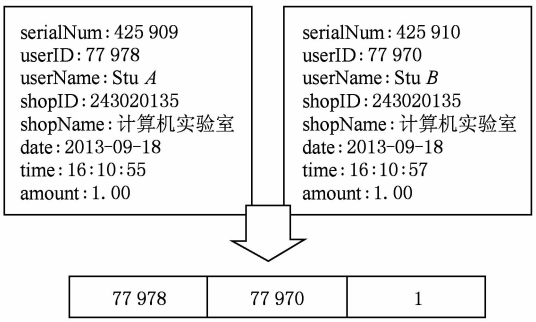


Fig. 5 Construct edge by co-occurrence.

图 5 利用共现构建边

将所有的刷卡日志序列中满足共现定义的人都找出来,将对应的边都添加到 PSSN 中,初始学生社会网络便构建完成,PSSN 的矩阵表示形式即为学生的共现矩阵.

2 社会网络优化

在 PSSN 构建的时候,我们为所有满足共现定义的学生之间都添加了 1 条边,2 人共现的次数越多,边的权重越大. 但是,在学生卡使用场景中,有 2 种情况能使 2 个学生产生共现:1)好友间同时同地点消费,抽取出的是该网络真实边;2)2 人只是偶然前后同时同地点消费,他们之间没有任何关系,这种情况抽取出来的边为这个网络的噪音边. 本节讨论如何识别 PSSN 中的噪音边,来优化社会网络. 本文将从 PSSN 优化剪枝后的学生社会网络简称为 SSN(student social network).

为了识别 PSSN 中的噪音边,我们为边定义 1 个存在系数 ρ,表示 1 条边在这个网络中存在的可能性. 将噪音边筛选问题转变成 1 个计算存在系数 ρ 值为前 M 的边集合的问题. M 为 SSN 中保留的边数,这是 1 个根据网络的实际应用场景确定的参数. 2.3 节将讨论影响存在系数 ρ 值的各个因素和 ρ 的定义.

2.1 边的权重

在 PSSN 中,1 条边的权重表示该边对应的 2 个学生共现的次数. 2 人共现的次数越多,边的权重越大.

当边的权重很小时,表明2人共现的次数很少,这样的边大多数都是偶然共现生成的,2人之间并不是好友关系.根据共现模型生成的初始网络中有很多这样的边,所以,对网络优化的第1步是将权重太小的边剪枝.在对PSSN优化过程中,可以将权重$w=1$和$w=2$的边剪掉,然后再进行后续处理.

权重大小是衡量1条边存在可能性的重要因素,但不是唯一因素.由于消费习惯不同,不同学生在刷卡日志中有不同的刷卡基数,这个刷卡基数直接影响了节点对应边的权重大小,所以边的权重不能唯一决定边的存在系数.

2.2　节点间相关性

为了消除刷卡基数的影响,除了权重之外,还要看边的2点间的相关性,如果2点间关系紧密程度强于其他点,这条边存在的可能性越大.本节为边定义1个2点间的Jaccard相关性系数来衡量1条边的2个节点间相关性.

在PSSN中,设点u的度为k,与点u关联的每条边的权重分别为$w_{u,i}(i=1,2,\cdots,k)$.边$E(u,v)$为点u与点v之间的边,权重为$w_{u,v}$.如下定义系数μ:

$$\mu(E,u)=\frac{w_{u,v}}{\sum_{i=1}^{k}w_{u,i}}. \tag{4}$$

其中,$\mu(E,u)$为边E的权重与点u关联的所有边权重之和的比值,表示这条边相对于与这个节点关联的其他边的重要程度,即点u和点v之间关系紧密程度与点u和其他相邻节点关系紧密程度的比较.

1条边有2个点,我们在μ值定义的基础上定义Jaccard系数来表示1条边2点之间的相关性.其定义如下:

$$J_{u,v}=\frac{w_{u,v}}{\sum_{i=1}^{k_u}w_{u,i}+\sum_{j=1}^{k_v}w_{j,u}}, \tag{5}$$

其中,k_u和k_v分别表示点u、点v的度数.

Jaccard相关性系数表示了2个节点之间关系的紧密程度,相对于其他的相关性度量方式,本文中的Jaccard相关性系数是基于边的权重和与点关联的所有其他边权重之和的比值系数μ定义的,能够更好地度量2点间的相关性;同时,Jaccard相关性系数只依赖于与这2点有关的局部数据,不依赖于整个网络中其他点的刷卡基数.

2.3　边的存在系数

由2.1节和2.2节可知,判断1条边是否为噪音边,与2个因素有关:边的权重以及点之间的相关性系数.结合这2个因素,我们为边定义了1个存在系数ρ.识别网络中的噪音边的过程,即求该网络中每一条边的存在系数ρ的值,1条边的ρ值越小,是噪音边的可能性越大.存在系数ρ的定义如下:

$$\rho(u,v)=\frac{1}{Z}w_{u,v}^{r}\times J_{u,v}. \tag{6}$$

1条边$E(u,v)$的存在系数为$\rho(u,v)$;Z为归一化因子,$Z=\sum_{\substack{i,j=1\\ i\neq j}}^{i,j=N}(w_{u_i,u_j}^{r}\times J_{u_i,u_j})$,其中$N$为点总个数;参数$r$是1个杠杆系数,决定边的权重在存在系数中的重要程度,取值范围为$r\geqslant 0$,r越大边的权值在边存在系数中越重要.当$r=0$时,$w_{u,v}^{r}=1$,ρ值等于Jaccard相关性系数的值.参数r决定了保留的边中不同权值的边所占比例,r取值越大,倾向于保留权重大的边;r取值越小,倾向于保留相关性系数大的边.

根据实际的社会网络节点规模和网络使用场景,可以决定最后抽取出的网络的边的规模,边存在系数ρ的阈值可以根据边的规模决定.比如,网络场景需要保留M条边,计算出所有边的ρ值,ρ值为前M的边所构成的网络即为最终抽取的社会网络.

3　实　　验

本节使用国内某高校所有34 375个在校学生2个月内产生的3 794 703条刷卡事务日志为例,进行学生社会网络抽取.通过抽取出来的网络的同班级比率和分析网络拓扑结构对提出的方法进行验证.实验结果表明,抽取出来的网络具有较高的同班级比率,同时,该网络度分布满足幂率分布(power-law distribution),具有无标度网络特征,且具有较高的聚集系数和较短的平均最短路径,是1个小世界网络(small world network).该网络满足常见社会网络特征.

3.1　初始网络抽取

实验数据如图1所示,包含所有与学生卡消费相关的属性.按照社会网络抽取模型的定义,将2个月的刷卡日志按照不同消费场所POS机类型,分为不同的刷卡日志序列.选取每个日志序列的所有刷卡时间间隔的中位数(证明见1.2节)作为这个序列的共现时间间隔阈值ΔT.

对所有日志序列抽取之后,产生的初始学生社会网络PSSN中有$M=6\,984\,528$条边,权重大小范围为1～117,但是,大多数边集中在权重很小的范围是1个长尾分布.这个网络的平均度分布$\langle K\rangle=2M/N=406.37$.显然,PSSN平均度分布远远超过

1 个正常的社会关系网络,在正常的学校环境中,不可能平均每个人拥有 400 多个好友,所以这个网络存在噪音边,需要将噪音边识别并删除.

3.2 网络优化

在学校中经常与某个学生一起进行吃饭、打水、去图书馆等活动的好友个数不会太多,一般只有几个人.我们假设在这些学生中,平均每人的这种好友有 k 个,平均度分布$\langle K\rangle=k$,优化后的网络中应保留存在系数 ρ 最大的前 $M=kN/2$ 条边,N 为剪枝后的学生社会网络 SSN 的节点个数.在本实验场景中,以 $k\approx4$ 为例,通过筛选噪音边,保留 $M=60\,000$ 条边.在计算每条边的存在系数 ρ 之前,先将 $w=1$ 和 $w=2$ 的小权重边剪枝.

图 6 表示了在 r 的不同取值下边存在系数 ρ 值为前 M 的边组成的网络 SSN 的边权重分布.实验结果中,所有 r 的取值下,权重大于 20 的边都被保留下来了,所以图 6 中省去了权重大于 24 的分布.

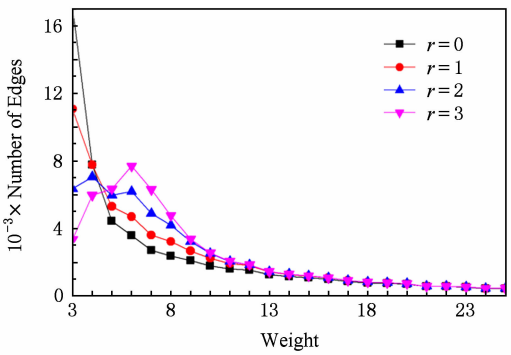


Fig. 6 Weight distribution of SSN with different r.

图 6 不同 r 取值下 SSN 的权值分布

由图 6 可知,当 $r=0$ 时,$w^r_{u,v}=1$,ρ 值等于 Jaccard 相关性系数值,权重小的边保留较多;当 r 的取值越大时,最终网络保留的边中权重较大的边所占比例越大;当 $r=3$ 时,权值大于 6 的边基本上全部被保留.

3.3 同班级比率分析

学生社会网络数据量庞大,无法通过实证调查的方法来获取真实网络情况.我们通过比较抽取出来的边对应学生同班级比率(见 1.2 节定义 2),来近似检验抽取出来网络的正确性.在学生中,大多数好友来自同一个班级,如果抽取出来的网络中边对应的 2 人为同班同学的比率越高,那么这个网络越接近于真实网络.

图 7 为参数 r 选取不同值时所抽取的网络中的同班级比率.图 7 中,Random 图示计算了在初始学生社会网络 PSSN 中随机选取 M 条边所组成的网络对应的同班级比率.由随机选取的边所组成的网络的同班级比率很小,只有 0.018 左右,远远小于使用优化算法抽取出来的网络中的同班级比率.这也充分说明了我们抽取模型的有效性.

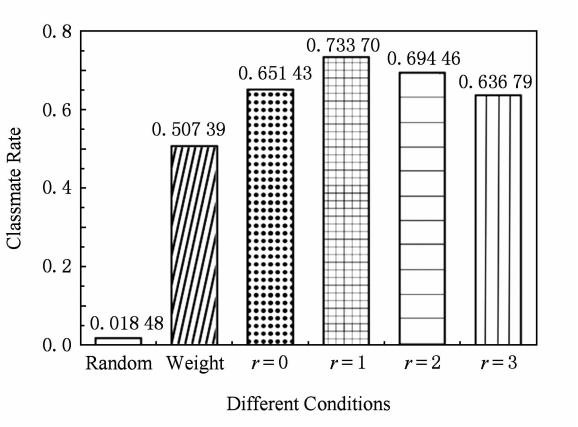


Fig. 7 Co-class rate.

图 7 同班级比率

图 7 中,Weight 图示表示直接截取权重最大的 M 条边所组成网络的同班级比率.这个比率小于所有 r 取值下的同班级比率,这表明,如果只将边的权重作为优化因子,网络抽取效率较低,验证了边的权重不能唯一决定 1 条边的存在性.同样,当 $r=0$ 时,边存在系数中 $w^r_{u,v}=1$,ρ 值等于 Jaccard 相关性系数的值,这种情况下同班级比率也较小.

在 r 的其他取值中,当 r 越大时,权值占的比重越大,倾向于保留权重大的边集合.实验表明,在参数 r 的各个不同取值下,当 $r=1$ 时,抽取出来的网络同班级比率最大,最接近真实网络.

图 7 中,$r=1$ 时同班级比率为 73.4%左右,最接近于真实网络,这是 1 个合理的比率.因为我们抽取的是好友关系,在学生中并不是所有好友都来自同一个班级,还存在一些跨班级好友.

同班级比率检测结果表明,在参数 r 的适当取值下,我们的网络抽取模型具有较好的效果.

3.4 SSN 网络分析

本节将对参数 $r=1$ 时实验得到的网络 SSN 进行拓扑结构分析,了解学生社会网络的特征.在抽取出的学生社会网络 SSN 中,有 $N=27\,755$ 个节点和 $M=60\,000$ 条边.为了方便对该网络的拓扑结构分析,我们将这个网络转换为 1 个无权网络,即所有的边权重变为 1.

3.4.1 网络度分布

1 个点的度数为直接与该点连接的点的数量,点 i 的度 $k_i=\sum_{j=1}^{N}a_{i,j}$.在 SSN 中,平均度大小为$\langle K\rangle=2M/N=4.324$.

度是衡量1个点在网络中中心性的指标,点的度表示该点在网络中的重要程度,度越大,该点越重要.网络的度分布能够体现1个网络的拓扑结构,在自然界中,大多数网络遵循无标度分布(scale-free degree distributions)[4],无标度分布可以用幂律定律(power law)$P(k)\sim k^{-r}$来界定.在学生社会网络SSN中,度分布如图8所示,可知SSN的度分布满足幂律定律,$r=2.7\pm0.1$,遵循无标度分布.

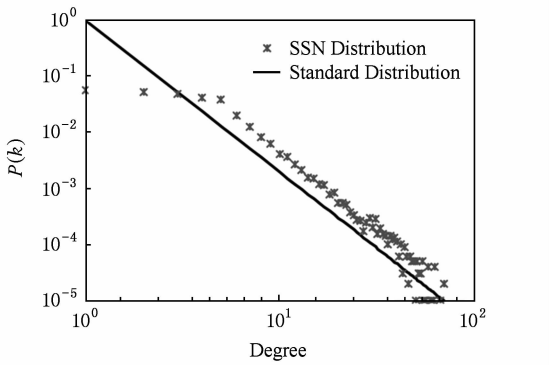


Fig. 8　Degree distribution of SSN.

图8　SSN度分布图

3.4.2　小世界网络

复杂网络中,点之间的平均最短路径定义为$L=\frac{1}{N(N-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{i,j=N}L_{i,j}$,其中$L_{i,j}$表示点$i$到点$j$的最短路径包含的边的数目.SSN的平均最短路径$L=4.2432$,小于1个随机网络的平均最短路径($L_{\text{rand}}\sim\ln N/\ln\langle K\rangle=6.9876$).

复杂网络中,1个点的聚集系数(clustering coefficient)C_i定义为它所有相邻节点之间连接边数目与可能的最大连接边数目的比值.1个网络的平均聚集系数为$C=\frac{1}{N}\sum_{i=1}^{N}C_i$.实验验证SSN的平均聚集系数为$C=0.6047$,远大于随机网络的平均聚集系数($C_{\text{rand}}\sim\langle K\rangle/N=1.5579\text{e}^{-4}$).

通常,把复杂网络的聚集系数较大和平均最短路径较小这2个统计特征结合在一起称为小世界效应,具有这种效应的网络称为小世界网络.SSN中,平均最短路径小于标准随机网路,聚类大于标准随机网络1个数量级.这2个统计值表明,我们抽取出来的社会网络SSN为1个小世界网络.

前人对社会网络的分析研究表明,社会网络是1个复杂网络,通常具有无标度分布和小世界网络特征.本文实验分析结果表明,抽取出的网络具有这2个特征,符合常见社会网络特征.

4　相关工作

社会网络分析研究具有70多年历史,一直都是非常热门的研究领域,有很多相关的著作和研究论文,如文献[1,5-6].社会网络数据的规模和可信度是影响研究的重要因素.除了社交网站产生的社会网络之外,有很多人尝试利用真实社会环境下的数据构建社会网络.最开始,Zachary[2]构建的由跆拳道俱乐部34个成员组成的社会网络成为了社会网络分析的经典案例.后来,很多研究通过电子邮件通信记录构建社会网络,如Neustaedter等人[7]利用email通信记录构建社会网络,对用户的邮件进行分类.还有一些研究,如Jung等人[8]、Eagle等人[3,9]和Kazienko[10]通过采集手机使用数据来构建社会网络.国内也有很多学者进行相关研究,如文献[11].

Eagle等人的reality mining项目从2004年开始,通过采集手机数据建立社会网络进行分析,在文献[3]中,使用手机软件对94位参与者的9个月的手机使用数据进采集,构建1个社会网络.通过将通信人的位置区域、2人间距离,以及通信时间是白天、晚上、工作日或周末等信息结合,为不同场景赋予不同的权重,抽取出2种社会关系:公事关系(in-role)和私人关系(extra-role).与本文对比,他们的工作生成的网络规模小,数据采集代价太大,而且使用GPS、蓝牙等工具采集的距离数据也不够精确.

Ishizuka等人[12-15]的大量工作中,利用共现特征从Web数据、搜索引擎数据中抽取社会网络,如科学家合作关系网络等.但是从搜索引擎中抽取的社会网络代价远远大于从事务日志中抽取的社会网络代价.在Ishizuka等人的工作中,主要利用互信息(mutual information)、Jaccard相关性系数、Cosine相关性系数、辛普森系数(Simpson coefficient)等相关性系数对使用共现抽取出来的边进行重要性度量.本文根据事务日志的特征,定义1个基于Jaccard相关性系数和边的权重的存在系数,度量边的重要性,为初始社会网络筛选掉噪音边.

5　总结和工作展望

当前社会网络分析研究急需高质量、具有一定规模的真实社会网络数据.现代软件系统产生的大量事务日志数据,让抽取基于真实社会环境的社会网络成为了可能.本文通过以高效的学生卡管理

系统事务日志为例,研究了如何从事务日志中抽取社会网络.本文利用共现特征,建立了1个从事务日志中抽取社会网络的模型;定义了1个基于边的权重和Jaccard相关性系数的边存在系数,为社会网络筛选噪音边;本文通过分析同班级比率和网络拓扑结构,对实验结果进行验证.实验中结果表明,我们抽取出来的社会网络具有较高的同班级比率,符合预期假设;同时,拓扑结构分析表明,该网络满足无标度特性和小世界特性,符合常见社会网络特征.

本文只讨论了从学生卡管理系统产生的事务数据抽取学生关系网络的情况,在将来工作中,将会把相应理论运用到其他系统产生的事务数据上.

参 考 文 献

[1] Furht B. Handbook of Social Network Technologies and Applications [M]. Berlin: Springer, 2010: 10-25

[2] Zachary W. An information flow model for conflict and fission in small groups [J]. Journal of Anthropological Research, 1977, 33(4): 452-473

[3] Eagle N, Pentland A S, Lazer D. Inferring friendship network structure by using mobile phone data [J]. Proceedings of the National Academy of Sciences, 2009, 106(36): 15274-15278

[4] Albert R, Barabási A L. Statistical mechanics of complex networks [J]. Reviews of Modern Physics, 2002, 74(1): 47-55

[5] Wasserman S. Social Network Analysis: Methods and Applications [M]. Cambridge, UK: Cambridge University Press, 1994: 45-50

[6] Scott J, Carrington P J. The SAGE Handbook of Social Network Analysis [M]. New York: SAGE Publications, 2011: 120-150

[7] Neustaedter C, Brush A J B, Smith M A, et al. The social network and relationship finder: Social sorting for email triage [C] //Proc of Conf on Emailand Anti-Spam (CEAS'05). Piscataway, NJ: IEEE, 2005: 122-129

[8] Jung J J, Choi K S, Park S H. Discovering mobile social networks by semantic technologies [G] //Handbook of Social Network Technologies and Applications. Berlin: Springer, 2010: 223-239

[9] Eagle N, Pentland A. Reality mining: Sensing complex social systems [J]. Personal and Ubiquitous Computing, 2006, 10(4): 255-268

[10] Kazienko P. Expansion of telecommunication social networks [G] //Cooperative Design, Visualization, and Engineering. Berlin: Springer, 2007: 404-412

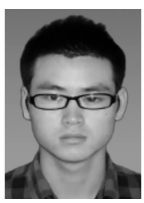
[11] Dou Binglin, Li Shusong, Zhang Shiyong. Social network analysis based on structure [J]. Chinese Journal of Computers, 2012, 35(4): 741-753 (in Chinese)
(窦炳琳, 李澍淞, 张世永. 基于结构的社会网络分析[J]. 计算机学报, 2012, 35(4): 741-753)

[12] Ishizuka M. Exploiting macro and micro relations toward Web intelligence [C] //Proc of Pacific Rim Int Conf on Artificial Intelligence (PRICAI'10). Berlin: Springer, 2010: 4-7

[13] Matsuo Y, Mori J, Hamasaki M, et al. POLYPHONET: An advanced social network extraction system from the Web [J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2007, 5(4): 262-278

[14] Mori J, Tsujishita T, Matsuo Y, et al. Extracting relations in social networks from the Web using similarity between collective contexts [C] //Proc of 2006 Int Semantic Web Conf (Web-ISWC'06). Berlin: Springer, 2006: 487-500

[15] Jin Y, Ishizuka M, Matsuo Y. Extracting inter-firm networks from the world wide Web using a general-purpose search engine [J]. Online Information Review, 2008, 32(2): 196-210



Chen Chuang, born in 1990. Master candidate. His current research interests include social network analysis, data mining and knowledge graph.

Xu Bo, born in 1988. PhD candidate. His current research interests include social network analysis, data mining and knowledge graph (bolang1988@gmail.com).

Xiao Yanghua, born in 1980. PhD, associate professor. Member of China Computer Federation. His current research interests include big data and knowledge graph.

Shi Quan, born in 1973. Master, associate professor. Member of China Computer Federation. His current research interests include social network analysis, Web database (sq@ntu.edu.cn).

Wang Wei, born in 1970. PhD, professor. Fellow member of China Computer Federation. His current research interests include database and knowledge graph (weiwang1@fudan.edu.cn).