单位代码 10006
学　　号 12211129
分 类 号 TP391

北京航空航天大學
BEIHANG UNIVERSITY

毕业设计 (论文)

基于移动社交网络中的关系识别的研究与实现

学 院 名 称　软件学院
专 业 名 称　软件工程专业
学 生 姓 名　冯溢濠
指 导 教 师　吕云翔教授

2016 年 06 月

北京航空航天大学

本科生毕业设计（论文）任务书

I、毕业设计（论文）题目：

基于移动社交网络中的关系识别的研究与实现

II、毕业设计（论文）使用的原始资料（数据）及设计技术要求：

原始资料及设计要求第一行

原始资料及设计要求第二行

原始资料及设计要求第三行

原始资料及设计要求第四行

原始资料及设计要求第五行

III、毕业设计（论文）工作内容：

工作内容第一行

工作内容第二行

工作内容第三行

工作内容第四行

工作内容第五行

工作内容第六行

IV、主要参考资料：

参考文献第一行

参考文献第二行

参考文献第三行

参考文献第四行

参考文献第五行

参考文献第六行

参考文献第七行

参考文献第八行

软件 学院 软件工程 专业类 122115 班

学生 冯溢濠

毕业设计 (论文) 时间： 2015 年 11 月 15 日至 2016 年 05 月 15 日

答辩时间： 2016 年 06 月 01 日

成　　绩：

指导教师：

兼职教师或答疑教师（并指出所负责部分）:

系（教研室）主任（签字）:

注：任务书应该附在已完成的毕业设计（论文）的首页。

本人声明

我声明，本论文及其研究工作是由本人在导师指导下独立完成的，在完成论文时所利用的一切资料均已在参考文献中列出。

作者： 冯溢濠

签字：

时间： 2016 年 06 月

基于移动社交网络中的关系识别的研究与实现

学　　生： 冯溢濠
指导教师： 吕云翔教授

摘　要

这部分当然是待写啦，等着所有的内容写完了再写摘要。

关键词：关系识别，社交网络，概率图模型

Inferring social ties in Mobile Social Networks

Author: Yihao Feng
Tutor: Prof.Yunxiang Lu

Abstract

Here is the Abstract in English. And this is a test sentence, You maybe enjoy the process of writting a thesis You can just ignore this.

This is another pargraph.

Key words: Social Ties, Social Networks, Probabilistic Graphical Models

目 录

1 绪论

1.1 研究背景

截至 2016 年 3 月, 全球最大社交网络平台 Facebook 活跃用户量已经突破 15.9 亿。中国最大的社交媒体微博用户也早在 2013 年突破了 6 亿用户，国际知名社交平台 Twitter 也在 2016 年突破了 13 亿的注册用户量。随着这些在线社交网络的迅猛发展以及移动智能电话的大规模普及，社交网络分析引起了越来越多的来自计算机、社会学、数学等领域的学者广泛关注。社交网络的原始定义[1, 2] 来自于社会学, 表示社会角色以及其交互关系的集合。而社会角色可以定义为独立的个人，也可以定义为家庭、学校或者国家等社会群体。而社会角色之间的的联系，则可以是任何无形 (如两个人之间的朋友关系) 或者有形 (如国与国之间的合作) 的交互关系，这些关系完全都可以由研究问题的学者自己定义。因此，由多个点 (即社会角色) 以及表示各个点之间关系 (即为交互关系) 的边所构成的网络，即为社交网络。在我们所生活的世界中，社交网络无处不在，如 Email 网络、学术网络或手机电话联系网络。虽然互联网的发展，出现了许多的在线社交网络, 如 Facebook、Twitter、Weibo 等等。这一系列的社交网络的兴起促进了海内外各个领域的学者对其的研究，而其研究结果又被用到广告营销、社会服务、公共安全等各种不同的领域。如图 1.1 即为一个来自 Friendster 网站的一个社交网络实例。图中可以看到，每个人都是一个点，而每条边表示两个人之间为朋友关系，将它们整体结合起来就构成了一个社交网络。

在以前的研究中，研究人员主要以在线社交网络为主要研究对象。但在移动互联网出现以前，用户只能通过 Web 页面登录到相应的社交网站。因此，以前的大多数研究都将重点放在了社交层次的用户交互上，而脱离于现实世界，从而限制了研究人员的研究思路与研究方法。但这几年，随着移动互联网的迅猛发展，越来越多的用户开始在移动终端使用相应的服务。同时, 移动开发者也开发了很多基于地理位置服务 (LBS, Location Based Service) 的移动应用。这一切使得研究社会网络有了新的方向与思路。移动社交网络 (Mobile Social Network) 是一种以移动终端为媒介、基于地理位置的社会网络。该网络相对于传统的社交网络更偏重于虚拟社会网络与现实世界之间的交互与联系，从而更加接近现实生活中的网络，从而研究者能研究的内容更加广泛、更加贴切现实生活中的实际情况。图 1.2 则展示了一个典型的移动社交网络。对于该网络中的用户来说，用

图 1.1 来自 Friendster 网站的社交网络实例[3]

户之间的虚拟联系 (如两人是否为朋友关系) 以及他们之间的联系属性构成了社交拓扑图。而对于现实生活中的用户来说，他们在现实世界中具有一定的时间、空间特征，而他们的位置轨迹以及关联等则构成了一幅位置移动图。在传统的社交网络研究中，研究者要么将重点放在社交拓扑图上，要么则主要研究用户之间的位置移动，很少有考虑两者之间的内在联系与共性。而移动社交网络研究中，研究者会同时研究用户的社交与位置轨迹等信息，将虚拟的社会网络与现实的物理世界有机地结合起来，从而能有效的研究用户虚拟世界与现实世界之间的内在共同特征与联系，更加真实的反映用户在现实世界中的行为与特征，为广告投放、社会服务等领域提供了更精准的信息与数据支撑。

当前，因为智能手机的大规模普及，和社交媒体 (如 Facebook、微博等) 的飞速发展，很多用户选择在移动端发状态或者消息。许多类似的应用在发状态消息的时，提供了是否要共享地理位置信息的选项。但很多用户基于隐私安全等考虑，并不愿意共享他们的地理位置，因此该数据的位置信息等大多处于缺失状态，很难构建用户的整个地理位置轨迹分布等信息。另外一点，虽然有用户在发状态时选择了共享他们的地理位置信息，但该模式很少和其他用户进行交流、通讯，因此实际上该交互模式中虚拟社交层次与地理交互层次是隔开的，研究者很少能够使用该数据来挖掘两者之间的内在联系，更不能挖掘他们之间的交互特征。因此，本文研究所采用的是移动手机通话数据。在移动社交

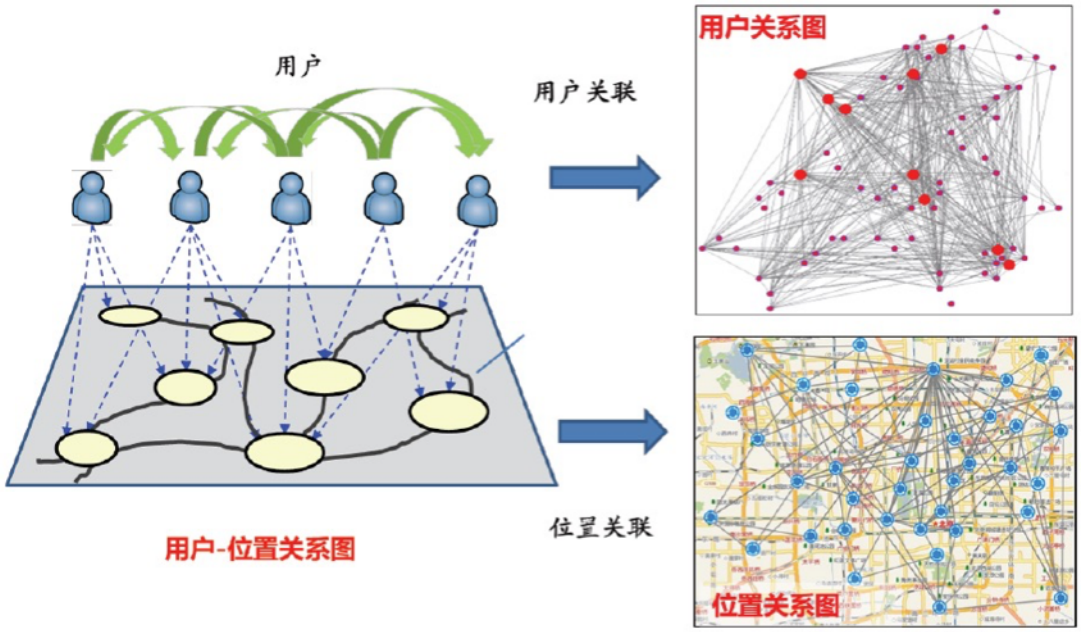


图 1.2 移动社交网络

网络中，移动通话数据具有其它数据不具备的优势，它记录丰富，时空信息完整，采样规律并且频繁，覆盖的人群阶层广泛，能有效的将用户的社交关系与地理轨迹交互联系起来。根据工信部 2014 年通讯运营业统计公报，我国截至 2014 年年底移动手机通话用户数量已经达到 12.86 亿户，普及率达到 94.5 部 / 百人[4], 并且增长十分迅速。如图 1.3, 可以看出这些年来我国居民拥有的移动终端的数目增加迅速，覆盖面越来越广。平均情况下已经接近人手一部的水平。因此使用移动通话数据网络具有非常高的普适性和广泛性 (而其它较大线上社交网络，如微博、Facebook、Twitter 等并不具有这些特点)。随着移动通话数据的进一步爆发式增长，基于移动通话数据的移动社交网络研究分析必将更加流行，当然也会带来了更大的机遇和挑战。

目前，移动社交网络的研究主要包括对社交网络中人群的画像识别、行为识别、关系预测以探讨社交网络与真实时空之间的关系。这些相关的研究在个性化推荐、用户轨迹预测、可疑用户监测等领域有着广泛的应用。例如在广告营销方面，确定了用户画像，就可以针对特定的用户人群投放更加精准的广告。另外，在追踪和调查犯罪嫌疑人的时候，我们确定了犯罪嫌疑人以及周围的社交关系，就能进一步的帮助相关执法机构缩小侦查范围，帮助公安机关迅速确定犯罪嫌疑人。

随着手机实名制的普及，越来越多的用户在运营商等急了个人的身份信息。当前很多家庭或者公司办理了家庭套餐、工作套餐等业务。这一套餐为我们精准把握用户之间的关系提供了数据支撑。尽管我国在大力推广手机实名制，但是全国仍然有将近 3 亿的

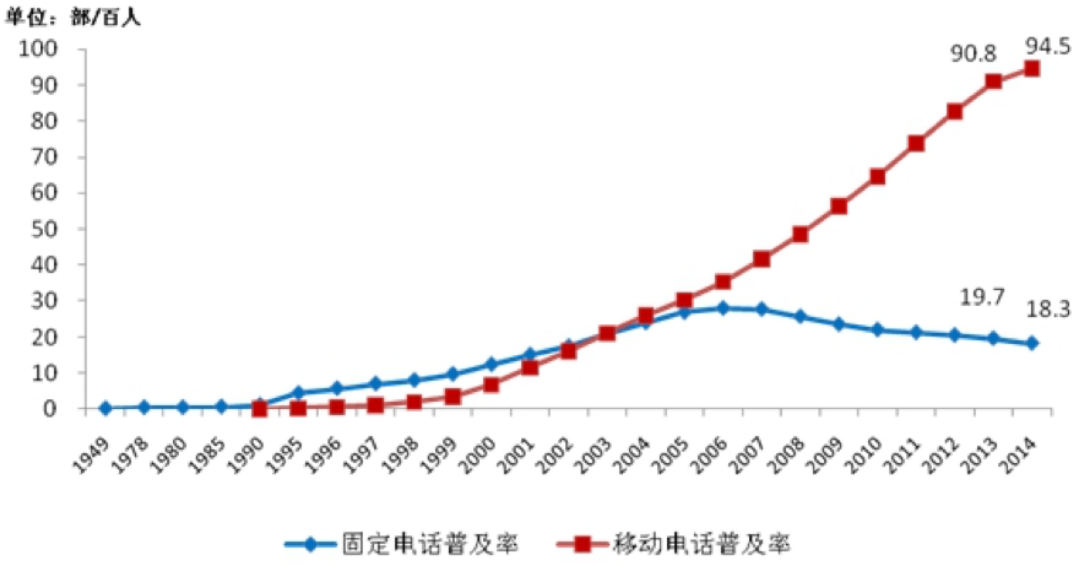


图 1.3 1949-2014 年固定电话、移动电话用户发展情况

用户的信息处于缺失状态。因此，如何利用这些信息来推断人与人之间的关系，即是当前社会的迫切需求，也是本文的挑战之一。

1.2 国内外研究现状

社交关系识别可以归结于关系预测等社交网络领域的经典研究问题。

1.3 本文主要内容

本文主要利用移动通话数据构建了一个移动社交网络，并对该网络中用户与用户之间的社交关系进行了探讨。其主要内容如下：

1) 问题定义及介绍

本文从当前研究移动社交网络的热点出发，定义了我们所要研究的问题，即关系识别在移动社交网络中的研究。除此之外，详细介绍了我们的移动社交网络的数据以及该网络的特点。

2) 移动社交网络特征

本文针对我们所研究的移动社交网络的特点，结合社会学、空间学等理论，提出了一系例具有时间、空间、网络结构和用户交互的特征，并对这些特征在真实数据集上进行验证了其有效性。

3) 关系识别模型

本文针对我们所研究的问题，在基于概率图等模型的基础上，提出了我们用于关

系识别的模型，并在真实数据集上测试了我们的算法，具有较高的识别准确率。

4) 结果分析及改进

本文针对模型的结果，详细的分析了各类因素对结果的影响，并针对这些结果分析了背后所存在的原因，并给出了一个高效并行的算法实现。

1.4 本文组织安排

本文共六章，总体的组织安排如下：

第一章为绪论，阐述了本文的研究背景以及研究意义，并介绍了当前国内外关系识别的研究现状。最后简单介绍了一下本文研究的主要内容。

第二章为基于社交网络的的研究与挑战，介绍了当前常用于该领域的几种模型。

第三章为针对移动社交网络所提出有特色的特征，介绍完了我们综合时间、空间、网络结构等方面所提出的特征。

第四章为我们针对所提出特征而建立的模型，该模型是基于概率图模型并针对我们的特征而构建的。

第五章为结果分析，针对我们模型的结果进行详细分析与并分析了对比实验。

第六章为总结与展望，主要总结了本文的研究工作与创新点，同时也指出了当前研究工作的不足，并提出了未来的研究方向。

2 相关理论与技术

移动互联网的迅速发展、社交网络的不断扩大、可研究数据的日益丰富以及机器学习、统计数学、数据挖掘等技术的引入，给社交网络这个领域带来了广泛广泛丰富的研究课题，如信息传播、动态网络演化，网络可视化、Top-K 节点挖掘、社群发现等等。而近些年来关于移动社交网络的研究主要集中在网络的空间性质，物理空间与社交网络的交互以及链接预测。关系识别是对现有网络或者某一个特定时期网络拓扑图中用户之间的关系判别，或者预测等等。从机器学习的角度来看，该问题可以看做是一个分类问题，判断网络中每条边的类型。而分类问题作为机器学习、社交网络中的一个基本问题，一直是该领域研究的热门之一。本文主要介绍移动社交网络的主要特点，并对当前用于该领域的模型及方法进行简要介绍。

2.1 社交网络的数据及研究特点

由于当前研究者掌握研究社交网络的数据各种各样，因此他们的研究也全然不同。有些小组里面的数据多对网络中用户属性的描述，如用户的年龄，性别等，则他们的研究主要在于对用户画像的识别等；另外一些小组的数据如有社交网络整体的变化等数据，则这些小组的研究重点主要在网络的演化等研究领域等。而为了充分利用我们所获取的数据，即社交网络中用户与用户之间的关系类别 (如家庭、朋友、同事等)，我们的研究重点放在了社交网络中的关系识别上。

从数据分析以及机器学习的角度来看，我们的研究可以转换为一个传统的分类问题 (分类出网络中不同的关系类别)，许多研究者也的确将这个问题看作分类问题[5, 6]，并且将研究重点放在对移动社交网络的内在特征与特点进行研究上，所以采用的方法大多是基于统计、时间序列的方法，或者传统的机器学习方法，如支持向量机 (Support Vector Machine)、决策树 (Decision Tree)、罗辑回归 (Logistic Regression) 等。这些分类方法大多假设数据分布之间是独立同分布的，即每个样本之间并不存在关系 (处理时间序列等方法可能会假设服从马尔科夫分布等等)，而对于类别的判断主要基于研究者对每个独立样本所提出的特征。而在实际世界中，特别是在移动社交网络拓扑图中，样本之间往往是具有一定的联系的，如 Web 网页数据、通话数据以及引用数据等等。在通话数据中，每个用户除了拥有自身独特的属性之外，其标签属性还与周围的通话对象有很大的联系。例如，某个用户的联系对象大多为二十岁左右的年轻人，从我们的经验可以基本

判断这个人的年龄也大致在二十岁左右。同样在引用网络里面，如果一篇论文的属性属于数据挖掘类，那么它所引用的文章很有可能属于这一大类。这一类数据中节点的标签不仅仅可以从自身的属性上进行推断，也可以从节点所在的拓扑图结构以及周围的信息来进行推断，因为节点与节点之间的信息关联并不是人为假设的，而是在现实世界中自然而然形成的，表示了拓扑图中每个节点之间的相关性与联系。传统方法如上文提到的支持向量机、决策树等，都集中于数据独立同分布，不会采集节点与节点之间的相关信息，这一信息的却是会对关系识别的准确率造成一定的影响。

由于社交网络中数据节点之间的相互关联性，在使用和研究解决社交网路问题的机器学习模型时也需要充分考虑其特性。网络中节点的关联性，往往体现在两点：不同节点之间存在不同的关系，总体的结构非常复杂，需要模型很好的处理这些依赖关系；节点的标签也往往具有关联性，即当我们推断某个 A 节点时，往往希望能从 A 节点周围的节点的标签分布并获取一定的信息，从而增加预测 A 节点标签的准确度。从这两点我们可以看出，我们的模型要么具有很强的联合推断能力，要么能在某些特定的假设情况下，能够忽略这些复杂的依赖关系，并且不会给模型分类的准确度带来影响。这两种模型的思路也代表了两种不同类的模型，在机器学习中，第一类模型常常被称为生成模型 (Generative Model)，解决问题思路常常从联合概率推断出发，而第二类模型常常被称为判别模型 (Discriminative Model)，解决问题思路从条件概率推断出发。

2.2 概率图模型

3 移动社交网络中的关系识别分析

关系识别是当前社交网络的重要研究课题之一。在一个社交网络中，人们会因为不同的关系而联系在一起，如家人、朋友、同事关系等等。明确社交网络中用户之间不同的关系类型，有利于其它领域的深入研究与发现。如在线或者移动广告营销中，如果知道用户家人、朋友的兴趣爱好以及其常购买的商品类型，那么就能更加准确的给该用户推荐相关的商品与广告，反之亦然，知道用户的喜好，也可以给其朋友、家人等推荐相应合适的商品。在协助公安侦破并抓捕犯罪嫌疑人时，如果能够掌握犯罪嫌疑人其家庭、朋友，则能更快协助相关部门侦破案件，有效的抓捕犯罪人员。由前面介绍可以得知，随着移动智能手机的大规模普及，移动通话数据的人群覆盖率已经接近 100%，具有相当的普适性。除此之外，移动运营商所提供了家庭套餐、集团套餐等营销套餐，如果研究者能够和移动运营商进行合作，则研究者能够利用从移动运营商中获得的关系数据，作为其训练数据。

从第二章可以得知，从机器学习的角度来看，关系识别实质是一个分类问题。基于目前的研究现状，已经有相当多的学者对此进行了研究。但大多数此类研究都是将关系拆分为简单的“信任与不信任”，“强关系与弱关系”，“友好与敌对”关系，并没有将关系具体到一个明确的网络当中去 (如具有家庭、同事、朋友的关系网)。还有部分研究对对关系分类赋予了特定的寓意，但这些研究主要有“指导 - 被指导 (Advisor-Advisee)”[7]、“讲授 (Teaching) - 指导 (Advisor) - 助教 (Teaching Assistant)”关系[8]，比较适用于有向关系，而并非特别适合我们所研究的家庭、同事和朋友关系集。另外一些研究则是基于特别的数据集，如恐怖分子网络数据集分布[9]，和我们所要进行研究的通话网络数据结构和性质相差太大，并且这些性质的研究，大多仅从社交层面上对关系进行阐释，而不能从模型的角度充分挖掘社交与空间地理位置之间的联系，而我们所要做的工作则需要从这两个角度同时进行考虑。

移动社交网络提供了非常丰富的信息，可以用来挖掘人们在真实日常生活中的社交关系。在本章，我们首先对基于通话数据中移动社交网络的关系识别问题进行论述和定义，然后将我们所研究的数据进行详细介绍。最后，我们针对从用户通话角度、地理位置同现两个角度进行出发，研究不同交互特征下同事、家庭、朋友关系之间的显著差异，并对特征进行相应的分析。我们用通话数据展示我们的发现。由于篇幅的限制，我们不展示在短信息中的发现，但两者的特征发现比较相近。

3.1 关系识别问题定义

很显然，社交网络是一个图模型，因此不同的问题的基本构成都可以用图 $G = (V, E, W)$ 来进行表示，其中网络图中的每个点 $v_i \in V$ 表示该网络中的用户，图中点与点的边 $Edge(v_i, v_j) \in E$ 表示用户 i 与用户 j 之间存在某种联系 (这种联系可以自己定义，如在我们的问题中即两人存在社交关系)，而 W 则表示了这种点与点之间的关系强度 (如在我们的问题中，则可以定量描述为两用户之间的通话频率与强度等)。

具体到我们的问题当中，我们让 $G = (V, E, X, Y)$ 代表无向移动社交网络，这里的 V 是 $|V| = N$ 数量的用户集合，而 $E \subset V \times V$ 是表示用户之间社交联系边的集合，每一条边 $e_i \in E$ 都有一个相应的社交关系 $y_i \in Y$ 与之对应，这里的 $Y \in$ {家庭关系, 同事关系, 朋友关系}。需要注意的是，这里的朋友关系定义为联系较为频繁的用户。$\mathbf{X}$ 是特征矩阵，$\boldsymbol{x_i}$ 代表了 $|\boldsymbol{x_i}|$ 维特征向量，为每条边 e_i 的特征。因此在解决最终问题，推断移动社交网络前，我们需要选取合适的特征，即 $\boldsymbol{x_i}$ 的值。

3.2 数据介绍

在本论文中所使用的数据集是从 2010 年 10 月 1 日到 2010 年 10 月 25 日采集的中国河南省某县级市的移动手机通话短信数据，包含了 30 万用户超过六千万 (67,630,000) 条的通话记录，三千万 (31,560,000) 条的短信记录，四百万 (4,420,000) 条的手机开关机纪录，一千二百万条的基站切换纪录。该县级市总共有 354 座基站，而且每一座基站都有相应的经度和纬度。其中通话、短信的格式如表 3.1，开机关机、基站切换的纪录格式如表 3.2.

除此之外，我们还有由移动运营商提供的家庭集团和同事工作集团的数据。为了更加精确、更加合理的预测用户之间的关系类别，我们移除了那些家庭集团和工作集团大小为 1 的孤立点，因为这些点不会对我们所分析的问题构成任何贡献 (我们研究的问题本身就是边的关系)。除去这些无用的用户之后，我们可以发现大多数的集团由两个或者三个构成，这类型的集团占了所有家庭和同事集团总数的 83%。并且我们从数据分布上可以发现，同时集团的大小大多小于 10 人。

表 3.1 短信/通话记录格式

主叫号码	被叫号码	通话时长	主叫基站	被叫基站
1597128XXXX	1565295XXXX	2010-10-20 18:12:34	60234	60183

表 3.2 事件纪录格式

事件发生时间	用户手机号码	时间类型	起始基站	终止基站
2011-10-20 10:10:13	135XXXXXXX	1	60284	74856

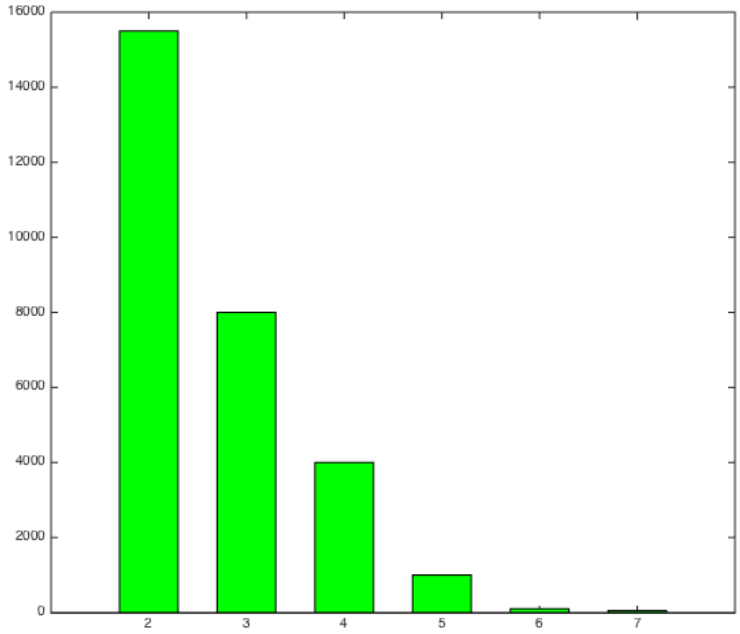


图 3.1 家庭集团大小分布

3.3 社交关系识别中的特征分析

本节主要从移动社交网络中的用户交互、时空交互、社交与时机地理空间交互等角度来分析影响社交关系识别的关键特征，充分利用移动社交网络的社交、空间结合等特性。本小节先从通话社交的角度进行分析，主要分析基本的通话特征对关系识别的影响，以及引入通话熵的概念，然后分析不同关系用户之间时间和通话强度的稳定性与差异性。然后从时空交互的角度来分析, 分析了用户时空同现性。随后，分析用户出现地理位置的寓意分析，了解用户同现所在实际位置的具体含义。最后，从经典的社交结构论扩充到地理空间的范围内，提出新的结构洞理论。

3.3.1 基本社交通话特征行为分析

从以前的研究中，我们可以知道，通过用户的通话记录，用户之间的关系 (朋友、非朋友) 能够很好的被识别出来[10]。但在那篇文章所用到的数据集都非常的小，并不能代表用户交互之间公共的特点，并不能直接认为围绕通话记录所得到的通话特征行为在我们的数据集上也有同样的效果。因此，我们在我们文章的数据集中重新考虑了用户之间的通话记录对关系识别的影响。

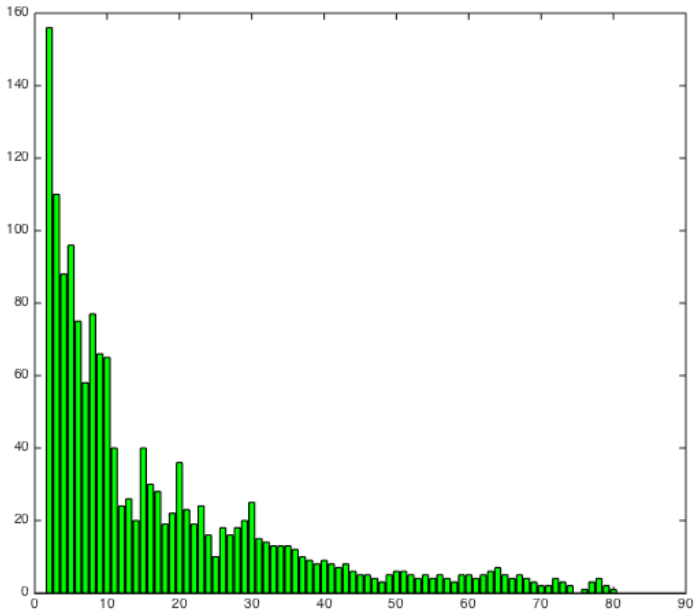


图 3.2 同事集团大小分布

在图 3.3 中，我们分析了不同的社交关系在一天之内，不同小时段的通话概率分布情况。这里我们定义忙时为每天的 10AM 到 12AM，6PM 到 9PM 期间。可以看到，无论哪儿种关系，在通话整体分布上呈现双峰分布，而曲线的最高点即为每天的忙时期间。在图中我们不难发现到，那些具有家庭关系的用户会比具有同事关系的用户拥有更高的交流通话频率。除此之外，我们也可以观察到，家人之间往往会选择在忙时进行通话。而具有同事关系的用户，往往选择的是在工作时间段 (每天的 9 点到 12 点，以及下午 14 点到 18 点) 内进行通话，而到了下班时间，我们会选择与具有家庭关系的用户进行通话，很有可能要通知家人是否要回去吃晚餐，大概几点会到家等话题。另外，朋友关系并不具有很明显的时间段分布，这反映了朋友关系往往是有事了才会选择通话，而并不具有每日特定时段的规律性，反映了朋友通话的随机性。从中我们可以知道，不同社交关系在不同时段之间的通话关系在一定程度上，反映了我们日常的行为，如从上述分析中我们可以知道人们往往会在下班之后选择和家人进行通话。这也充分反映了通话行为在识别不同社交关系的有效性。

从日常生活经验中，我们可以知道，人们在周末和工作日的通话行为往往是不一样的，为此，有必要分析不同社交关系在工作日和周末的通话行为。如图 3.4，图 (a) 为在工作日的不同社交关系通话频率分布差异，图 (b) 为在周末不同社交关系的通话频率分布差异。除了在上述图 3.3 的分析中所发现的特性，我们分析工作日和周末通话频率的共性和差异可以知道，无论是周末还是工作日，具有家庭关系的用户的通话频率都远远

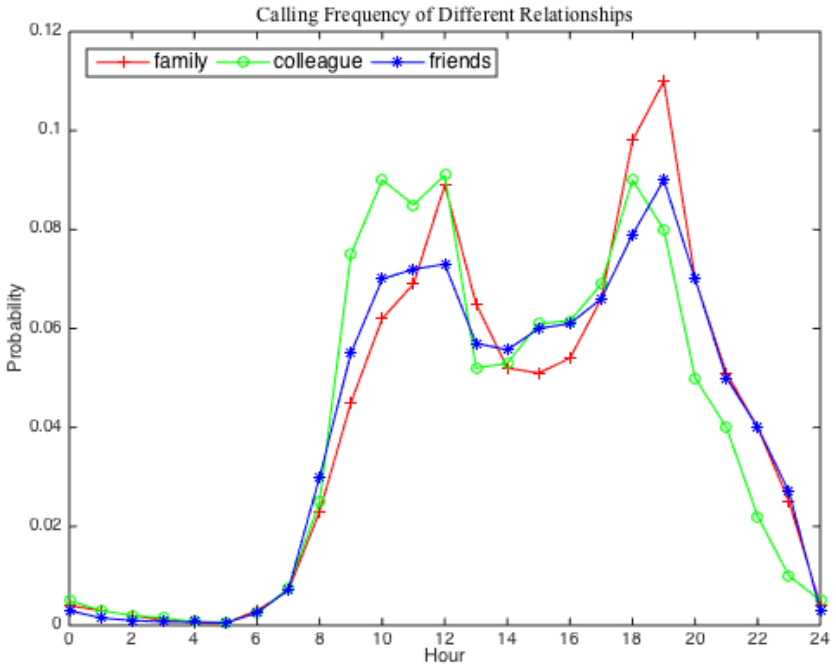
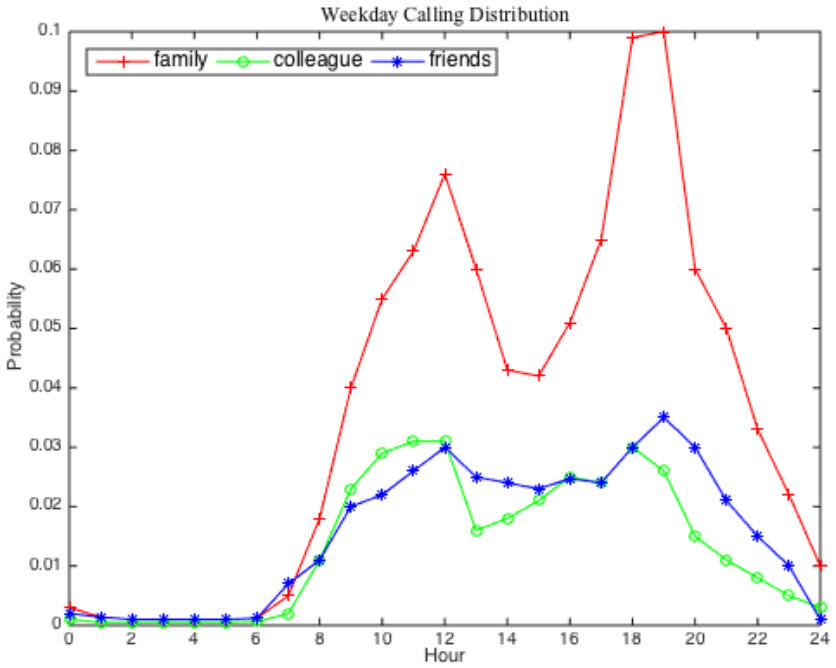


图 3.3 社交关系与不同小时段通话频率之间的联系

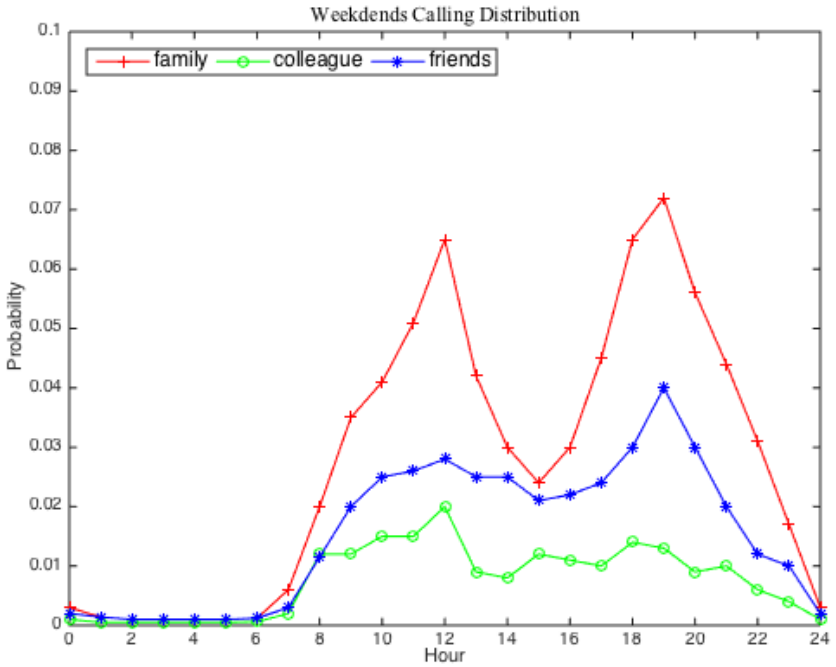
的超过具有同事或朋友关系的用户，这也充分体现了人们往往与自己的家人联系数量最多、最为紧密。同时，在周末，无论是具有家庭关系还是同事关系的用户，相对于工作日，两类关系的通话频率均有下降。这并不难发现，在周末家人大多数时间都是呆在一块，并不需要通过电话来进行联系。而同事之间在休息日若无工作，则很少通过电话来进行沟通。而朋友关系，正如在图 3.3 中所发现的，在周末和工作日的分布较为平稳，没有很大的区分，这也反映了朋友关系之间的稳定性。

3.3.2 通话熵分布分析

我们知道，我们会在不同的时间段内与不同的人通话联系。通常人们会在他们工作的时间与他们的同事通过手机联系。虽然说家人们主要集中在忙时通话，如下班之后。但是大多数的人们如果需要和他们的家人联系则不会等到某个特定的时间段内再联系，而是会直接打电话联系。为了定量描述这一特性，我们提出了通话熵的概念，可以用来描述用户之间通话的稳定性。熵的概念来源于信息领域，常常用来描述变量的随机性。由熵的概念而联想到的，我们定义了，通话熵的概念，用来描述通话在时间分布上的随机性。



(a) 工作日通话频率分布



(b) 周末通话频率分布

图 3.4 社交关系与不同小时段通话频率之间在工作日与周末的区别与联系

定义 3.1： 通话熵计算公式,

$$CallEntropy = -\sum_{i=1}^{T} p(x_i) \cdot \log p(x_i) \tag{3.1}$$

其中，$p(x_i)$ 代表了用户在第 i 小时段内通话的概率，T 的值通常被设置为 24，代表了一天里面有 24 小时。其计算结果代表了用户通话在一天内时间上的不确定性。如果通话熵的值比较小，那么用户的通话时间分布相对来说比较集中，这也意味着用户的通话时间分布相对比较稳定一些。反之，如果用户的通话熵越大，那么用户通话的时间越分散，即反映了该用户群体的通话时间越不确定。

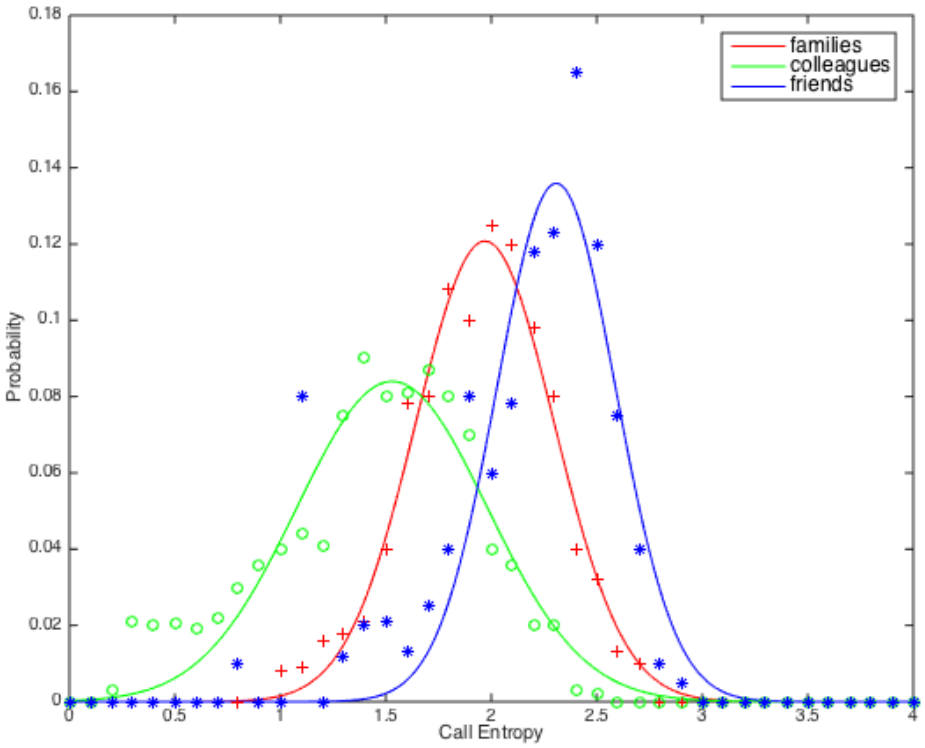


图 3.5　社交关系与通话熵之间的联系

在此概念的基础上，图 3.5 统计了不同社交关系用户的平均通话熵分布。我们可以发现基本上所有不同的社交关系的通话熵分布都遵循高斯分布。但是，具有朋友的社交关系的通话熵值是最大的，而具有家庭关系的用户的通话熵值小一些，而同事关系的通话熵值是最小的。这一现象揭示了同事之间往往会选择在特定的时间段 (从前面的分析中我们可以知道主要在工作时间段内) 进行通话，而家人和朋友则更多会随意一些。这也充分证明了朋友关系是作为一种非常重要的社交关系，因为人们在想起了事情或者想朋友了就会随时给他们打电话。这一区分度非常符合我们在实际生活中的通话习惯，以及处理不同社交关系的基本策略等等。

3.3.3 空间位置同现性分析

与传统的社交网络相比，移动社交网络不仅仅能够揭示用户之间基本通话行为所具有的特性，更能够从他们的地理位置分布等角度来挖掘新的特性。在我们的数据中，正如前面所介绍的，每一次用户的通话和短信，开关手机时，用户所在最近的基站都被记录了下来。由此，我们就能够得到用户每次呼叫的地理位置信息。基于这一其它社交网络并不具备的优势，我们进行分析不同社交关系与他们所具有的空间特征之间的联系时，具有得天独厚的条件。

在进行空间特征分析之前，我们需要分析我们的移动通话数据在多大程度上可以用来刻画用户在空间上的行为特征。为此，我们需要从数据集中所有带有地理位置信息的事件 (短信、通话、基站切换、开关机等等) 计算所有记录前后的时间差，以此确定用户出行行为规律。如图3.6，我们统计了用户的位置间隔分布，可以看出，绝大多数的用户记录间隔都在一个小时之内，这也表明我们的数据集记录比较频繁，可以用来研究用户的出行行为。

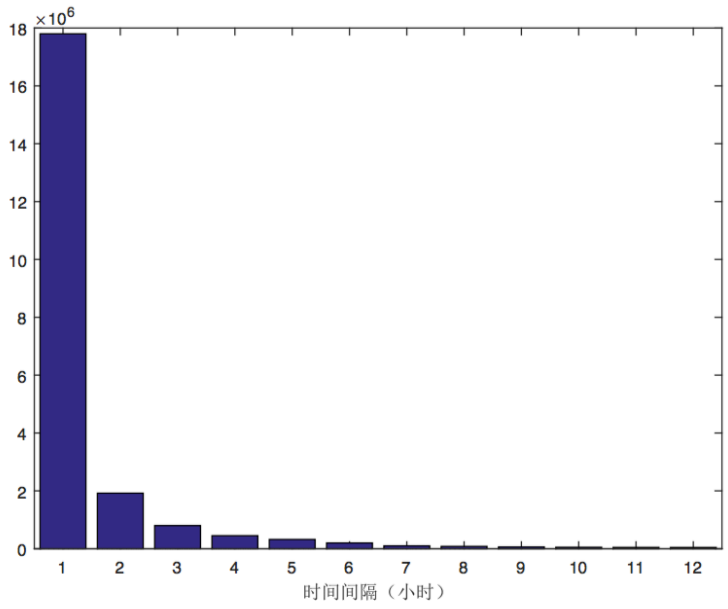


图 3.6 数据集中的用户位置间隔

接下来我们对用户的位置同现条件进行定义。这里的位置同现不同于字面上理解的两人在同一时刻出现在了同一地点，因这一类的数据在整个数据集上的分布较少，同时我们也无法确定同一地点的具体含义，因此我们需要重新定义同一时刻，即时间相邻的时间段，以及统一地点，即空间相邻的地理范围。在以前的研究中，以小时为时间粒度

进行聚合，算作时间相邻，算是一种比较有效的手段[6]。另外考虑空间相邻的定义，为了得到用户的地理位置，我们往往是考虑用户所使用的基站的位置，因此只要两个不同的用户使用同一个基站，就可以认为他们在空间上具有位置同现的特性。但是，在真实世界中，因为有的地方基站分布较为密集，即是两个人处于确切的同一位置，他们也有可能暴露给两个不同的基站，因此很有必要合并一些距离非常近的基站。因基站合并算法在当前研究中较为成熟[11]，我们采用现成的算法，不再进行研究，直接使用大多数研究中所使用的基于 *Voronio* 图的基站临近合并算法[11]。在进行这些工作之后，如果两个用户在一小时的时间间隔内处于同一基站下 (合并之后的基站)，那么我们可以认为，这两个用户时空位置同现。

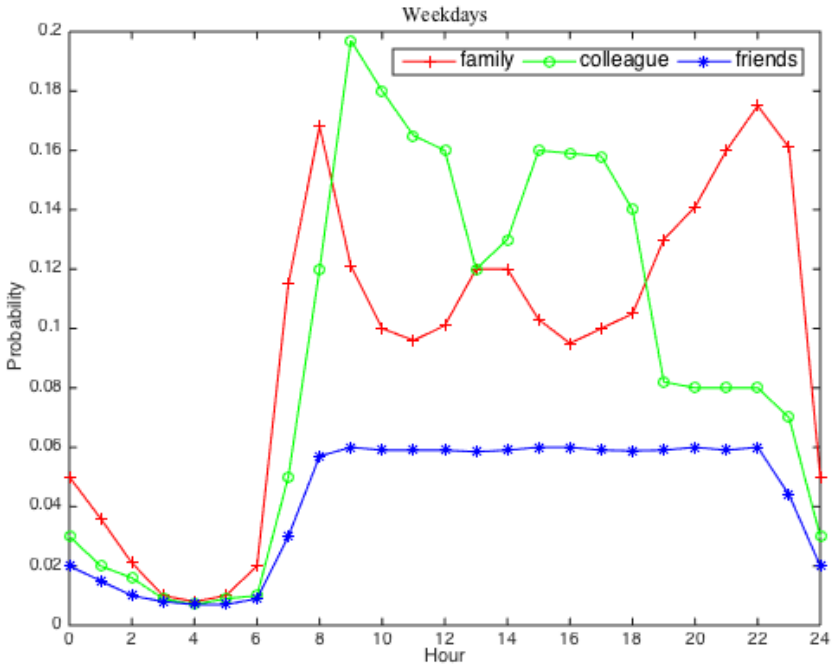
图3.7则显示了不同社交关系的用户在一天内同现的可能性分布，图3.7(a)为在工作日的同现特征分布，图3.7(b)为在周末的同现特征分布。图中横坐标为一天中的 24 小时，纵坐标为同现概率，这里我们定义同现概率。

定义 3.2： 同现概率定义，

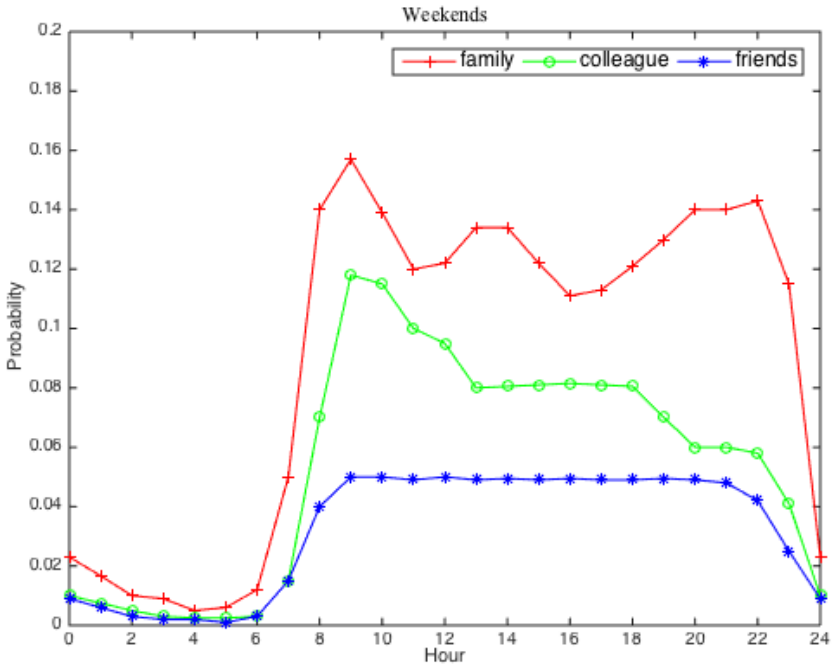
$$C^h(x,y) = \sum_{l \in L} p_x^h(l) \times p_y^h(l) \tag{3.2}$$

其中 L 为基站总集合，$p_x^h(l)$ 代表了用户 $\boldsymbol{x}$ 在时刻 h 内出现在地理位置 l 的概率与可能性。

从图中的曲线我们可以观察到，无论在工作日还是在周末，三种社交关系在午夜凌晨同现的概率都比较低，但在其它的时间段内，以及在工作日和周末都呈现一定程度上不同的同现特征。在工作日的时候，同事之间同现的可能性明显比其它两种关系的可能性要高。但是在工作日的中午，具有家庭关系的用户同现的概率反而增加了，而同事之间同现的概率出现了一个小的低谷状态。分析其中缘由，我们不难发现这个现象只有在中国小城市里面的传统家庭才会出现的。和许多国内外大城市不同的是，在中国中小城市里的人们往往会选择中午回家吃饭，并且进行短暂的休息。我们的数据集以及曲线很好地验证了这一现象。当到了工作日的晚上，家庭关系同现的概率突然呈现剧增状态，而同事之间同现的概率却急剧下降。造成这一现象的原因很简单，正是因为许多用户这个时候都已经下班回家了，不会再和同事呆在工作的场所，反而会和家人在一起。而在周末，家庭关系的用户同现的可能性最高，而朋友关系的用户同现可能性最低。这揭示了人们往往会在周末陪伴家人，呆在一起，而同事之间周末有时候会选择聚会等等活动。这些在曲线上所挖掘出来的信息，在现实生活中都得以体现，证明我们所提出来的同现特征，在我们的数据集上对关系识别有着很好的的效果，在一定程度上能对关系识别进行有效的区分。因此，我们可以进一步在同现特征的基础上进行扩展，挖掘出更多基于



(a) 工作日不同社交关系同现特征分布

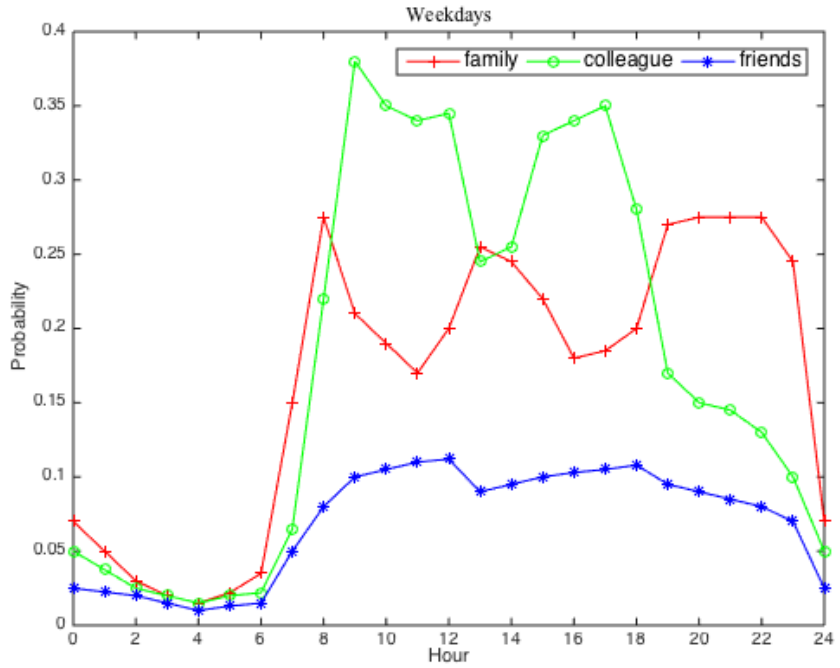


(b) 周末不同社交关系同现特征分布

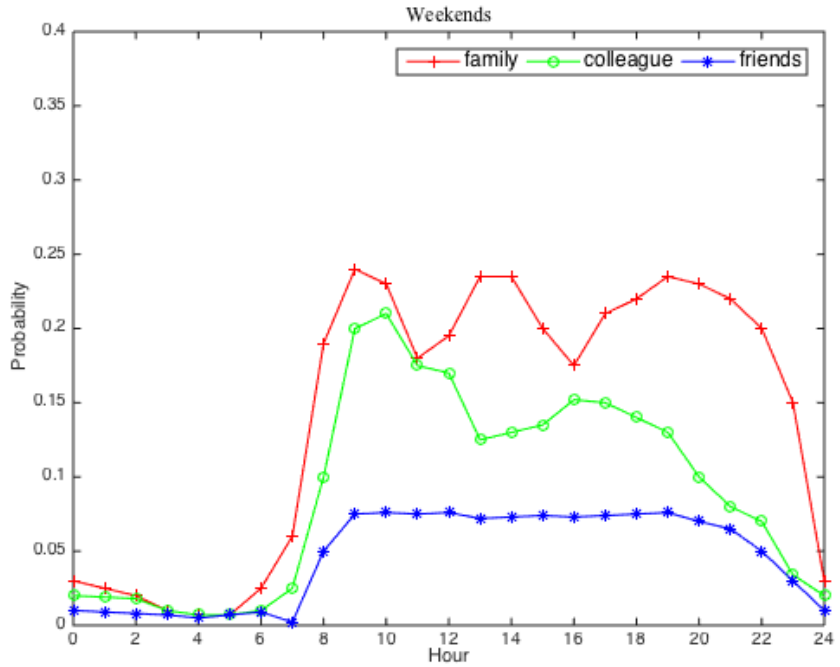
图 3.7 社交关系与位置同现特征在工作日与周末的区别与联系

此思路的同现特征。

从前面的分析中我们知道，从用户的总体轨迹的同现特征能够发现一定的规律。接



(a) 工作日不同社交关系在出现最频繁的基站同现特征分布



(b) 周末不同社交关系在出现最频繁的基站同现特征分布

图 3.8 社交关系与在最频繁出现的位置同现特征在工作日与周末的区别与联系

着，我们分析用户出现最频繁的基站所附近同现特征的规律。从实际生活中我们可以得知，如果一个用户有工作，那么他在白天出现最频繁的基站很有可能就是他工作的地点

附近。而晚上出现最频繁的基站则很有可能就是他家里面，所以单独选出着用户出现最频繁的基站有很大研究意义。由于时间和篇幅的限制，这里仅仅对用户出现最频繁的基站进行分析，而并非分为用户在白天和夜晚出现最频繁的基站进行研究。这里对一些概念来进行定义，定义 l_x^h 为用户 x 在 h 时刻出现频率最高的基站的 ID，则用户两个用户出现最频繁的基站同现的概率

定义 3.3： 最频繁基站同现概率定义，

$$P^h(x,y) = (l_x^h == l_y^h)?1:0 \tag{3.3}$$

如图3.8即为不同社交关系在最频繁出现的基站的概率统计。可以观察到，在工作日，具有同事关系的用户在白天同现率较高，而具有家庭关系的用户在晚上在最频繁基站具有较高的同现率。这也验证了我们之前的猜想，我们在白天一般出现最频繁的位置就是我们的工作地点，而晚上出现最频繁的地点则为家所在的地点。而在周末，具有家庭关系的用户同现率也比较高，这也验证了前面的结论：人们周末更愿意花时间陪伴家人，而且更愿意在家里面和家人相聚。

3.3.4 空间地理语意分析

前面的分析我们主要从同现的角度来分析用户的特征行为，得到了一些有用的结论。接下来的分析，我们希望能够深层次的探讨不同关系用户出行的规律与含义。我们现在知道，用户拨打和接收电话的记录记录了每个用户所在的基站的信息，而每个基站都有相应的经纬度信息。直观上看这些经纬度信息除了统计实际物理世界的地理位置分布之外，并无其他作用。但是由于移动互联网的发展，有了基于位置的服务。从这些基于位置的服务 (Location Based Service) 中，我们可以得到一些有实际意义的信息，进而通过分析用户的的出行位置，从而能得到用户的出行兴趣与爱好。

结合我们特有的国情，我们使用国内知名互联网地理位置服务提供商百度地图 (中国最大的 LBS 提供商) 获取基站周围 100m 范围内的 *POI*(Point of Interest, 地理热点) 信息，对于每一个输入的经纬度信息，百度地图 API 返回一个响应的数据如表3.3所示。

由表中数据信息我们可以知道，我们可以得到该城市中每个基站周围 100M 范围内主要的 POI 热点的具体信息。而我们所想研究的是用户的出行目的具体代表了什么如家人一起去了旅游景点、休闲娱乐等等含义。这样能进一步分析不同社交关系的出行行为特征。从表中可以看出，POI 类型即为类似的信息。从我们抓去的结果大致可以知道 POI 类型由政府机构、旅游景点、美食、休闲娱乐、购物、金融、运动健身、学校、行政

表 3.3 百度地图 API 返回 POI 信息

名称	类型	备注
Status	enum	结果状态值，成功为 0，否则-1
Location	Lat	纬度坐标
Location	Lon	经度坐标
Address	查询点的具体地址	
POIS(周围的 POI 信息群)	Addr	详细地址信息
	Source	数据来源
	Direction	和查询坐标的方向
	Distance	和查询点的距离
	name	POI 名称
	POINType	POI 类型，如学校、办公楼等
	POINT	POI 坐标 (x, y)
	tel	电话
	uid	百度标识 ID
	POSTAL	邮政编码

地标、住宅、医疗、丽人、道路、公司企业、汽车服务、酒店、运动健身、自然景色、文化传媒等 20 个分类。我们知道，一个地理位置往往会返回好多个 POI 点，而每个 POI 热点都含有一个 POI 类型，如运动健身等。因此，我们会发现基站及其附近区域往往包含几个不同的地理含义，即语义信息。例如，一个在商业街的基站，附近往往会包括美食、休闲娱乐、购物等区域。因此，如不对这些 POI 信息进行一定的处理，则这些信息无法得到有效利用。这里我们借鉴信息检索中常用的方法，用来衡量基站地下周围 POI 的语义信息对基站语义含义的贡献程度。转换为一个信息检索问题，那么就是由一篇文本所构成词的类型，来确定文章的类型。这里我们采用最简单有效的 $TF-IDF$(Term Frequency-Inverse Document Frequency) 方法。这种方法解决的问题正如前面所阐述的一样，常常用来判断某类词对一份文件或者语言库的重要程度。其基本思想即为一类词对文档的重要程度随着其在文档中出现的次数成正比，而与其在整个语料库中出现的频率成反比[12]。我们很容易知道，这一思想来自于假设对于文档最有意义的词语来自于那些在文档中出现的次数最多的词，而在整个语料库中出现频率较少的词语集合[13]。由文本挖掘的场景扩充到我们的问题中，TF 即为某种地理语义在特定某基站出现的的频率，IDF 即为所有基站总数与含有该语义的基站数比值的对数。譬如，在某特定基站下，“休闲娱乐”在此基站下出现了 10 次，那么该基站的 $TF=10$，并且这座城市里面包含“休闲

娱乐”的基站数为 100，而总基站数为 2000，那么该 IDF 值为 $\log \frac{2000}{100} = \log 20$。则“休闲娱乐”对于此基站的贡献程度 $TF-IDF$ 为 $10 \times \log 20$。下面我们具体定义 $TF-IDF$ 计算公式

定义 3.4： $TF-IDF$ 计算公式

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{3.4}$$

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|} \tag{3.5}$$

$$tf-idf = tf_{i,j} \times idf_i \tag{3.6}$$

我们选取对于识别不同社交关系有意义的语义，并对某些比较相似、比较相近的语义进行合并，之后我们计算每个基站下的每一类语义 $TF-IDF$ 值，可得到该基站在不同语义下的分布，同时为了计算的方便，我们将改分布进行归一化处理并与不同社交关系的用户同现的出行位置分布相乘, 可得到图3.9中各种不同社交关系出行的语义分布。

在图3.9中我们可以观察到，具有家庭关系的用户出现在住宅区、旅游景点等地方。这也验证了我们更多愿意和家人出去旅游，或者在家里陪伴家人。另外，具有同事关系的用户出现在商业街或者酒店，这也反映出了同事们最常上班的地方即在商业街，或者一起出差去了酒店等等。这一粗略的分析也反映了具有不同社交关系的用户往往会选择在不同的地点进行聚集，这也粗略地反映了不同社交关系的群体往往具有不同的社交行为，和群交关系爱好。

3.3.5 社交结构团分析

在以前的很多研究中[14, 15]，都或多或少从社会学理论角度进行分析社交网络, 社交平衡理论[16](Social Balance Theory) 则是被成功运用到社交网络中的经典知识之一。社交平衡理论阐释在实际社交网络中，随着不同用户之间的交流增多，一个初始的网络最后往往会发展成稳定的网络，而且最终往往会形成比较稳定的网络结构。图3.10则为社交平衡理论在我们的移动社交网络中的实际应用。图中，我们采用三元团来举例社交平衡理论，也因为三元团是社交平衡理论中最简单的结构之一。对于一个三元团所构成的闭环，每一条边都代表社交网络中用户的关系，如在我们的网络中，每条边的关系可以是家庭、同事或者朋友。如此一来，在实际社交网络中所有可能存在此三元团结构为 10

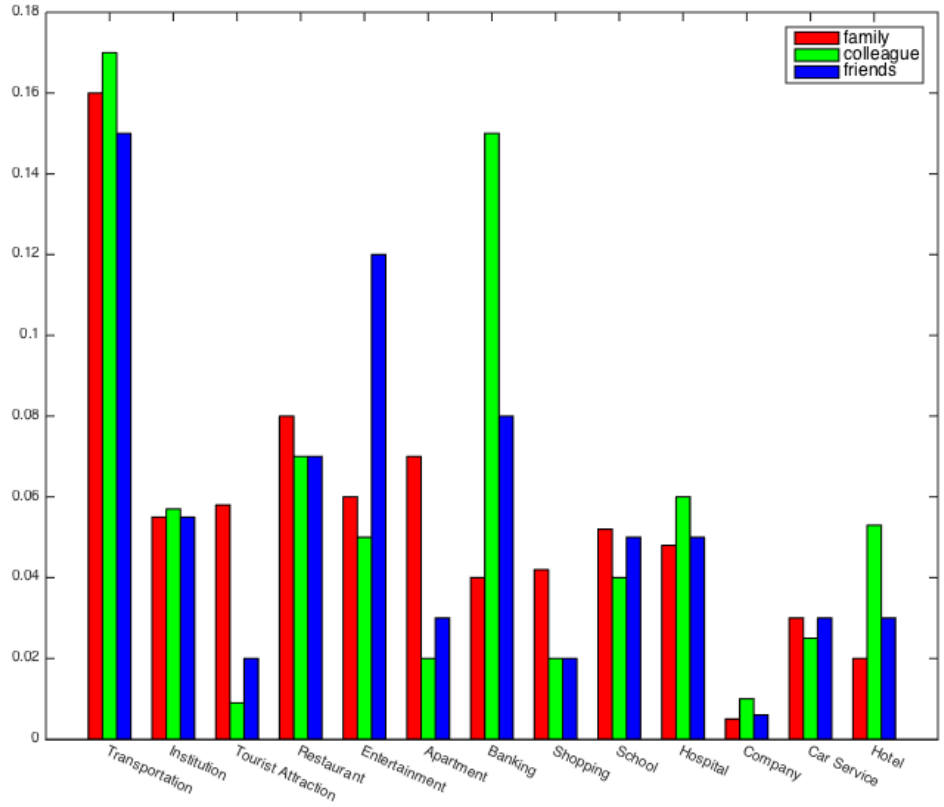


图 3.9　不同社交关系与位置同现语义之间的关系

中，正如图中所示。但是，社交平衡理论告诉我们，并不是所有的三元团都是稳定的，可保持的。随着时间的推移，网络拓扑图中不平衡的三元团会越来越少，而平衡的三元团会越来越多，理想情况下最终一个稳定的社交网络中只会存在极少数的不稳定三元团。

具体到我们的网络中实际分析，我们可以认为家庭关系、同事关系属于较强、比较稳定的关系类型，而朋友关系属于不稳定的关系类型 (类似对比到一个仅仅由信任-不信任网络中，信任属于较强、较为稳定的关系类型，而不信任属于较弱、不稳定的关系类型)。但同时考虑家庭关系、同事关系的不同点，我们认为家庭关系比同事关系的稳定性更强 (同事关系最终可能发展为家庭关系，也可能因为离职等因素而破坏)。以此类推，我们认为在所有的三元团里面，由三条边都是家庭关系、三条边都是同事关系、一条边是家庭关系和两条边是同事关系、一条边是家庭关系和两条边是朋友关系、一条边是同事关系和两条边是朋友关系所构成的三元结构团为稳定的三元团，而其他的都是不稳定的三元团。在此基础上，我们统计了哪些满足稳定条件的三元团在整个移动社交网络中所占的比例，看是否符合我们所提出的平衡理论。图清晰的展示了我们统计的结果。如图3.11，我们知道我们所认为具有平衡结构的三元团大约占了总的三元团 85% 的比例，

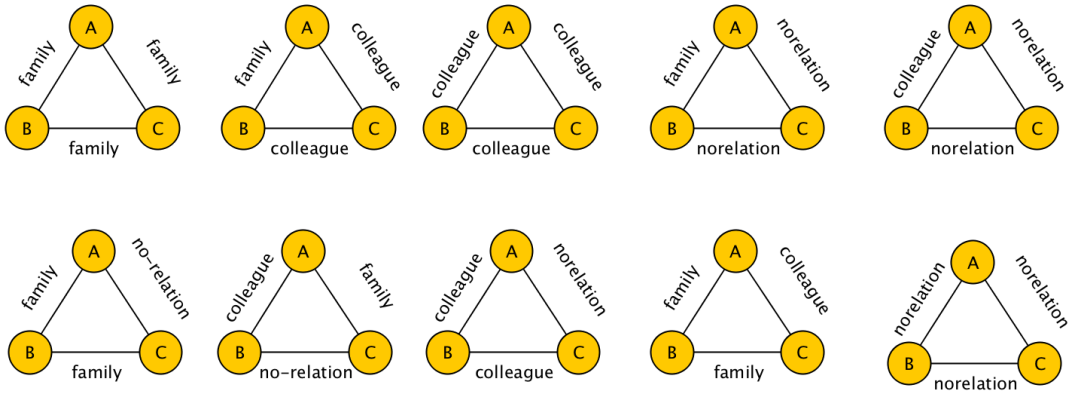


图 3.10　结构平衡理论举例

这也证明了在一个较为稳定的社交网络中，人们会逐渐形成较为稳定的结构，最终整个网络中所占的不稳定三元团总数相对来说较少。

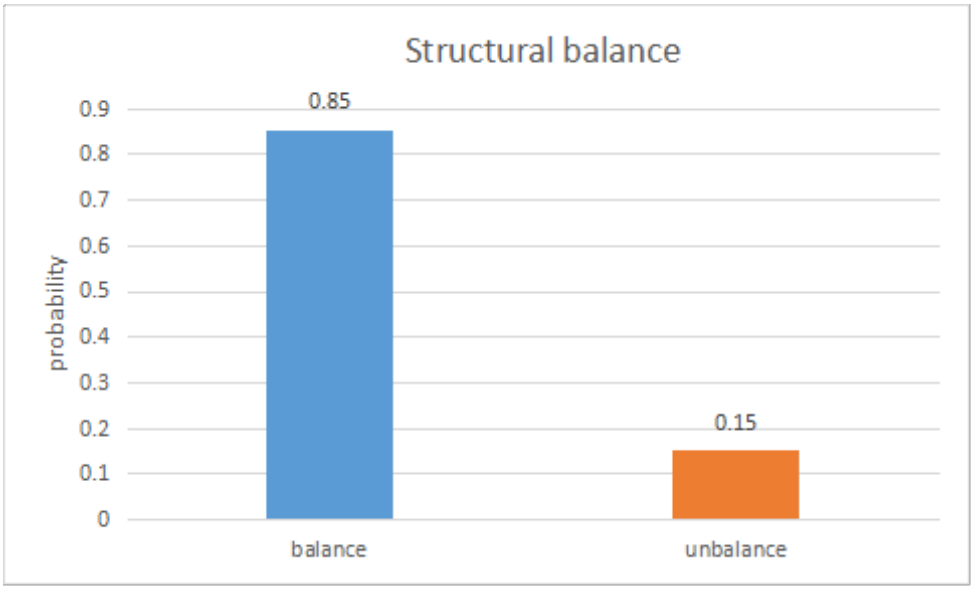


图 3.11　平衡与不平衡三元团结构在社交网络中所占的比例

由此延伸，我们将社交理论中的三元团进行扩展，放到真实世界中。我们分析这些在社交层面上具有的三元团，将三元团中的三个人同现的地理位置与时刻单独挑出来进行分析。由于篇幅的限制，我们这里不再展示结果，仅仅说一下思路。将具有三元团结构的用户提取出来，并分析他们同现的地方的地理语义，去其最常去的地方的地理语义，即为特征。这部分社交三元团、地理语义三元团与后面模型中的团特征相关。

4 关系识别模型

移动社交网络中的关系识别本身的特点让我们在选取基本模型的时候面对很多挑战：1) 移动社交网络中的数据结构以及依赖关系异常复杂，每个变量之间的关系以及结构往往是相互依赖、相互共存的，因此选取的模型需要由很强的表达能力来学习这些依赖，或者模型本身能够通过一定的手段忽略这些依赖关系；2）移动社交网络中的边的标签并不是相互独立而不受影响的，构成边的用户所在的边之间往往是相互联系，之间存在很强的关联性，如何充分利用这种关联性也是我们所需要解决的。

因此，结合第二章我们所介绍的主要几种概率图模型，我们认为我们问题中各类边和点的相互依赖关系难以做出一定的假设来构造马尔科夫网络，故最佳的方式是能够忽略变量之间的关联性，在此基础上充分利用边与团的关联性。因此，我们决定基础模型采用广义的条件虽机场模型，以此基础上构建我们的关系识别模型。由前面第二章的知识我们可以知道，条件虽机场是判别式模型，因此建模的基础事条件概率，这一优势能够有效的忽略变量之间的相互关联性。除此之外，条件虽机场提供了模版团，这正好能够让我们有效利用三元团等特征来提高关系识别的准确率。

模型本身对于各类特征的有效学习之外，我们还希望模型能够在一定程度上减轻标签分布不平衡问题。从我们的数据中统计可以知道，我们的移动社交网络中的关系分布是不均衡的，从整个大的网络分布来看，大约近 80% 的关系均为朋友关系，而大约各有 10% 左右为家庭关系和同事关系。从前面的知识我们可以知道，概率图模型是基于概率的模型，因此对于标签分布不平衡问题非常敏感。另外，对于我们的问题，这些在总体分布较少的关系往往在实际应用中扮演着至关重要的角色，因此不能说整个关系识别的准确率较高则认为我们的模型较好，而是要具体到这些分布较少的关系识别的准确率当中去。如果其识别的准确率较高，那么我们的模型则较为成功。否则，再高的准确率也说明不了什么问题。因此，在此问题中，我们也必须解决标签分布不平衡问题，并努力提高分布较少的关系类型的识别准确率。

基于前面关系识别特征的发现，我们提出了一个基于条件虽机场 (**Conditional Random Fields**) 的因子图模型 (**Factor Graph Model**)，我们称这个模型为 **BTFG** 模型，为 Balanced Triadic Factor Graph Model 的缩写。在这个模型中，我们从模型角度解决了数据挖掘中常见的数据标签不平衡问题，并且结合了前面的三元团等理论，将团的特性与概率图模型中的模板等概念结合，充分考虑这些团的特征，最大程度上提高关系识别的

准确率。

4.1 关系识别问题定义

问题 4.1： 关系识别预测：给定一个一部分被标注的网络 $G = (V, E^L, E^U, \boldsymbol{X})$，以及一些已经知道的变量 $y = 1, 2, 3$（*1* 代表家庭关系，*2* 代表同事关系，*3* 代表朋友关系），还有一些尚未知道的变量 $y =?$，我们的目标就是要预测这些未知的变量。在我们的模型中，抽象为一个数学问题，则需要学习以下的函数：

$$f : G = (V, E^L, E^U, , Y^L), \boldsymbol{X} \rightarrow Y^U$$

用来预测用户之间的关系类型。这里的 Y 即代表了边的关系类型。特别的，Y^U 代表了那些尚未知道关系类型的边 E^U。

在我们研究的工作中，我们将社交关系识别问题看作一个三元分类问题，即将关系类型分为三个群体：朋友，家庭和同事。

下面我们阐述具体模型的推导。

4.2 *BTFG* 模型框架

我们总体的目标是设计一个新颖的模型框架，这个模型框架能够利用和捕捉到上文我们所描述的所有特征，如各类时空特征、三元团特征等等。我们提出了一个能解决标签不平衡、采集三元特征的因子图模型。同时，我们为了解决大规模网络等问题，我们进一步提出了一个高效的分布式学习算法。

4.2.1 *Balanced Triadic Factor Graph*

上文中我们提到采用判别式模型，故我们采用条件概率对我们的问题进行建模。正如前面所说到的一样，采用条件概率进行建模能最大程度忽略不同变量之间的相互依赖关系。在给定用户之间关系的特征和输入移动社交网络的结构条件下，为了使得社交关系变量 Y 的可能性最大，从而我们可以本问题的目标函数。因子图模型提供了一套非常简单、有效的求解全局函数的计算框架，即现可计算不同本地函数的值，然后求各个不同本地函数的乘积的积分，即可得到全局函数。换句话说，即可讲全局函数分解为几个不同本地函数的乘积。因子图模型的这一性质使得最大化目标函数更佳简单，求解更加方便。因此，根据因子图的定义，我们有以下的目标函数：

$$P(Y|\mathbf{X},G)=\frac{P(\mathbf{X},G)P(Y)}{P(\mathbf{X},G)}\propto P(\mathbf{X}|Y)\times P(Y|G) \tag{4.1}$$

其中，$P(Y|G)$ 代表了给定网络结构下标签的概率，$P(\boldsymbol{X}|Y)$ 代表了由属性 $\boldsymbol{X}$ 在给定边的标签 Y 的条件下所贡献的概率。下面，我们假设在给定标签的条件下，每条边的属性之间是条件独立的，因此由属性所贡献的概率即可看作每个属性所贡献的概率相乘总的积，因此我们有

$$P(Y|X,G)\propto P(Y|G)\prod_i P(\mathbf{x}_i|y_i) \tag{4.2}$$

在我们的模型中，我们总共设计了三种不同的因子。第一个因子则是属性因子 (Attribute Factor) $f(y_i,x_i)$，用来采集两个用户之间的社交关系与其边的基本属性之间的联系。第二个因子是平衡因子 $g(y_i)$，用于从全局和模型的角度来解决标签分布不平衡问题，具体我们在下面进行阐述。第三个因子是三元结构因子 $h(y_c)$，用来采集社交关系与我们移动社交网络中三元结构特征之间的关系，这里的 y_c 代表了 y_i,y_j,y_k 所组成的集合，如果这三条边能够形成一个封闭的三元三角关系闭元结构。

因此，综上所述，结合我们前面所提到的因子图特性，我们可以进一步降目标函数的联合概率分解为以下的表达形式，

$$P(Y|G,X)=\prod_{e_i\in E}f(y_i,\mathbf{x}_i)\times\prod_{e_i\in E}g(y_i)\times\prod_{c_{ijk}\in G}h(y_c,\mathbf{x}_c) \tag{4.3}$$

从公式中我们可以看出，基于所有随机变量的联合概率分布能够进一步分解为所有局部因子的乘积集合。结合我们的实际社交关系识别问题，我们将三个因子进行的实例化。

4.2.2 模型中的三个因子

从前面的分析中，我们将全局函数分解为三个不同的因子，加下来我们将对这三个因子，结合我们的问题对他们进行实例化。

属性因子。我们使用这个因子来代表用户之间的社交关系 y_i 和它的移动网络的社交时空特征 $\boldsymbol{x}_i$ 之间的关系。更进一步，我们用以下的线性指数函数来实例化这个因子：

$$f(y_i,\mathbf{x}_i)=\frac{1}{Z_e}\exp(\alpha_i\cdot\mathbf{x}_i) \tag{4.4}$$

这里的 α 是我们所提出的模型中的的一个参数，而 Z_e 则为标准正则化常量。对于每一条边来说，α_i 是一个长度为 $|\boldsymbol{x}|$ 的向量。而这个参数的第 k 维代表了 x_{ik}(即 x_i 的第 k 个属性) 对于预测边的标签的贡献程度。比如说，x_{ik} 代表了两用户之间关系的同现概率，那么这一个因子可以采集不同的社交关系在移动社交网络中所具有的的同现特征。同理，属性因子能够采集其他我们在前面所提到的各种社交时空特征。这部分是所有的概率图模型均会具有的部分，即为我们模型学习的基础，其对社交关系的判别依赖于我们所提出特征的准确性与有效性。

平衡因子。接下来我们定义平衡因子。在定义平衡因子之前，我们现对现有的标签不平衡问题做一个调查回顾。

不平衡问题是数据挖掘领域一个比较经典的问题，曾经在 IJCAI 和 KDD 等数据挖掘国际定会上有过专门针对不平衡分类问题的研究主题和讨论。从当前解决不平衡问题的方案来看，主要从数据层面或者算法层面来考虑。

数据层面的方法的目的是通过对数据的采样，改变数据分布从而使原来不平衡的数据分布改变为平衡的数据分布的方法。数据采样的方法又可以分为上采样，下采样和混合采样三种方法。上采样方法又称为过采样方法的目的是增加少数类样本数目，从而改善不平衡分布。下采样方法又称为欠采样方法，其目的是通过减少多数类样本数目的方法使数据分布趋于平衡。两种方法各有优缺点，对于哪种方法更胜一筹也没有严格的证明，于是有研究者将两种方法结合起来提出了混合采样的方法。算法层面的方法主要考虑代价的学习模型，即代价敏感学习方法。此类方法常常对错误分类进行修正，达到对数据集训练重新分分布的目的。除了从代价敏感的角度，最近很多解决方法也 boosting 算法来考虑解决。

具体到我们的问题当中，我们希望我们的模型能自己学习数据集中标签不平衡的特性，而不希望破坏网络的整体结构 (采用采样的方法必须得劈坏整个网络总体的结构)。因此, 我们想能够从第二种方法来考虑, 即考虑代价敏感的学习方法, 并将此方法加入到我们的模型当中去。结合 Yale Song 等人[17] 应用在隐式条件随机场 (**Hidden Conditional Random Fields**) 的标签分布敏感参数，我们将其思想推广到我们的因子图模型当中来，这也算是代价敏感学习方法的一种。

这里我们定义平衡因子 $g(\boldsymbol{y}_i)$, $\boldsymbol{y}_i$ 代表了边 $e_i \in E$ 的社交关系类型。特别的，我们有、

$$g(y_i) = \frac{1}{Z_n} \exp(\beta_i \cdot \frac{\overline{N}}{N_{y_i}}) \tag{4.5}$$

其中我们 $\overline{N}$ 为所有与边 e_i 有公共顶点的边的总数，而 N_y 则为与边 e_i 有着同样标签(社交关系类别)、公共顶点的边的总数。这样以来，我们的平衡因子即相当于一个标签分布学习因子，能够有效的抑制标签不平衡问题所带来的负面影响。

三元结构因子。最后我们定义三元结构因子来采集社交关系与其社交平衡结构之间的关系。这里，我们有

$$h(y_c) = \frac{1}{Z_c} \exp(\sum_c \sum_k \gamma_c \cdot h_k(\mathbf{Y}_c)) \tag{4.6}$$

对于三元结构因子函数 $h_k(\mathbf{Y}_c))$，我们定义 10 个特征函数，包含 5 个平衡结构因子函数以及 5 个不平衡结构因子函数，如在图3.10所示。并且这 10 个函数都被定义为二元函数。更确切的说，如果一个三元结构满足某个二元函数，那么对应的结构因子函数的值为 1，否则为 0。这一定义参照了概率图模型[18]，如条件随机场中常用的函数定义方法，简单有效。

最终，我们结合公式4.4、4.5、4.6, 将它们带入到公式4.3中，并将目标函数定义为我们所提出模型的似然，可以得到

$$\begin{aligned} \vartheta(\alpha,\beta,\gamma) = &\sum_{e_i \in E} \alpha_i \cdot \mathbf{x}_i + \sum_{e_i \in E} \beta_i \cdot \frac{\overline{N}}{N_{y_i}} \\ &+ \sum_{c_{ijk} \in G} \sum_k \gamma_c \cdot h_k(\mathbf{Y}_c) - \log Z \end{aligned} \tag{4.7}$$

这里的 $Z = Z_e \cdot Z_n \cdot Z_c$ 全局标准化变量。

我们所提出的模型能够很好的吸收和消化我们前面所提出的时空以及三元结构特征，并且在训练模型的同时就能解决标签不平衡问题，而不需要破坏网络的整体结构。并且，基于条件概率的判别式模型往往会比基于联合概率的更优，因为我们无需假设某些变量以何种结构依赖于另外一些变量结构，因此我们仅仅需要整体上对所有的变量进行建模，并对社交网络的具体结构进行进一步的分解以适应我们的特征模型，使得模型能够充分利用和学习我们在第三章所提出来的所有类型的特征。这样一来，整个模型的具有一定的实际意义，也是比较容易理解的。

完成对模型的构建构建之后，后续则需要对模型进行求解，以及后续对社交关系的预测。

4.3 模型的学习与预测

现在我们需要来估计参数以及对未知社交关系的用户进行推断。这两个问题其实可以看作一个问题，都可以算作为概率图模型中的推断 (*Inference*) 的过程，如果我们把参数也看作变量的话，即利用一些已知的变量推断另外一些未知的变量。从概率论的角度来看，学习 *BTFG* 模型就是估计合适的一组参数 $\theta = \{\alpha, \beta, \gamma\}$，来最大化似然概率函数 $\vartheta(\alpha, \beta, \gamma)$。即

$$\theta^{\star} = \arg\max \vartheta(\theta) \tag{4.8}$$

4.3.1 参数学习

为最优化这个参数学习问题，我们采用一种梯度下降的方法 (也被称为 **Newtown-Raphson** 方法)。特别的，我们对每个参数进行求导分解得到，

$$\frac{\partial \vartheta(\theta)}{\partial \alpha} = \boldsymbol{E}[\sum_{e_i \in E} \alpha_i \cdot \boldsymbol{x}_i] - \boldsymbol{E}_{P_\alpha(Y|X)}[\sum_{e_i \in E} \alpha_i \cdot \boldsymbol{x}_i] \tag{4.9}$$

$$\frac{\partial \vartheta(\theta)}{\partial \beta} = \boldsymbol{E}[\sum_{e_i \in E} \beta_i \frac{\overline{N}}{N_{y_i}}] - \boldsymbol{E}_{P_\beta(Y|X,G)}[\sum_{e_i \in E} \beta_i \frac{\overline{N}}{N_{y_i}}] \tag{4.10}$$

$$\frac{\partial \vartheta(\theta)}{\partial \gamma} = \boldsymbol{E}[\sum_{c_{ijk} \in G} \sum_{k} \gamma_c h_k(\boldsymbol{Y}_c)] - \boldsymbol{E}_{P_\gamma(Y|X,G)}[\sum_{c_{ijk} \in G} \sum_{k} \gamma_c h_k(\boldsymbol{Y}_c)] \tag{4.11}$$

其中，$\boldsymbol{E}[\sum_{e_i \in E} \alpha_i \cdot \boldsymbol{x}_i]$ 是在给定的数据 Y 和 $\boldsymbol{X}$ 的条件下，属性因子函数求和的期望值。而 $\boldsymbol{E}_{P_\alpha(Y|X)}[\sum_{e_i \in E} \alpha_i \cdot \boldsymbol{x}_i]$ 则是给定的参数模型下属性因子函数的期望值。对于参数 α, β 是同样的道理，平衡因子函数与三元结构因子函数在两者条件下的期望值。

由于在我们模型中的图结构是任意的，那么很有可能含有闭环结构。这使得三个公式中的第二项非常难以计算，因为其时间复杂度是对数级别的。为此，们必须采用近似推断的方法来解决这个问题。这里我们采用的是 LBP(*Loopy Belief Propagation*)[19] 来计算边缘概率。因 LBP 容易实现，并且非常的高效，计算方便。

整个参数的学习过程可以被描述为一个迭代算法。在每一次的迭代过程中，都包含两步计算：第一步, 我们调用 LBP 三次, 来计算未知变量 $P_\alpha(Y|X), P_\theta(Y|X,G), P_\gamma(Y|X,G)$ 的边缘分布；第二步，我们使用公式4.12中的学习更新参数 η 来更新 α, β, γ。整个学习算法会等到迭代到一定的次数，或者更新参数的幅度过小的时候，会最终停止。

$$\theta_{new} = \theta_{old} + \eta \cdot \frac{\partial \vartheta(\theta)}{\partial \theta} \tag{4.12}$$

4.3.2 社交关系预测过程

有了上述的过程, 我们即可以算出估计的参数 θ, 那么我们的模型即可以确定了。由前面的介绍我们可以知道，其实求参数过程和预测过程基本上都可以算作是一个过程, 即利用已知变量推断未知变量, 因此我们可以使用刚才求参数类似的思想来预测我们移动社交网络中未知的社交关系 y_i =?，即找到一组社交关系值，使得下面的目标函数的似然最大,

$$Y^* = \arg\max \vartheta(Y|\mathbf{X}, G, \theta) \tag{4.13}$$

由上述可知，即可同样采用 *LBP* 来计算边缘概率，最终来估计参数。特别的，我们我们计算每种关系的边缘概率分布 $P_(y_i|\mathbf{x}_i, G)$ ，最终我们给社交关系赋予那些能够使得最大似然函数的标签。

至此，我们即可以估计出移动社交网络中未知的社交关系。

4.4 并行算法实现

呵呵

致谢

感谢国家

参 考 文 献

[1] 刘军. 社会网络分析导论: An introduction to social network analysis[M] : 社会科学文献出版社, 2004.

[2] TREVISAN T S. An Introduction to Social Network Data Analytics[M] : Springer, 2011.

[3] HEER J, BOYD D. Vizster: Visualizing online social networks[C]. Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on. 2005 : 32 – 39.

[4] 工业与信息化部. 2014 年通信运营业统计公报 [R]. 2015. http://www.miit.gov.cn/n11293472/n11293832/n11294132/n12858447/16414615.html.

[5] CHO E, MYERS S A, LESKOVEC J. Friendship and mobility: user movement in location-based social networks[C]. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. 2011 : 1082 – 1090.

[6] WANG D, PEDRESCHI D, SONG C, et al. Human mobility, social ties, and link prediction[C]. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. 2011 : 1100 – 1108.

[7] WANG C, HAN J, JIA Y, et al. Mining advisor-advisee relationships from research publication networks[C]. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. 2010 : 203 – 212.

[8] TASKAR B, WONG M-F, ABBEEL P, et al. Link prediction in relational data[C]. Advances in neural information processing systems. 2003 : None.

[9] ZHAO B, SEN P, GETOOR L. Entity and relationship labeling in affiliation networks[C]. ICML Workshop on Statistical Network Analysis. 2006.

[10] MIN J-K, WIESE J, HONG J I, et al. Mining smartphone data to classify life-facets of social relationships[C]. Proceedings of the 2013 conference on Computer supported cooperative work. 2013 : 285 – 294.

[11] AURENHAMMER F. Voronoi diagrams—a survey of a fundamental geometric data structure[J]. ACM Computing Surveys (CSUR), 1991, 23(3) : 345 – 405.

[12] SALTON G, MCGILL M J. Introduction to modern information retrieval[J], 1986.

[13] SALTON G, BUCKLEY C. Term-weighting approaches in automatic text retrieval[J]. Information processing & management, 1988, 24(5) : 513 – 523.

[14] TANG J, LOU T, KLEINBERG J. Inferring social ties across heterogenous networks[C]. Proceedings of the fifth ACM international conference on Web search and data mining. 2012 : 743 – 752.

[15] DONG Y, YANG Y, TANG J, et al. Inferring user demographics and social strategies in mobile social networks[C]. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014 : 15 – 24.

[16] EASLEY D, KLEINBERG J. Networks, crowds, and markets: Reasoning about a highly connected world[M] : Cambridge University Press, 2010.

[17] SONG Y, MORENCY L-P, DAVIS R W. Distribution-sensitive learning for imbalanced datasets[C]. Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on. 2013 : 1 – 6.

[18] SUTTON C, MCCALLUM A. An introduction to conditional random fields for relational learning[J]. Introduction to statistical relational learning, 2006 : 93 – 128.

[19] MURPHY K P, WEISS Y, JORDAN M I. Loopy belief propagation for approximate inference: An empirical study[C]. Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. 1999 : 467 – 475.

[20] KOTTWITZ S. LaTeX Beginner's Guide[M] : Packt Publishing, 2011. http://books.google.com.hk/books?id=rB1Cb62dVnUC.

[21] ACHARYA A, SETIA S. Availability and Utility of Idle Memory in Workstation Clusters[J]. ACM SIGMETRICS Performance Evaluation Review, 1999.

[22] ANDERSON E A, NEEFE J M. An Exploration of Network RAM[R] : UC Berkley, 1994.

[23] BODEN N, COHEN D, FELDERMAN R, et al. Myrinet: A Gigabit-per-Second Local Area Network[J]. IEEE Micro, 1995, 15(1) : 29 – 36.

[24] BOVET D P, CESATI M. Understanding the Linux Kernel[M]. 3rd : O'Reilly, 2005.

[25] CORBET J, RUBINI A, KROAH-HARTMAN G. Linux Device Drivers[M]. 3rd : O'Reilly, 2005.

[26] FEELEY M J, MORGAN W E, PIGHIN F H, et al. Implementing Global Memory Management in a Workstation Cluster[J]. ACM SIGOPS Operating Systems Review, 1995 : 201 – 212.

[27] FLOURIS M D, MARKATOS E P. The Network RamDisk: Using Remote Memory on Heterogeneous NOWs[J]. Cluster Computing, 1999, 2(4) : 281 – 293.

[28] FRANKLING M J, CAREY M J, LIVNY M. Globla memory management in client-server

DBMS architectures[C]. Proceeding of the 18th VLDB Conference. 1992.

[29] HAN J, ZHOU D, HE X, et al. I/O Profiling for Distributed IP Storage Systems[C]. Proceeding of The Second International Conference on Embedded Software and Systems. 2005.

[30] HE X, YANG Q, ZHANG M. A Caching Strategy to Improve iSCSI Performance[C]. Proceeding of Local Computer Networks. 2002.

[31] IFTODE L, LI K, PETERSEN K. Memory Servers for Multicomputers[C]. Proceeding of the IEEE Spring COMPCON 93. 1993 : 538 – 547.

[32] KOUSSIH S, A. ACHARYAM S S. Dodo:A User-level System for Exploiting Idle Memory in Workstation Clusters[C]. Proceeding of the Eighth IEEE International Symposium on High Performance Distributed Computing. 1999.

[33] LIANG S, NOTONHA R, PANDA D K. Swapping to Remote Memory over InfiniBand: An Approach using a High Performance Network Block Device[J]. IEEE Cluster Computing, 2005.

[34] LOVE R. Linux Kernel Development[M]. 2nd : Sams Publishing, 2005.

[35] MARKATOS E P, DRAMITIONS G. Implementation of a Reliable Remote Memory Pager[C]. Proceeding of the 1996 Usenix Technical Conference. 1996.

[36] NEWHALL T, FINNEY S, GANCHEVM K, et al. Nswap:A Network Swapping Module for Linux Clusters[C]. Proceeding of Euro-Par'03 International Conference on Parallel and Distributed Computing. 2003.

[37] OLESZKIEWICZ J, XIAO L, LIU Y. Parallel Network RAM: Effectively Utilizing Global Cluster memory for Large Data-Intensive Parallel Programs[C]. Proceeding of International Conference on Parallel Proceeding. 2004 : 577 – 592.

[38] PETRINI F, FRACHTENBERG E, HOISIE A, et al. Performance Evaluation of the Quadrics Interconnection Network[J]. Journal of Cluster Computing, 2003, 6(2) : 125 – 142.

[39] SUN H T, CHEN M, FAN J. A Scalable Dynamic Network Memory Service System[C]. Proceeding of High-Performance Computing in Asia-Pacific Region. 2005.

[40] TREVISAN T S, COSTAL V S, WHATELY L, et al. Distributed Shared Memory in Kernel Mode[C]. Proceeding of Computer Architecture and High Performance Computing. 2002.

[41] XIAO L, CHEN S, ZHANG X. Adaptive Memory Allocations in Clusters to Handle Unexpctedly Large Data-Intensive Jobs[J]. IEEE Transactions on Parallel and Distributed

Systems, 2004, 15(6): 577–592.