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)

基于移动社交网络的关系识别研究与实现

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 05 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：

基于移动社交网络的关系识别研究与实现

II、毕业设计（论文）使用的原始资料（数据）及设计技术要求：

1. 数据来源：某移动运营商提供的中国河南省某县级市移动通话数据

2. 开发语言：C/C++

3. 数据分析：Weka，LightSide

4. 改进基础：基础机器学习分类算法，基础概率图算法

III、毕业设计（论文）工作内容：

1. 数据清洗与整理

2. 特征分析与提取

3. 识别算法的设计

4. 识别算法的高效实现

5. 算法有效性验证

6. 模型评估与分析

IV、主要参考资料：

Koller, D. and Friedman, N., 2009. Probabilistic graphical models. MIT press.

周志华著. 机器学习, 北京: 清华大学出版社,2016.

Murphy, K.P., 2012. Machine learning: a probabilistic perspective. MIT press.

Easley, D. and Kleinberg, J., 2010. Networks, crowds, and markets:Reasoning about a highly connected world. Cambridge University Press.

李航著. 统计学习方法. 清华大学出版社, 2012.

Bishop CM. Pattern Recognition. Machine Learning. 2006.

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 05 月

基于移动社交网络的关系识别研究与实现

学　　生：冯溢濠
指导教师：吕云翔教授

摘　要

众所周知，不同的社交关系往往对人们有着不同的影响。然而，由于受到可靠数据的限制，许多研究都将他们的研究重点集中放在在推测朋友关系，或者在一个很小的数据集上进行研究。与此同时，随着移动位置服务 (Location Based Service) 的兴起，越来愈多的研究者开始利用一些新的附加信息：人们去过的地方。

在我们的工作中，我们研究如何为移动社交网络设计一个社交关系识别算法。我们使用由中国某移动运营商所提供的中国河南省某县级市的通话网络来进行研究工作，而这些通话网络数据附带有特别的关系 (家庭，同事和朋友关系)。我们先从总体上分析了时间和空间，还有各种社交因素对于关系识别的影响。于此同时，我们提出了一个可以定义人们日常经常去过的地方类别的方法 (如学校，购物中心等)，并在此基础上，我们找到了几个比较有效的时空特征，来预测移动网络中的关系类型。进一步，我们分析了用户最常去的地方，发现了不同的社交关系往往具有不同的行为特征。

基于这些发现。我们阐述了一个可以通过学习我们先有得移动网络特点的模型框架，用来区分不同社交关系类型。这个框架能够将我们之前所发现的特征糅合到一个因子图里面去，这样能够有效提高关系识别的准确性。我们的实验评估显示了时空因素与社交理论是如何显著提高了关系识别的准确度，以及我们的模型在关系识别任务上的有效性与准确性。这一系列的结果证实了我们的算法的优越性，为基于移动社交网络的关系识别研究开启了一扇新的大门。

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

It is well known that different types of social ties have essentially different influence on people. However, due to the limitation of reliable data, a bulk of research has focused on inferring friend relationships or has done on a small dataset. What's more, with the soaring adoption of location based social services it becomes possible to take advantage an additional source of information: the places people visit.

In this work, we study the problem of designing a social tie recognition system for mobile network. We used a dataset of a middle city in China provide by China Mobile, with specific relationships(families, colleagues and friends) in the network. We analyze the spatial and temporal influence on recognition and development of relationships. What's more, we proposed a method to define place categories(such as schools, malls) that users visit in their daily activities. From this perspective, we find several effective spatial features to infer social ties in the mobile network. Further, we study a special place that user visit most in their daily activities, home and workplace, which shows different charateristics on inferring different relationships.

Building on these findings, we describe a framework for classifying the type of social relationships by learning the charateristics of our existing mobile network. The framework incorparates what we find into a factor graph model, which effectively improives the accuracy of inferring the types of social relationships in the mobile network. Our evaluation shows how the inclusion of inforamtion about spatial factors and related user activities offer high social relationship recognition performance. These results open new directions for realworld relationship recognition system on location-based social network.

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

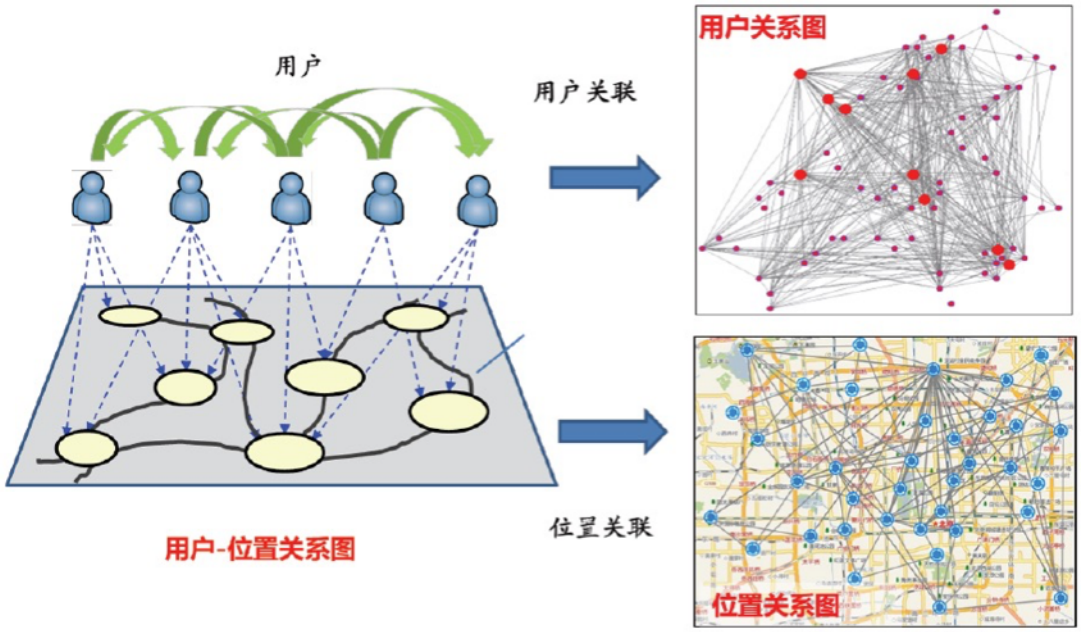


图 1.2　移动社交网络

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

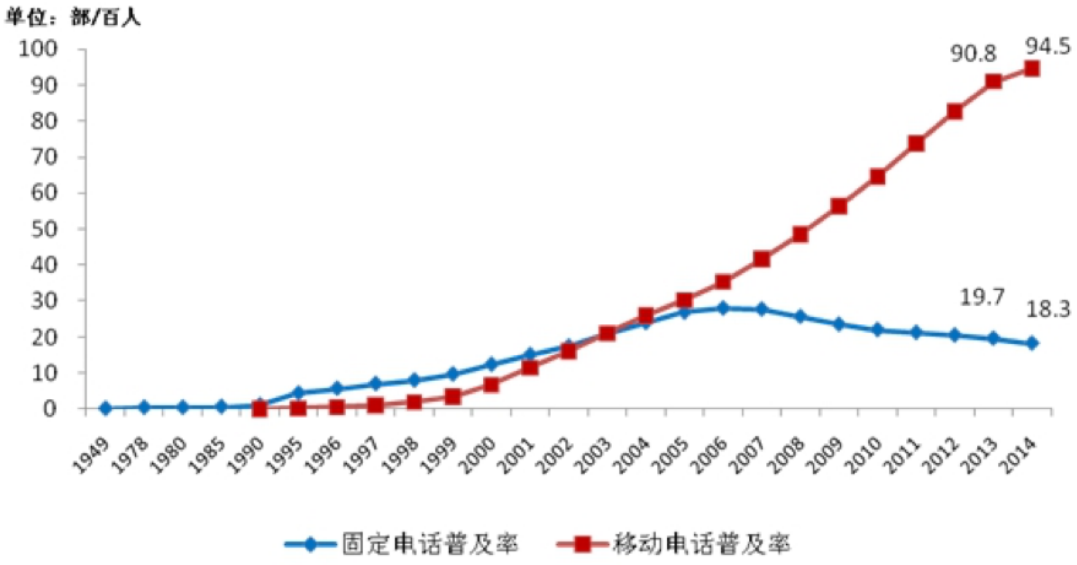


图 1.3 1949-2014 年固定电话、移动电话用户发展情况

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

目前，由于大型网络数据集的兴起，如移动手机网络与全球定位系统 (GPS) 的数据，使得越来越多的研究者能够分析基于位置的网络。极大程度的加深了我们对于人们移动通话行为与他们之间关系的理解程度。同时，不同形式的数据集决定了研究者研究的内容不同，因此移动社交网络中的研究内容千变万化。本小节结合当前主流研究热点，介绍当前国内外移动社交网络研究热点与现状。

目前移动社交网络中的热点研究主要分为关系链接预测、社交网络中空间地理性质研究以及网络中用户画像的识别[5]。下面我们分别从三个方面进行介绍。

首先是关系链接预测。本文所研究的关系识别工作与关系链接预测最为相关。在这些研究中，Liben-Nowell 和 Kleinberg[6] 采用了一种非监督的学习方法，来预测社交网络中新的关系出现的可能性，这一方法主要通过计算所有可能出现的关系候选所对应的概率，然后给这些候选关系进行一个排序，由此获得最有可能、即概率最高的新的关系，即为社交网络中可能出现的新的关系。这一方法在预测动态网络中新的关系的形成有一定的效果。Cranshow 等人[7] 则通过跟踪一些非常少数量的手机用户的通话行为，来预测他们在线上社交网络中的关系。虽然该研究中所用的数据量较少，但所采用的方法可进行迁移到大型数据集中进行验证。Eagle 等人则考虑通过利用手机用户人们之间的交

互信息，来预测人们之间的社交关系[8]。另外与该研究类似的是 Crandal 等人[7] 研究发现了从人们的行为中挖掘出来的时间和空间等因素能够用来帮忙预测人们之间的社交关系，这一研究也为我们考虑移动社交网络中时空特性提供了一定的参考。然而，我们的工作需要处理复杂社交关系，并且需要将网络的社交性和空间特性进行充分考虑。

当然了，也有相关的研究将研究重点全部集中于当前移动社交网络中的空间属性以及其周边扩展。一部分研究选择全部刨除网络的社交性质，将问题研究的重点都放在网络在真实世界的时空性质，探讨网络的时空性质对研究问题的具体影响。其中 Liben-Nowell 等人研究了两个独立个体之间的朋友关系是如何与他们之间的地理距离产生相互影响的[9–11]。更进一步，其他的研究人员发现了具有短距离的用户往往会发展成朋友关系，而长距离保持的朋友关系长久[12]。除此之外，研究者也有将移动网络数据来预测用户的地理空间行为等研究。如在网络中利用人们的移动行为规律来预测用户未来所在的地方[13–15]。这一系列最近的研究要么集中于一小堆的志愿者来进行实验，这样关系类型以及所使用的手机应用或者位置服务已经被定义好了，或者采用大型线上社交去如 Flickr[16] 等。尽管这些研究并没有提供一个大的社会群体范围内的社交联系或者用户日常行为规律，但是这些研究的的确确证明了物理实际空间与网络结构之间在许多方面的强烈联系。这些研究也为我们的关系识别提供了一些可以进行验证和改进的特征，是非常有价值的工作。

其他的一些人则将研究重点放在了研究用户的行为特征以及用户画像之间的预测。许多工作都做了关于人们出行移动特征与网络、物理特征之间的联系[17–19]。而至于用户画像的刻画，现有的工作有基于用户在线浏览记录或者它们搜索的行为[20, 21] 来刻画他们的画像。Leskovec 和 Horvitz 检测了 MSN 用户网络以及他们用户画像之间的内在关联[22]。Tang 等人则分析大型学术合作网络中研究者的档案，并对此建模来研究该网络中的研究者以及研究者之间的性质[23]。除此之外，研究者也有利用邮件网络[24]、LinkedIn[25] 网络等来验证用户的状态等等。诺基亚研究院在 2012 年组织的移动数据挑战赛中，使用了 200 个用户的交流信息来预测他们的年龄和性别[26, 27]。Kovanen 等人则利用了短期人流移动等性质来分析了动态社交网络中的画像同质性。Tang 等人也通过考虑不同用户之间画像的属性之间的关系 (如用户的年龄、性别等)，从而取得了比单独分析这些用户的年龄性别等等状态要好的结果[28]。

综合以上研究热点，我们当前研究的问题需要综合偏重地理空间特性以及社交性质的各类研究，并从中发现有效的各类特征，在我们的数据集上进行验证他们的有效性。除此之外，我们要在这些有效的特征基础上，提出一个能够有效综合这些时空特征、社交特征的模型，将他们的特性都得一充分利用，从而能够最大程度上预测移动社交网络

中的社交关系。除此之外，我们需要保证我们模型算法的时间复杂度不能太高，否则在实际应用中算法并不占优势。因此，如何针对我们的模型进行优化，如进行分布式、并行化或者近似计算等等，都是我们论文需要研究的主题。

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, 29]，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，如网页链接关联数据、移动通话网络数据以及学者合作网络数据等等。在移动通话数据网络中，每个手机用户除了具有自身独特的特征之外，其标签还与他通话人群有着很大的联系。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：网络中节点之间存在各种复杂的关系，导致刻画他们之间的结构非常困难，则相应的模型要有很好的处理这些复杂关系；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

因为社交网络本身就是一个图网络，因此在很多研究中，采用图模型对社交网络建模是非常有道理的。本节内容主要介绍在社交网络研究中运用比较广泛的概率图模型。这里我们主要介绍概率图模型框架中最上层的模型框架，而至于如何进一步详细求解概率图模型，即网络参数推断以及后面的学习算法等，因为涉及的预备知识以及数学概念较多，我们这里不进行深入讨论，仅仅给出该概念以及常见的几种方法。

因此，在本章节中，我们重点介绍概率图模型中最经典的三种模型，即贝叶斯网络，马尔科夫随机场，条件随机场。贝叶斯网络是有向图模型，常用于一些有向图，如 twitter 网络中粉丝和大 V 之间的关系建模。而马尔科夫随机场和条件随机场则是无向图模型，运用较广泛，常常可以用于表示无向关系的一些问题，比如我们的关系识别问题等。

2.2.1 贝叶斯网络

贝叶斯网络 (Bayesian network) 也被称为信¨念网 (Belief network)。它是利用有向无环图来刻画属性之间的依赖关系。并使用条件概率表来描述属性的联合概率分布。

具体来说一个贝叶斯网络是由其结构 G 和其参数部分 Θ 共同构成，即为 $B =< G, \Theta >$。其中网络结构 G 为一个有向无环图，每个结点之间对应一个属性，如果两个属性之间有直接依赖关系，则它们直接用一条有向的边连接起来；参数 Θ 定量地描述了这种依赖关系。比如假设属性 x_i，在 G 中的父节点集为 π_i，则 Θ 包含了每个属性的条件概率表 $\theta_{x_i|\pi_i} = P_B(x_i|\pi_i)$。

下面我门就具体的贝叶斯网络的结构进行介绍。

贝叶斯网络的结构有效地表达了属性间的条件独立性。给定父节点集，贝叶斯网络假设每个属性与它的非孩子属性独立。于是 $B =< G, \Theta >$ 将属性 $x_1, x_2, ..., x_d$ 的联合概率分布定义为

$$P_B(x_1, x_2, ..., x_d) = \prod_{i=1}^{d} P_B(x_i|\pi_i) = \prod_{i=1}^{d} \theta_{x_i|\pi_i} \tag{2.1}$$

图2.1给出了贝叶斯网络中典型的三个变量之间的依赖关系。前两个已经在式2.1中体现了。

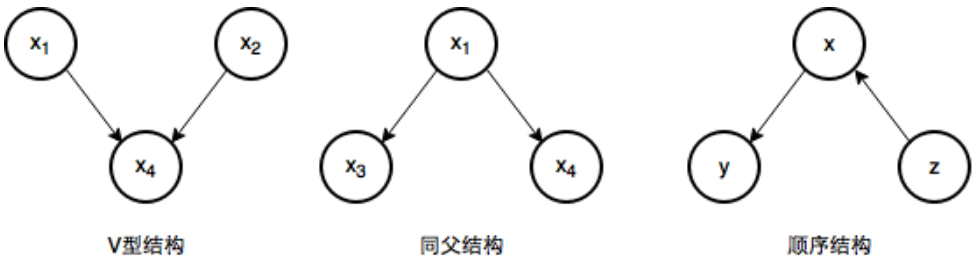


图 2.1 贝叶斯网络中变量的典型依赖关系

在同父结构中，给定父节点 x_1 的值，则 x_3 和 x_4 条件独立，在顺序结构的依赖关系中，给定 x 的值，则 y 与 z 相互独立。在 Y 型结构中，如果说给定节点 x_4 的值，那么 x_1x_2 必然不独立。但是需要我们注意的是，如果说 x_4 的值未知，那么在 V 型结构下 x_1x_2 是互相独立的。这种独立性称为边际独立性 (*marginal Independence*)，记作 $x_1 \perp x_2$。

实际上，一个变量的取值的确定与否，能对另外两个变量之间的独立性发生影响，并非在这几个结构中特有的。一般来说，某个值确定与否，会对其他变量的独立性产生非常大的影响。

为了分析有向图中所有的 V 型之间的独立性,我们可以利用有向分离 (D-Separation)。

首先，我们将有向图转为一个无向图:

- 找出有向图中所有的 V 型结构，在 V 型结构的两个父节点之间架上一条无向边；
- 将所有有向边改为无向边。

因此，由此产生的无向图称为道德图 (moral graph)，我们使得父节点相连的过程称为道德化 (*moralization*)。

基于 *moral graph* 我们能够非常直观地、迅速地找到变量之间的条件独立性，假定 *moral graph* 中有变量 x, y 以及变量集合 $\boldsymbol{z} = z_i$，如果变量 x 和 y 能够在图上被 z 分开，则从 *moral graph* 将变量集合 $\boldsymbol{z}$ 除去之后，x 和 y 分属两个连通分支，则称两者为有向分离，$x \perp y|\boldsymbol{z}$ 成立。

2.2.2 马尔可夫随机场

马尔科夫随机场 (**Markov Random Field**，简称 MRF) 是最经典的马尔科夫网络模型，是我们前面所说的无向图模型。模型图中每一个节点均表示相对应的某个变量，并且图中节点间的边表示节点所代表的两个变量的依赖关系。马尔科夫随机场模型中有称为势能函数 (potential function)，也称为因子 (factor) 函数，即为定义在变量集合上 main 的非负实函数，用于定义所研究问题所需要的概率分布函数。

在马尔可夫随机场中，多个变量的联合概率分布能够基于场中的图团的分解为多个因子 (factor) 的乘积，即为每一个因子仅仅和一个团结构相关。具体来说，对于变量集合 $\boldsymbol{X} = \{x_1, x_2, x_3..., x_n\}$，所有的团结构的结合为 C，和团 $Q \in C$ 对应的变量集合记为 $\boldsymbol{x}_Q$，则所定义的联合概率 $P(\boldsymbol{X})$ 定义为

$$P(\boldsymbol{X}) = \frac{1}{Z} \prod_{Q \in C} \psi_Q(\boldsymbol{X}_Q) \tag{2.2}$$

$$Z = \sum_{\boldsymbol{x}} \prod_{Q \in C} \psi_Q(\boldsymbol{X}_Q) \tag{2.3}$$

其中 ψ_Q 为和团 Q 一起对应的势能函数，用于对团 Q 中的各种变量的关系进行建模，Z 为规范化因子，以确保 $P(\boldsymbol{X})$ 被正确定义的概率。在实际运用的时候，计算 Z 的值通常比较困难，但幸运的是，许多方法都可以通过近似计算来获取 Z 的取值。

显然易见，当网络中变量的数量越来越多的时候，则团里面的数目将会变得非常多(如两个相连的变量都会变成团结构)，这就说明式2.2会有非常多的乘积项，这显然会给模型带来非常多的计算量。并且，如果团结构 Q 不是极大团的话，则它一定会被一个极大团 $Q^\star$ 所包含。既有 $\boldsymbol{X}_Q \subseteq \boldsymbol{X}^\star$；这也说明变量 X_Q 之间的关系不仅仅在势能函数 ψ_Q

中进行了计算，也在势能函数 $\psi_{Q^\star}$ 中进行了计算。由此我们可以将联合概率用极大团来重新进行定义。这里我们假设所有的极大团构成的集合为 $C^\star$，则我们前面定义的联合概率可以重新改写为

$$P(\boldsymbol{X}) = \frac{1}{Z^\star} \prod_{Q \in C^\star} \psi_Q(\boldsymbol{X}_Q) \tag{2.4}$$

$$Z^\star = \sum_{\boldsymbol{x}} \prod_{Q \in C^\star} \psi_Q(\boldsymbol{X}_Q) \tag{2.5}$$

这里的 $Z^\star$ 为规范因子。

下面我们对马尔科夫随机场的独立性进行说明

- 全局马尔科夫性 (*Global Markov Property*)：给定两个变量集的分离集，则这两个变量集条件独立。
- 局部马尔科夫性 (Local Markov property)：给定某个变量的近邻变量，或者称为该变量的马尔科夫毯 (Markov blanket)，则该变量独立于其他的变量。用形式化语言来描述即为，让 V 为网络中的结点集合，$n(v)$ 为结点 v 在网络图中的近邻节点，$n^\star(v) = n(v) \cup \{v\}$，有 $\boldsymbol{x}_v \perp \boldsymbol{x}_{V \setminus n^\star(v)} \mid \boldsymbol{x}_{n(v)}$
- 成对的马尔科夫性 (*pairwise Markov property*)：给定所有其他变量，两个非邻近变量条件独立。形式化语言来说，即令图中的节点集合和边的结合分别为 V 和 E，对于网络图中的两个节点 u 和 v，如果 $< u, v > \notin E$，则 $\boldsymbol{x}_u \perp \boldsymbol{x}_v \mid X_{V \setminus <u,v>}$。

除此之外，我们最后来对马尔科夫随机场的势能函数进行具体说明。很明显，势能函数 $\psi_Q(\boldsymbol{x}_Q)$ 是定量的对变量集合 X_Q 中的变量之间的关系进行说明，即该函数应该为非负函数，并且在所喜好的变量取值有着较大的函数值。

而为了满足势能函数的非负性质，我们常常使用指数函数来定义势能函数，即为

$$\psi_Q(\boldsymbol{X}_Q) = \exp^{-H_Q(\boldsymbol{x}_Q)} \tag{2.6}$$

$H_Q(\boldsymbol{x}_Q)$ 函数是一个定义在变量 $\boldsymbol{x}_Q$ 上面的实数函数，常见的形势可以为

$$H_Q(\boldsymbol{x}_Q) = \sum_{u,v \in Q, u \neq v} \alpha_{uv} x_u x_v + \sum_{v \in Q} \beta_v x_v \tag{2.7}$$

其中 α_{uv}、β_v 为参数。上面式子中第二项仅仅考虑单个节点，而第一项则考虑每对

节点之间的关系。

2.2.3 条件随机场

条件随机场模型 (Conditional Random Fields，简称 CRF)[30] 是一种判别式的，且为无向图模型。前面第一张的研究中我们提到过生成式模型以及判别式模型。判别式模型为对条件概率分布进行建模，而生成式模型则是直接对联合概率分布进行建模。马尔科夫条件随机场以及贝叶斯网络均是生成式模型，而条件随机场为判别式模型。

与马尔科夫随机场不同的是，条件随机场试图在给定的观察值的条件下，对多个变量的条件概率进行建模。具体来说，让 $x = \{x_1, x_2, ..., x_n\}$ 为观察序列，$\boldsymbol{y} = \{y_1, y_2, ..., y_n\}$ 为其与之对应的标记序列，那么条件随机场的目标首先即为构建基于条件概率分布的模型 $P(\boldsymbol{y} \mid \boldsymbol{x})$。值得说明的是，这里的标记变量集合 $\boldsymbol{y}$ 可以是结构性的变量，即有可能其分量之间本身就具有一定的相关性。例如在 NLP 领域当中的词类别标注任务当中，如果说可观测的变量为语句 (即为单词序列)，那么标记相应的词类别序列，即 $\boldsymbol{y}$ 具有线性结构。

下面我们对条件随机场模型进行定义。

让 $G =< V, E >$ 表示图中的节点与标记变量 $\boldsymbol{y}$ 中的元素对应的无向网络图。y_i 表示与节点 i 对应的标记变量，$n(i)$ 表示节点 i 的近邻节点集合，如果图 G 中每个节点都满足马尔科夫性质，即

$$P(y_i \mid \boldsymbol{x}, \boldsymbol{y}_{V\setminus\{i\}}) = P(y_i \mid \boldsymbol{x}, \boldsymbol{y}_{n(i)}) \tag{2.8}$$

则我们称 $(\boldsymbol{y}, \boldsymbol{x})$ 构成一个条件随机场。

理论上来说，图 G 可以拥有任何结构，只需要它满足图中变量之间的条件独立性关系即可。但是，在现实世界应用中，特别是前面的标记序列建模的时候，最常用的仍然是我们前面所说到的链式结构，即链式条件随机场 (chain-structured CRF)。下面我们仅仅对该形式的条件随机场进行具体分析，而任意图结构的条件随机场会在我们建模的时候得以体现，所以这里我们不进行介绍。

和我们前面介绍的马尔科夫随机场定义联合概率的方法相似，这里条件随机场定义使用势能函数以及图中的团结构来定义条件概率 $P(\boldsymbol{y} \mid \boldsymbol{x})$。在图2.2所显的链式条件随机场主要有两种已知标记变量的团结构，即单个标记变量 $\{y_i\}$ 以及近邻的标记变量 $\{y_{i-1}, y_i\}$。选择合适的势能函数，则可以得到如式2.2的条件概率形式。在条件随机场里面，我们通过使用指数势能函数，并且引入特征函数的概念 (feature function)，条件概率

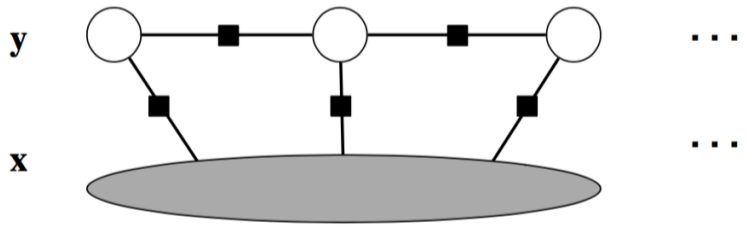


图 2.2 链式条件随机场模型实例

即可以重新改写为

$$P(\boldsymbol{y} \mid \boldsymbol{x}) = \frac{1}{Z} \exp\left(\sum_{j} \sum_{i=1}^{n-1} \lambda_j t_j(y_{i+1}, y_j, \boldsymbol{x}, i) + \sum_{k} \sum_{i=1}^{n} \mu_k s_k(y_i, \boldsymbol{x}, i) \right) \tag{2.9}$$

其中 $t_j(y_{i+1}, y_j, \boldsymbol{x}, i)$ 为定义在观察序列变量上面的邻近标记位置上的转移特征函数 (transition feature function), 这些函数常常用来描述相邻标记变量之间的定量关系以及观察序列对这些观察序列之间的影响。$s_k(y_i, \boldsymbol{x}, i)$ 为观察序列的标记位置 i 的状态特征函数 (status feature function)，这个是用来定量描述观察变量对于标记变量的影响。其中 λ_j 和 μ_k 为待求解参数，Z 为规范化因子，用于确保式子2.9的概率准确性。

对比2.9以及式2.2，我们可以知道，条件随机场以及马尔科夫随机场都用了团结构上面的势能函数来定义和求解概率，二者在形式上面并没有太大的区分，但是条件随机场使用的条件概率进行建模，而马尔科夫随机场使用的联合概率进行建模。

2.2.4 网络中的学习与推断

对于概率图模型中所定义的联合概率，我们能够用目标变量的边际分布 (marginal distribution) 或者用某些已知变量为条件来进行条件分布推断。在前面的介绍中我们已经知道了很多的条件概率建模，例如在隐式马尔可夫模型中估计观察序列在给定参数下的条件概率分布。这里的边际分布指的是对无关的变量进行求和或者进行积分之后而得到结果。例如，在马尔科夫网络中，网络中的变量的联合分布可以被表示成为极大团的势能函数的乘积，则在给定参数 Θ 的条件下，求解某个变量 x 的分布，即成为了对于联合概率分布中其他的无关变量进行积分的过程，也被称为边际化 (marginalization)。

对于概率图模型，求解的过程实质就是确定具体的未知参数分布，这也被称为参数估计或者参数学习问题。换个角度，如果将未知参数看作未知变量，那么参数估计的过程则和未知变量推断的过程非常相似。因此，两个问题都可以归结为概率图模型中的推

断问题。

概率图模型中的推断方法大致可以分为两大类[31, 32]，第一类为精确推断，即通过数学推导的方法精确计算出目标变量的边际分布的准确值。但是，一般来说，此类方法的计算复杂度为指数级别，该类方法的使用范围有限。第二类则为近似推断的方法，即能够在较低的时间复杂度内实现对原来问题的近似求解。此类方法在现实研究中运用的更加常见，该类方法也主要分为两大类：即第一类为采样 (sampling)，采用随机采样的方法对结果进行近似求解；另外一类是变分推断 (variational inference)，即近似进行推断求解。在我们的问题中，我们为了算法效率的高效性，采用的是信念传播的近似求解，置信信念传播 (Loopy Belief Propagation) 进行参数的学习与推断。

3 移动社交网络中的关系识别分析

移动社交网络中的关系识别是当前主流研究课题之一。在一个特定的网络中，用户之间往往存在各种的社交关系，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，由于移动智能手机的普及率日益上升，使用移动手机 (当然主要是智能手机) 的人群覆盖率将近 100%, 具有一定的普适性。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，关系识别问题为分类 (Classification)。由前面的第二章节介绍可以知道，当前已经有相当多的学者在进行该方面的研究工作。然而，但绝大多数此类的研究工作都是将关系拆分为简单的“信任与不信任”, “朋友与非朋友”关系，并没有将这种抽象的概念放在一个具有明确关系语义的网络当中去 (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“辅导 - 被辅导”[33]、“讲授 - 指导 - 助教”关系[34]，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[35], 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。具体的数据内容格式在表 3.1。开机和关机、通话基站的切换数据格式如表 3.2.

除此之外，我们的数据当中有由移动运营商提供的家庭 (Family Clique) 和同事集团 (Colleague Clique) 的具体关系。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/通话数据格式

主拨号码	接听号码	通话时长	主拨用户所在基站	接听用户所在基站
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

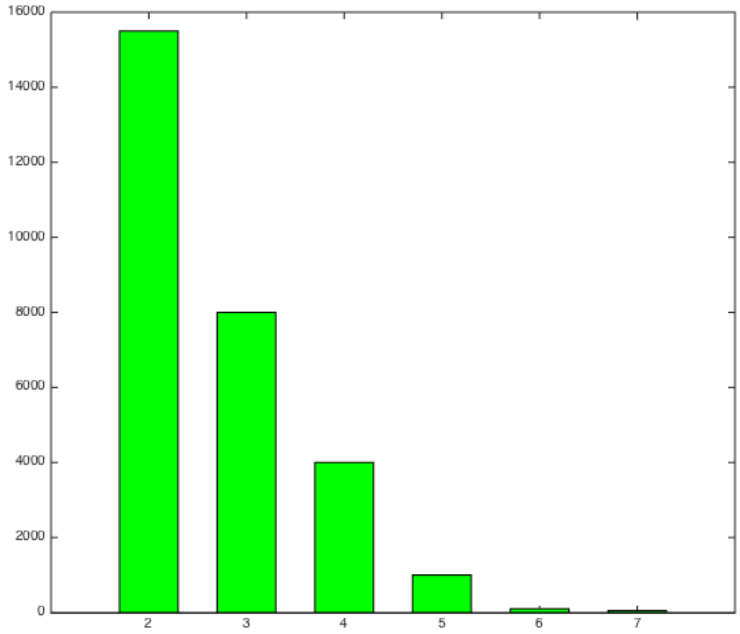


图 3.1 家庭集团大小分布

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[36]。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。

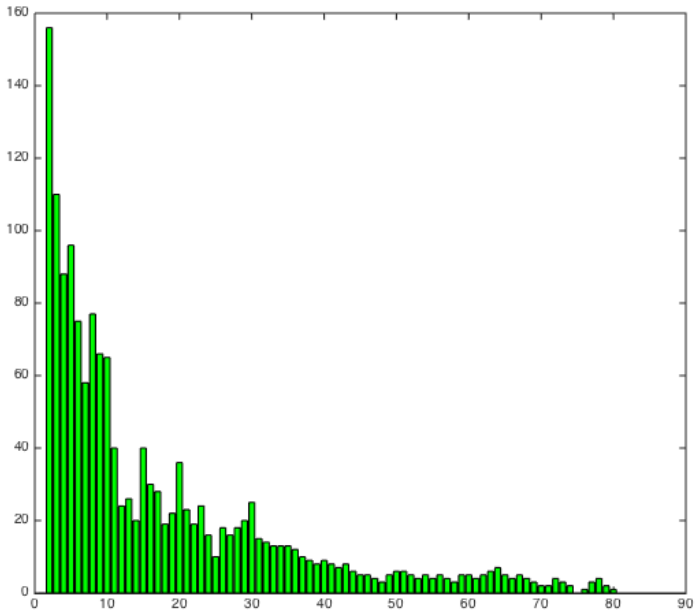


图 3.2 同事集团大小分布

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，不管是星期天还是周一到周五，具有家庭关系用户们之间的呼叫

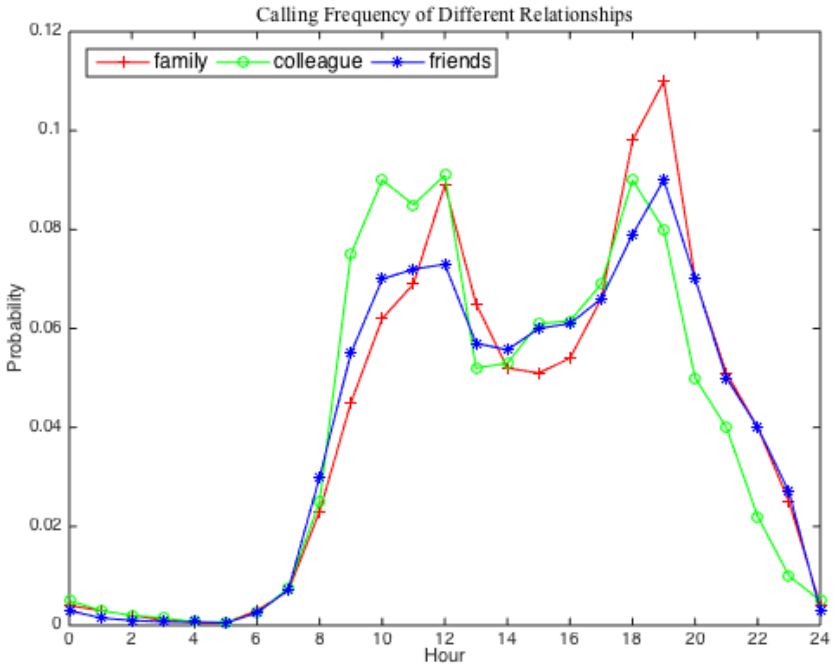
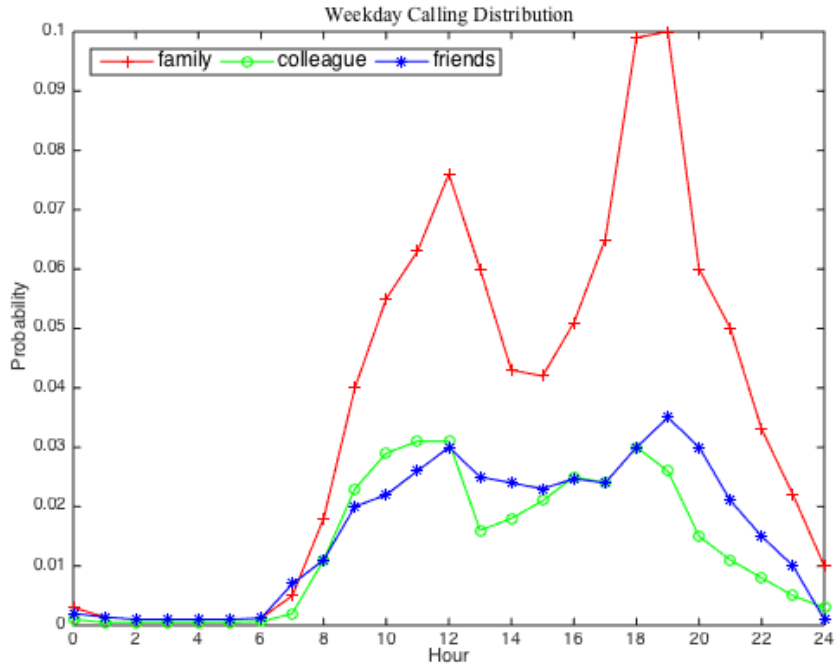


图 3.3 社交关系与不同小时段通话频率之间的联系

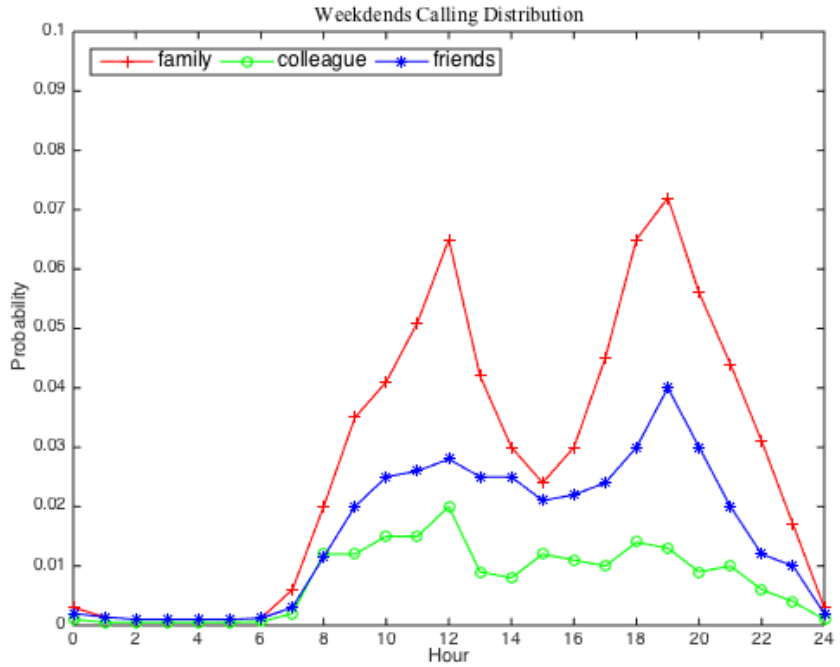
频率都远远的超过具有同事或朋友关系的用户，这也充分体现了人们往往与自己的家人联系频繁。在周末，具有家庭、同事两种关系的用户，相对于工作日，两类关系的通话频率均有下降。这并不难发现，在周末家人大多数时间都是呆在一块，并不需要通过电话来进行联系。而同事之间在休息日若无工作，则很少通过电话来进行沟通。而朋友关系，正如在图 3.3 中所发现的，在周末和工作日的分布较为平稳，没有很大的区分，这也反映了朋友关系之间的稳定性。

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，$p(x_i)$ 为了用户第 i 小时段内通话的概率，T 的值设置为 24，代表一天有 24 小时。如果计算结果 Call Entropy 的值小，那么说明移动通话用户的通话时间分布相对集中，这也意味着他的的通话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。

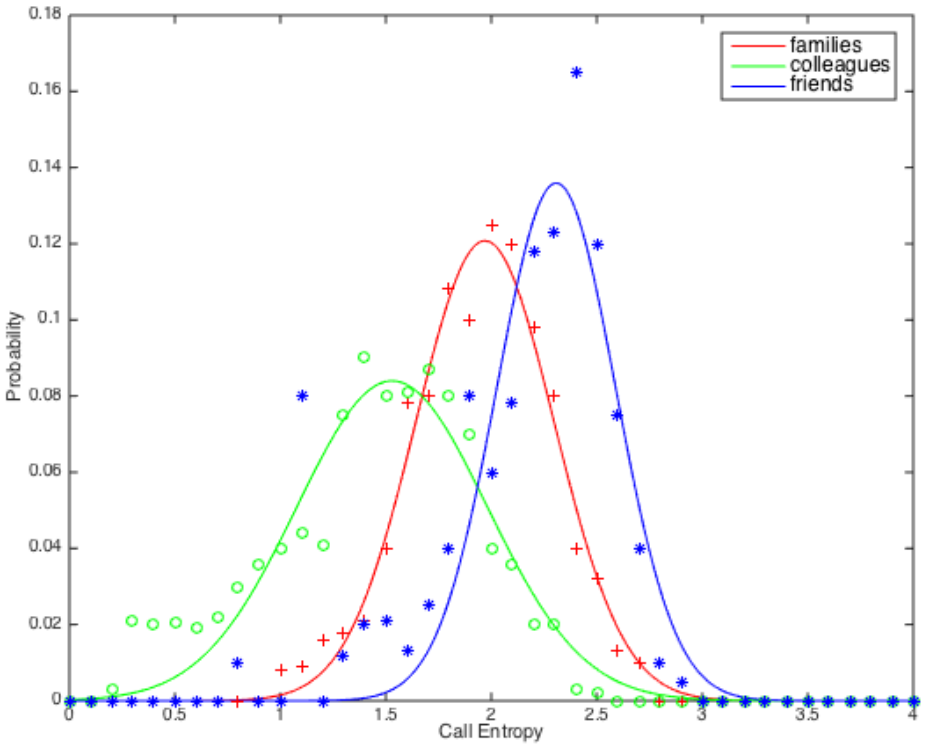


图 3.5　社交关系与通话熵之间的联系

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。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。

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。

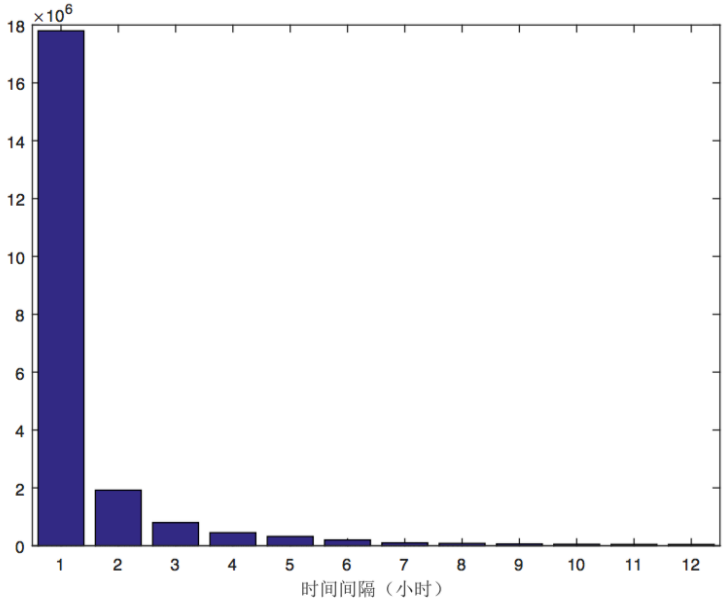
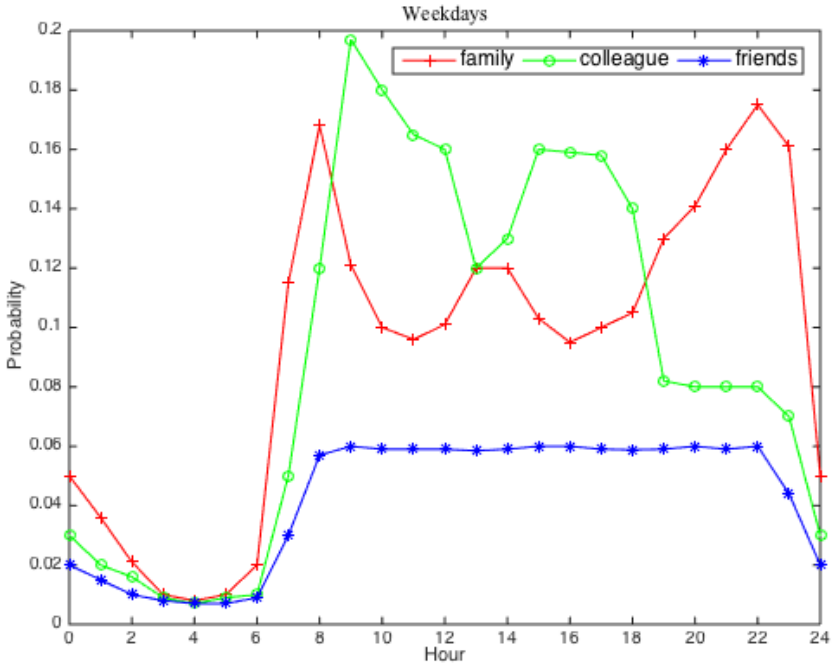


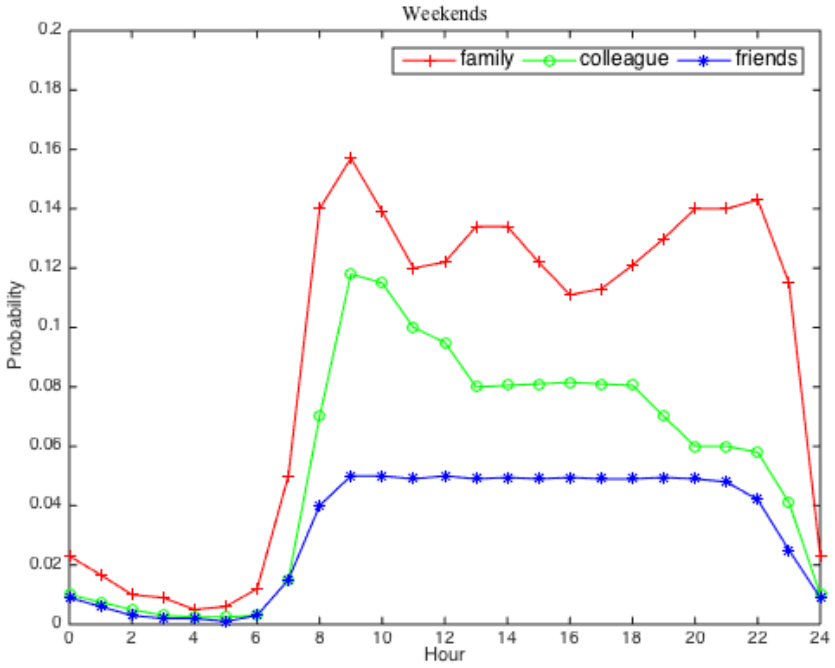
图 3.6　数据集中的用户位置间隔

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[5]。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[37]，我们采用当前已经由的基站合成算法，不再进行研究，用为 *Voronio* 图的基站临近合并算法[37]。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。



(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

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。

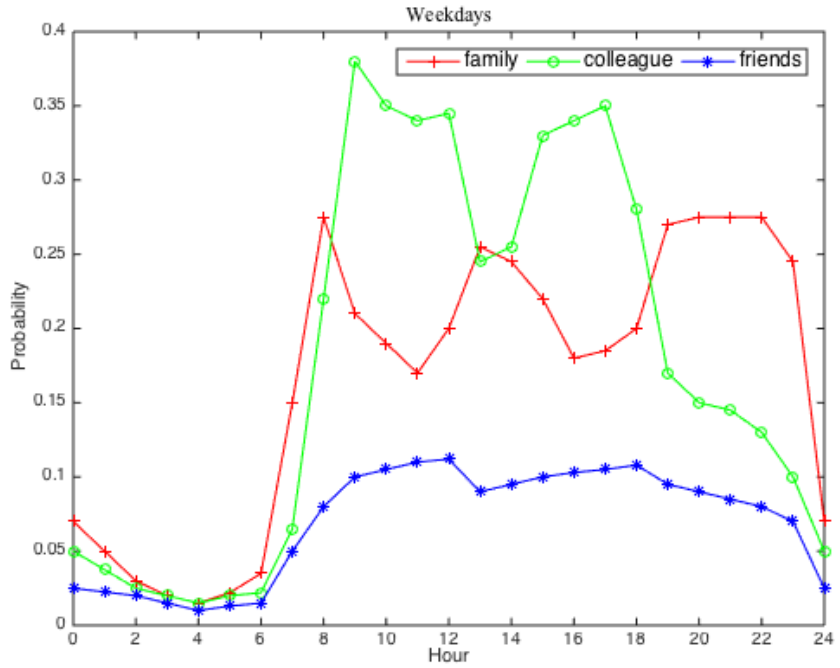
从图中的曲线我们可以观察到，无论在工作日还是在周末，三种社交关系在早上 0 点到 4 点的概率较低，其他时间以及在 Weekdays 和 Weekends 都呈现一定程度上不同的特征。在 Weekdays 的时候，同事在一起比其它关系的可能性要高。但是在工作日的中午，具有家庭关系的用户同现的概率反而增加了，而同事之间同现的概率出现了一个小的低谷状态。分析其中缘由，我们不难发现这个现象只有在中国小城市里面的传统家庭才会出现的。和许多国内外大城市不同的是，在中国中小城市里的人们往往会选择中午回家吃饭，并且进行短暂的休息。我们的数据集以及曲线很好地验证了这一现象。而在周末，家庭关系的用户同现的可能性最高，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，l_x^h 为用户 x 在 h 时出现可能性最高的基站，基站同现的概率

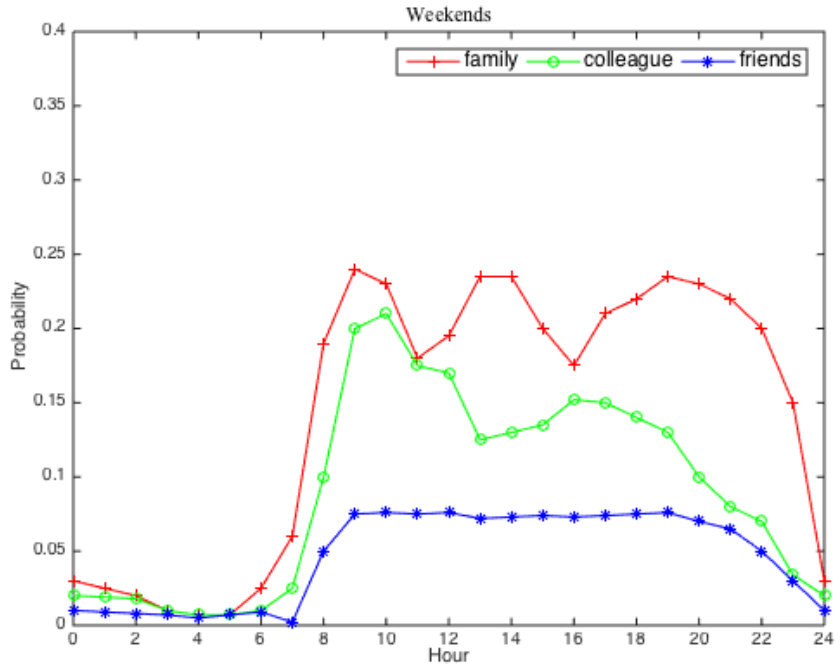
定义 3.3： 最频繁基站同现概率定义，

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

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



(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

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。

结合我们特有的国情，我们使用地理位置服务提供商百度地图 (中国最大的 LBS 提供商) 获取基站范围内的 *POI*(Point of Interest, 地理热点) 信息，对于输入的经纬度信息，百度地图 API 返回一个响应的数据如图3.9所示。

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

图 3.9 百度地图 API 返回 POI 信息

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 Government、Tourists、Food、Entertainment、Shopping、Finance、Health、School、Apartment、Medical、、Road、Company、Car Service、Hotel、Sports、Culture 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，基站及其附近区域往往包含不同的 Context。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 Word 对 Document 的重要程度随着其在 Document 中出现成正比，而与其在整个 Document 中出现成反比[38]。我们很容易知道，这一思想来自于假设对于 Document 中意义的 word 来自于那些在 Document 中出现最多的 word，而在 Document 中出现频率较少的词语集合[39]。由文本挖掘的场景扩充到我们的问题中，TF 即为某种地理含义在特定某 base station 出现的的概率，IDF 即为所有 base station 总数与含有该 context 的 base station 数比的 log。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： $\boldsymbol{TF-IDF}$ 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3.10中各种不同社交关系出行的语义分布。

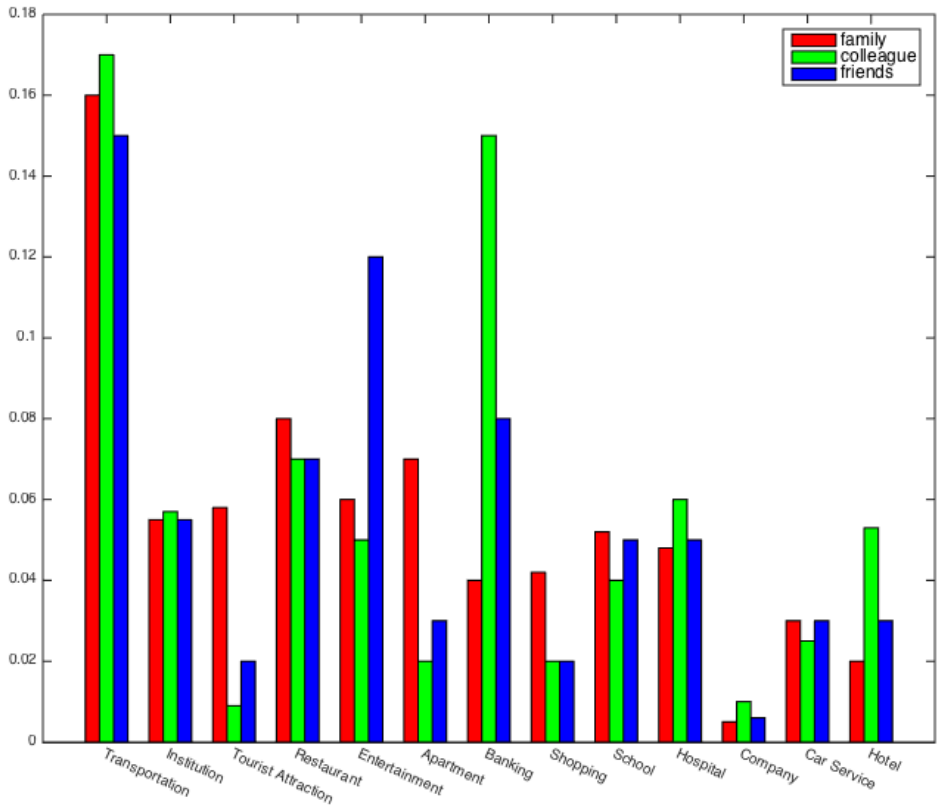


图 3.10 不同社交关系与位置同现语义之间的关系

在图3.10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[28, 40]，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[41](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3.11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 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。

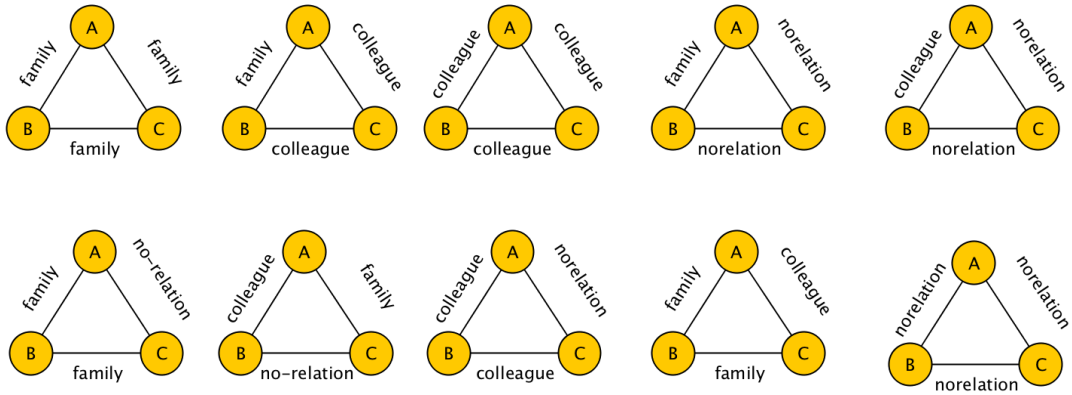


图 3.11 结构平衡理论举例

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3.12，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，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。

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

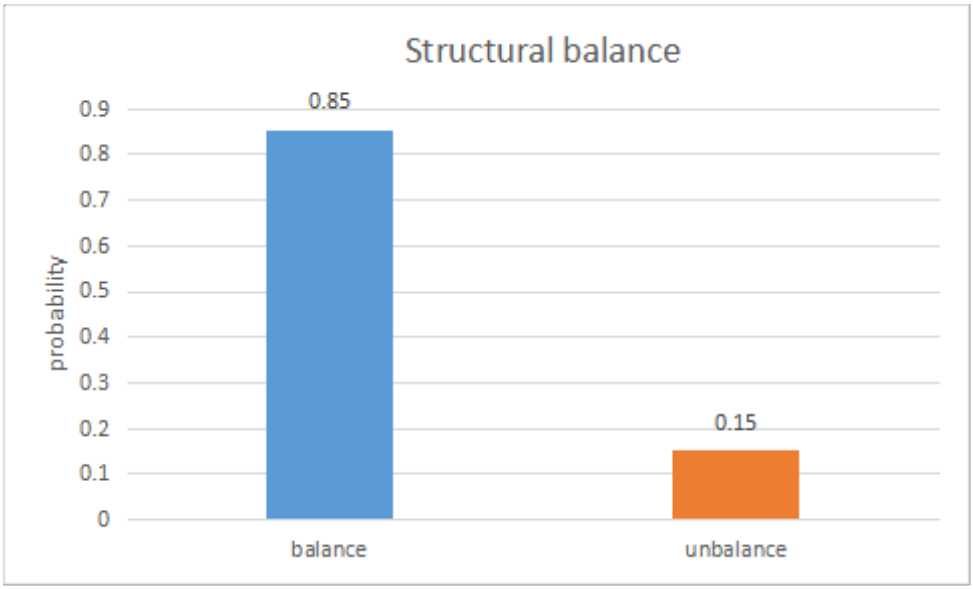


图 3.12 平衡与不平衡三元团结构在社交网络中所占的比例

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比较复杂，不同数据实体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[42] 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

$$h(y_c) = \frac{1}{Z_c}\exp(\sum_{c}\sum_{k}\gamma_c \cdot h_k(\boldsymbol{Y}_c)) \tag{4.6}$$

对于三元结构因子函数 $h_k(\boldsymbol{Y}_c))$，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3.11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[30]，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)[43]。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*)[44] 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

因我们的数据规模较大，而且我们的 *BTFG* 模型基于迭代算法，因此单机单线程实现算法所花费的时间较多。基于以上现状，我们有必要提出分布式算法或者并行算法来提高模型运行效率。

在这里，基于我们的图模型算法，我们提出了一种基于消息传递 (Message Passing Interface) 框架的分布式并行算法。我们所提出的并行算法的基本思想就是将整个移动社交网络拆分为若干个子网络，并在不同的线程中完成对于参数的学习，总体上来讲我们采用常见的分布式方法，MapReduce 的思想来完成并行算法的设计。

因为我们算法训练的消耗时间主要是在学习算法的第一步，所以我们主要加速的就是这一步的工作，即 Map 这一步，我们通过许多个奴¨隶线¨程 (核) 来实现这个过程。而在第二步，我们利用主节点的线程来收集次节点 (前面提到的奴¨隶线¨程) 的结果，并以此来计算每一步迭代中参数的更新结果。

特别地，我们的分布式学习算法总体上基于主人-奴隶结构的分布式框架，总体上能够被描述成两个阶段的算法。首先第一步，我们将整个大的移动社交网络 G 拆分为 K 个子网络 $G_1, ..., G_k, ...G_K$，这里的 K 代表了有多少个子节点线程 (核)。而在第二步，我

们我们分两个步骤来学习模型中的参数。首先，每个县城能够计算它子网络 G_k 的当地边缘概率。然后，主节点线程收集所有奴隶节点线程的子图概率并对所有的参数进行更新。第二步重复直到满足一定的条件，如到了迭代的次数，或者参数更新比较小的时候。

这里有两点值得我们注意。首先，在分割社交网络图的时候，我们根据之前其他研究者工作[45]，我们将移动社交网络根据不同的行政区进行划分，这一部分信息可根据用户注册的所提供的信息。第二个，我们将原来网络中的所有特征都提取出来供每个奴隶节点进行边缘概率的计算。这样，整个算法能够提升效率 10 到 20 倍 (主要根据所使用服务器的 CPU 核数来决定)。

5 模型试验以其评估

为了证明我们所提出模型的有效性以及合理性，我们设计了不同的实验来识别我们数据中的社交关系，并以试验结果来评估我们的模型。

5.1 实验准备工作

5.1.1 数据及评估方法

从前面的介绍中我们可以知道，我们构造的移动社交网络的数据集来自中国中部湖南省的一个县级市，具体的信息可见章节 3。为了更加有效的推断用户之间的社交关系类别，我们仅仅考虑在我们网络中较为活跃的用户。这里我们定义活跃用户为在三周内至少有 5 个联系人与该用户进行了联系，则我们可以认为该用户为活跃用户。而我们所采集具有社交关系的用户们，他们在三周内至少进行了 10 次以上的通话，才可以将这些社交关系放入我们的移动社交网络当中来。需要注意一点的是，在我们的实际数据集中，并不具有朋友关系这一实际关系集合，因此我们将两个通话次数在 15 次以上的用户当作具有朋友关系的用户，这一方法在许多其他的研究中都得到运用[40]。通过这种筛选工作，构成我们移动社交网络大概有 304,000 活跃用户，以及两千万条稳定的社交关系。为了使得我们的结果真实可信，我们将实验重述了 10 次，最终展示结果取 10 次结果的平均值，以准确率、召唤率以及 F1-Measure。这三个概念是常用来衡量机器学习分类器的指标量，具体的计算公式如下[46]，

$$Precision = \frac{tp}{tp + fp} \tag{5.1}$$

$$Recall = \frac{tp}{tp + fn} \tag{5.2}$$

$$\boldsymbol{F1} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5.3}$$

其中 tp 代表实际为正的标签预测为正的总数，tn 代表实际为正的标签预测为负的总数，fp 代表实际为负的标签预测为正的总数，fn 代表实际为负的标签预测为负的总数。

我们核心的代码均用 C++ 进行了实现，并且我们实现的 C++ 代码提供了 Shell 接口，方便用户直接在 Linux Shell 命令行下进行操作。同时，我们的代码具有相当的扩展性，因为时间与篇幅的关系，我们在这里不对其进行详细赘述。除此之外，我们实验的环境都是在一个有 16 核，2.5G 的因特尔顶级处理器，104G 内存，1T SSD 硬盘的谷歌服务器上进行的，充分保证了实验代码运行的速度，以及实验的稳定性。我们的代码没有采用分布式的方法，本身因为实验设备的限制，另外一个原因则是我们服务器的内存以及 CPU 性能相当好。因此，如果直接用并行的方法来处理实验，反而会比 10 台普通机器以内的分布式算法计算得更快 (在消息传递时，单机传递消息在内存中传递，所耗时间非常少)。

5.1.2 实验比较的方法

我们将我们所提出的 *BTFG* 模型与其他的分类器算法进行对比。出了常见的分类器算法外，我们也将我们的方法与最传统的条件随机场算法进行对比。因此这些算法包括朴素贝叶斯 (**Naive Bayes**)，随即森林 (**Random Forest**), 支持向量机 (**Support Vector Machine**) 以及条件随机场 (**Conditional Random Fields**)。对于朴素贝叶斯以及随机森林，我们使用 *Weka*[47] 来实现，并同时考虑了标签分布不平衡问题。对于支持向量机，我们采用 *LibSVM*[48] 的代码实现。对于条件随机场，我们使用所有的因子来进行训练，之后再来对测试集合来进行预测，所用工具为 Mallet[49]。注意到所有的比较方法我们均考虑到了标签分布不平衡问题。

对于所有的比较方法，我们全部使用了第三章所提出来的非结构特征，如时空特征等等。而对于图模型，条件随机场使用了所有的因子来构建模型，但是在训练集中进行训练，在测试集中进行测试，即测试集合训练集是两个不同的网络，并非在一个网络上进行的。而我们的 *BTFG* 模型，则使用了所有我们定义的因子函数来预测不同边的社交关系。这两个唯一不同的一点是，我们的方法直接在一个网络中进行推断，并没有像解决经典的分类问题一样，将数据分为训练集合测试集。因此，在训练的时候，我们可能会使用到未知的标签 (将未知也看作一种类型)。

5.2 实验结果

我们在数据集上面采用不同的方法来对我们数据集中的社交关系进行预测。在预测实验中，我们使用 80% 具有明确社交关系的数据，而将剩余 20% 的数据看作未知的社交关系，以此作为测试。

5.2.1 模型预测性能

表 5.1 不同分类方法在社交关系识别任务上的准确率

Algorithm	Prec.	Accuracy	F1-score
Naive Bayes	0.663	0.685	0.673
Random Forest	0.652	0.703	0.681
Support Vector Machine	0.726	0.720	0.724
Conditional Random Field	0.623	**0.903**	0.749
BTFG	0.762	0.853	**0.798**

表5.1展示了不同分类算法在我们数据集上面的预测任务的结果。很显然，我们的 *BTFG* 模型比其他任何一个分类算法的效果都要好得多。而支持向量机算法是在所有非图模型算法中预测效果最好的算法。条件随机场模型比一系列的非图模型算法都要好，因为条件随机场能够已定程度刻画复杂图结构与社交关系之间的关系，而这些复杂的结构在现实生活中往往是真实存在的。我们的模型比 CRF 的结果更好，原因在于我们运用了完完整整的社交网络结构，而并非将数据分为训练集和测试集。将我们的模型与非图模型算法进行对比，我们的模型在准确率、召回率和 F1-Measure 上面比基本的分类器算法要好得多，将结果提升将近 10% 左右。除此之外，我们的基本的基础算法的识别准确率相对于其他研究来说，算是很高的结果，这也说明我们提出的时空等特征在识别不同社交关系的任务上有着较高的识别度。从识别准确率的角度来看我们的结果，总体上我们能够识别将近 80%。分别看每一类关系，我们能够识别 83% 的朋友关系，70.8% 的同事关系，76.5% 的家庭关系。

5.2.2 特征贡献分析

在 *BTFG* 模型当中，我们充分考虑了通话特征，时空特征，还有前面我们提出的因子，包括平衡因子以及三元结构因子。这里我们实验来检测这些特征对于关系识别模型的最终识别结果的影响以及贡献程度。我们首先将所有的特征放在一起进行实验，即我们所进行的原始实验。之后，我们我们一个一个按照顺序来移除这些特征，这样我们就能清楚的看清楚每一类特征对于关系识别任务预测的重要程度。特别的，我们首先一处时空特征，并将其模型成为 BTFG-G，接着进一步移除三元结构特征，即 BTFG-GT。最后，我们移除平衡因子记为 BTFG-GTB。我们一次对每类模型进行训练任务以及预测任务。最终，我们可以在图5.1中观察到，每一类特征的减少，都会带来相应的识别准确率的降低，只有当所有的因子在一起的时候识别准确率爱最高。这也充分的说明了我们的

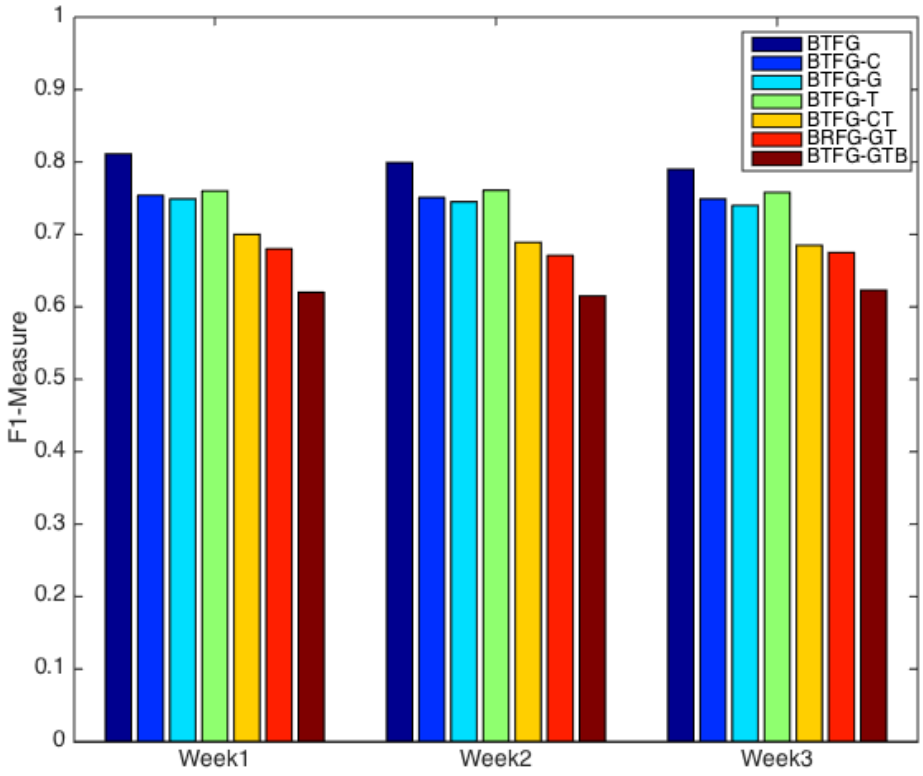


图 5.1 特征贡献分析。BTFG 是提出的模型。BTFG-C 是除去社交特征的模型。BTFG-G 是除去时空特征的模型。BTFG-T 是忽略三元结构特征的模型。BTFG-CT 是同事忽略通话特征和三元结构的模型。BTFG-GT 是忽略了地理特征与三元特征的模型。BTFG-BGT 则是忽略了平衡因子、时空特征以及三元结构的模型。

模型能够很好的将所有不同的因子函数结合起来，并且每一个因子在我们提出的方法中对于关系识别性能的预测都有提升。

5.2.3 标签分布不平衡方法比较

特别地，我们设计实验检验我们提出的解决标签分布不平衡问题方法的有效性。在表5.2中我们可以看到解决标签分布不平衡问题的各类方法的预测性能。我们在每种方法进行 10 次实验之后，去改方法在关系识别任务上面的平均 F1 score。这也显示了在我们的模型中，使用平衡因子不仅仅能够有效改进标签分布不平衡问题，而且我们的方法比传统解决标签分布不平很问题的方法进行比较的时候，我们的方法用来进行关系识别任务要好得多，这也证明了我们所提出的想法的正确性。而使用下采样的结果远比另外两种方法要差，我们认为造成这一现象的原因是由于下采样之后，整个网络中的数据量较少，破坏了原始的网络结构，删去了一些非常重要的点和社交关系边，而剩余的节点

与边并不能很有效的帮助模型来进行关系识别任务。

表 5.2 各种解决标签分布不平衡问题的 F1-Measure Score

Methods	Families	Colleagues	Friends
BTFG-B	0.781	0.743	0.92
Undersampling	0.701	0.642	0.803
Oversampling	0.787	0.741	0.89
BFTG	**0.798**	**0.754**	**0.92**

5.2.4 标签不平衡比例实验

我们进一步的测试了我们提出平衡因子方法的鲁棒性。图5.2显示了在不同关系比例的数据集上关系识别任务的 F1 Score。为了使得实验更加简单有效，我们让两种关系的比例保持一致的同时，再调整剩余的标签数据的比例。结果可以看到，在不同的情形下我们的结果稳定性相当好，最差的情形仅仅损失了大约 2% 的 F1 Score。

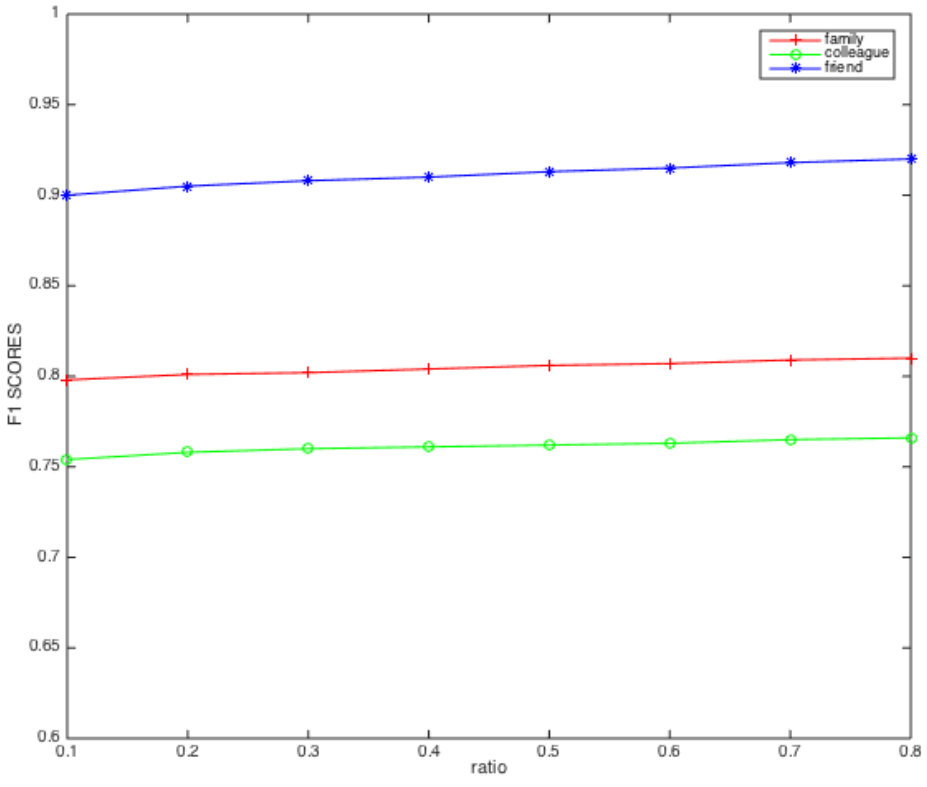


图 5.2 不同比例的不平衡数据集实验结果

6 总结与展望

6.1 完成工作总结

在本篇论文中，我们通过研究在中国河南省中某个县级市的移动社交网络通话数据，来学习不同社交关系在网络中所展示的通话、时空特征，并在此基础上，我们设计出了一个有效识别社交关系的算法。

在探索不同社交关系所具有的时空特征时，我们发现了一些非常有趣的人与人之间关系的特点。首先，不同社交关系的用户往往在沟通的时间段上各有不同，沟通的时间往往与他们的身份比价相符。第二，不同的社交关系往往会出现在不同类的地方，这侧面反映了他们的日常生活。第三点，不同的社交关系的人群往往在时间和空间上表现各不相同。第四，随着人与人之间交流的深入，人们往往会形成较为稳定的交流圈子。

基于这些发现，我们定义了移动社交网络中关系识别问题，并且提出了一个叫做平衡结构因子图的模型 (*Balanced Triadic Factor Graph Model*)，并用 C++ 对该算法进行了实现，尽量采用了并行化处理，提高了实现算法的效率。这个模型结合了我们在移动社交网络中各类发现以及标签不平衡问题，并且在关系识别任务上有较好的表现。

我们将我们提出的的模型与当前比较流行的几种算法进行对比，发现我们的模型在移动社交网络中的关系识别任务上能够显著的提高识别准确率。接着，我们又进行了一系列实验来证明我们模型以及我们理论的准确性以及有效性，最后我们可以得出结论，我们的 $BTFG$ 模型在移动社交网络中的关系识别上有着相当高的识别准确度以及稳定性。

6.2 未来工作展望

在移动网络中检测用户之间的社交关系，使得社交网络更加具有真实性，与我们人们生活的世界更加贴切。对于未来的工作，主要有以下几点需要进行：

- 首先是地理语义提取的方法需要改进。当前我们采用的方法是 $TF-IDF$，该方法较为简单，提取准确率无法得到保障。因此，我们需采用更加有效的方法来提取地理语义。初步的思路是采用 LDA[50] 来改进地理语义提取。
- 当前的时空与社会学理论交叉的不多，但是有很大的空间可以研究。如我们在运用结构不平衡理论的时候，我们其实研究这些稳定的三元团的出行特征，在时空

分布与地理去向上面有什么特点。

- 对标签分布不平衡问题解决方法的改进。当前，我们都是从模型的角度来改进标签分布不平衡问题。而在图模型中，我们可以尝试在网络推断的时候进行选择性消息传播，这样能够更加有效的改善标签分布不平衡问题。
- 移动社交网络中的分析不仅仅可以识别社交关系，还有许多有趣的应用可以研究。如谣言传播、用户画像刻画、网络推演等等，这些研究在当前社交网络的发展中扮演着越来越重要的角色。因此，下一步我们可以进一步扩展移动社交网络的主题研究。

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] 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.

[6] LIBEN-NOWELL D, KLEINBERG J. The link-prediction problem for social networks[J]. Journal of the American society for information science and technology, 2007, 58(7) : 1019 – 1031.

[7] CRANSHAW J, TOCH E, HONG J, et al. Bridging the gap between physical location and online social networks[C]. Proceedings of the 12th ACM international conference on Ubiquitous computing. 2010 : 119 – 128.

[8] 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.

[9] LIBEN-NOWELL D, NOVAK J, KUMAR R, et al. Geographic routing in social networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2005, 102(33) : 11623 – 11628.

[10] LAMBIOTTE R, BLONDEL V D, DE KERCHOVE C, et al. Geographical dispersal of mobile communication networks[J]. Physica A: Statistical Mechanics and its Applications, 2008, 387(21) : 5317 – 5325.

[11] KRINGS G, CALABRESE F, RATTI C, et al. Urban gravity: a model for inter-city telecommunication flows[J]. Journal of Statistical Mechanics: Theory and Experiment, 2009, 2009(07) : L07003.

[12] SCELLATO S, MASCOLO C, MUSOLESI M, et al. Distance Matters: Geo-social Metrics for Online Social Networks.[C]. WOSN. 2010.

[13] XUE A Y, ZHANG R, ZHENG Y, et al. Destination prediction by sub-trajectory synthesis and privacy protection against such prediction[C]. Data Engineering (ICDE), 2013 IEEE 29th International Conference on. 2013 : 254 – 265.

[14] NOULAS A, SCELLATO S, LATHIA N, et al. Mining user mobility features for next place prediction in location-based services[C]. Data mining (ICDM), 2012 IEEE 12th international conference on. 2012 : 1038 – 1043.

[15] MONREALE A, PINELLI F, TRASARTI R, et al. Wherenext: a location predictor on trajectory pattern mining[C]. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. 2009 : 637 – 646.

[16] CRANDALL D J, BACKSTROM L, COSLEY D, et al. Inferring social ties from geographic coincidences[J]. Proceedings of the National Academy of Sciences, 2010, 107(52) : 22436 – 22441.

[17] BROCKMANN D, HUFNAGEL L, GEISEL T. The scaling laws of human travel[J]. Nature, 2006, 439(7075) : 462 – 465.

[18] GONZALEZ M C, HIDALGO C A, BARABASI A-L. Understanding individual human mobility patterns[J]. Nature, 2008, 453(7196) : 779 – 782.

[19] SONG C, KOREN T, WANG P, et al. Modelling the scaling properties of human mobility[J]. Nature Physics, 2010, 6(10) : 818 – 823.

[20] BI B, SHOKOUHI M, KOSINSKI M, et al. Inferring the demographics of search users: social data meets search queries[C]. Proceedings of the 22nd international conference on World Wide Web. 2013 : 131 – 140.

[21] HU J, ZENG H-J, LI H, et al. Demographic prediction based on user's browsing behavior[C]. Proceedings of the 16th international conference on World Wide Web. 2007 : 151 – 160.

[22] LESKOVEC J, HORVITZ E. Planetary-scale views on a large instant-messaging network[C]. Proceedings of the 17th international conference on World Wide Web. 2008 : 915 – 924.

[23] TANG J, ZHANG J, YAO L, et al. Arnetminer: extraction and mining of academic social networks[C]. Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. 2008 : 990 – 998.

[24] HU X, LIU H. Social status and role analysis of palin's email network[C]. Proceedings of the 21st international conference companion on World Wide Web. 2012 : 531 – 532.

[25] ZHAO Y, WANG G, YU P S, et al. Inferring social roles and statuses in social networks[C]. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. 2013 : 695 – 703.

[26] YING J J-C, CHANG Y-J, HUANG C-M, et al. Demographic prediction based on users mobile behaviors[J]. Mobile Data Challenge, 2012.

[27] MO K, TAN B, ZHONG E, et al. Report of task 3: your phone understands you[C]. Nokia mobile data challenge 2012 workshop, Newcastle, UK. 2012 : 18 – 19.

[28] 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.

[29] 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.

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

[31] WAINWRIGHT M J, JORDAN M I. Graphical models, exponential families, and variational inference[J]. Foundations and Trends® in Machine Learning, 2008, 1(1-2) : 1 – 305.

[32] KOLLER D, FRIEDMAN N, GETOOR L, et al. 2 Graphical Models in a Nutshell[J]. Statistical Relational Learning, 2007 : 13.

[33] 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.

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

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

[36] 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.

[37] AURENHAMMER F. Voronoi diagrams—a survey of a fundamental geometric data struc-

ture[J]. ACM Computing Surveys (CSUR), 1991, 23(3): 345–405.

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

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

[40] 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.

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

[42] 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.

[43] LUENBERGER D G. Introduction to linear and nonlinear programming: Vol 28[M]: Addison-Wesley Reading, MA, 1973.

[44] 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.

[45] TANG J, WU S, SUN J. Confluence: Conformity influence in large social networks[C]. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. 2013: 347–355.

[46] OLSON D L, DELEN D. Advanced data mining techniques[M]: Springer Science & Business Media, 2008.

[47] HALL M, FRANK E, HOLMES G, et al. The WEKA data mining software: an update[J]. ACM SIGKDD explorations newsletter, 2009, 11(1): 10–18.

[48] CHANG C-C, LIN C-J. LIBSVM: a library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(3): 27.

[49] MCCALLUM A K. MALLET: A Machine Learning for Language Toolkit[H]. 2002.

[50] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. the Journal of machine Learning research, 2003, 3: 993–1022.

致谢

四年本科生生涯，如今也到了分别的时刻。从高中毕业到如今，一晃已经四年过去了。北航甲子，我只是这个学校的一个过客。四年的交情并不深厚，我也不是一个感性的人，但如今离别之际，竟然也浮生不少伤感。回望过去四年，太多的人给予我帮助，如今到了毕业之际，也是感恩之时。

首先要感谢的是吕云翔老师。本研究是在我的导师吕云翔老师的指导下完成的。在此之际感谢吕老师在过去四年里对我们悉心指导与关怀。虽然我的毕设是在校外做的，但吕老师仍然严格要求，无论是具体研究分析，还是论文写作，吕老师都给予我细致的指导。吕老师渊博的理论知识、严谨的科研态度以及丰富的学术经验给我留下了深刻的印象，使我受益终生。

感谢我在北大做研究时，指导我的宋国杰老师。当初宋老师不嫌弃我是北大校外学生，能力不够，反而委以重任，手把手教我做研究，让我从事社交网络的研究。是您带我走进科研的世界，让我感受到了科研的美妙。当我选择放弃读您的研究生反而选择出国时，您仍然鼓励我要做自己想做的事情，也继续帮助我从事科研的工作。您对我的这份恩德我无以为报，希望您在以后的事业中研究成果愈加丰硕，家庭愈加美满。

感谢大三暑假在 CMU 做研究时的导师，教授 Justine Cassell。您让我切身感受到了真正世界级学术大牛的风采，也让我萌生了以后走学术道路的想法。当然，也感谢您在我申请的时候给予的推荐信，让我有机会在北美直接读 PhD，继续从事自己喜爱的研究工作。同时也感谢博士后 Alex 和 Yoichi，使得我有机会在大三暑假做了一些非常有意义的工作。

感谢在暑期科研期间所有帮助我的学长和同学们。特别感谢徐可扬学长，作为我的的直系学长，你在学术上的追求以及在为人处事方面一直是我的学习榜样，同时感谢你在我申请和科研路上一直提供帮助，帮我解答人生疑惑，让我坚定了投身科研的信心；感谢实验室的赵冉学长，非常怀念和你在匹兹堡凌晨三四点的时候一起讨论学术问题，希望以后再有机会一起合作；还需要感谢生活上帮助很多的室友胡张柠、秦宇、戴自航学长，希望胡张柠学长在纽约 Google 工作顺利，秦宇学长顺利从 MCDS 项目毕业找到心仪的工作，戴自航学长在 CMU 读博一帆风顺，争取能以后和你有 Deep Learning 方面有意义的合作。

另外感谢在学术生活道路上一起成长的同学刘天毅、张元，祝天天到了清华后多发

paper，研究生申请到好学校，祝元神到北大多发 paper，以后希望能有一起合作的机会；同时也非常感谢同班的杨缘大大，大帝以及学佛同学，你们的存在让我明白大神非一日所成；同时需要感谢的基友李捷，本科最怀念的时光就是和你一起做项目熬夜的时间；另外感谢同班的朱公朴同学，祝你公司能够早日上市，和你女朋友感情长久；感谢申请路上一直陪伴的北邮大明，清华珊姐和加拿大滑铁卢大学的 ViVi，希望你们以后都前程似锦，生活如意。

最后需要感谢始终爱我、关心我、鼓励我的父母。虽然我的父母都没有上过大学，但是在我追逐梦想的道路上，我的父母从来都是支持我的选择，从而我有机会从事计算机科学的研究，走上学术这条道路。