分类号: 单位代码:
密 级: 学 号:

杭州師範大學
HANGZHOU NORMAL UNIVERSITY

硕士学位论文

中文论文题目：在线社交网络的结构分析、建模及预测

英文论文题目：The Analysis, Modeling and Prediction for The Structure of Online Social Network

申请人姓名：尤志强

指导教师：韩筱璞

合作导师：尚明生(电子科大)

专业名称：计算机应用技术

研究方向：复杂网络及数据挖掘

所在学院:

论文提交日期:

摘 要

在线社交网络是由多社交主体互动构成的社会关系结构,其作为人类自发形成的社会结构，蕴含着丰富的人类行为模式、规则、机制，以社交网络角度去深入分析结构性质，探究背后的机理，可以加强对人类行为的认知与理解，能够促进潜在商业价值的挖掘，因此这方面的研究有着巨大的理论与实践意义。

本文基于复杂网络理论方法，研究了在线社交网络结构的演化及预测机制。文章主要包含两方面内容，分别是针对QQ群网络上群结构的性质特点和演化机制的研究，以及以微博用户关注网络数据为基础，分析并设计用户关注行为的预测算法。

其中，针对 QQ 群社交网络的研究，一定程度上填补了当前社交网络研究存在的一大缺失。由于缺乏有效可信的群组社交关系数据，目前关于在线社交网络的研究主要集中在个体社交关系上，而较少涉足于群组层面的群体行为，使得我们对于社会群组的加群行为机制知之甚少。本文通过分析 QQ 群结构特性，发现两种不同的群生长规则，并进一步基于实证发现，提出基于渗流理论的兴趣扩散模型来解释群组演化机制。实验结果显示所提出的加群机制能够有效解释群组生长规律。

此外对微博用户关注网络上关注行为的链路预测算法研究则进一步弥补了复杂网络现有链路预测算法的不足。类似微博这样的社交网络上用户的关注行为主要是由兴趣驱动,并且用户所发表的微博文本内容可以显性得体现出个人的兴趣偏好。针对传统复杂网络链路预测算法不能有效利用文本信息的不足，本文在对用户兴趣及信息传播特点的实证研究基础上,提出一种新的局域计算链路预测指标。在 twitter 及新浪微博数据上的实验结果显示，该指标在 AUC、准确度及

召回率等方面的表现远远优于传统的链路预测算法。

关键词：在线社交网络；复杂网络；群组网络；加群行为；兴趣扩散；微博关注网络；文本信息；链路预测

Abstract

Online social network(OSN) is a social relationship structure which is formed by multi-agent interactions. As spontaneously formed social structure, OSN contains ample patterns, rules and mechanisms of human behaviors. To discover network structures and explore hidden principles can help us understand human behaviors which will bring potentially commercial values. Thus, studies on OSN have a greater theoretical and practical significance.

Using the theories and methods of complex network, we study the evolution and prediction mechanisms of OSN structures. This paper mainly includes two parts: the first part is the studies on the properties and evolution rule of the group structure in QQ group networks and the second part focuses on twitter social network to design a new link prediction algorithm for user following link.

The research on the QQ group networks offsets the deficiency of current studies about OSN. So far researchers have generally paid their attentions on individual social relationships, leaving participation in social groups less understood. It is because effective and credible data of collective human behaviors as a group is hard to collect. Through the analysis on the characteristics of QQ group structures, we find two different types of online social groups' growing rules. Further, on the basis of empirical findings, we propose a percolation-like diffusion model to explain the social groups' evolution rule. The model results indicate that the proposed mechanism is an important driven-factor for the growth of real social groups.

Moreover, the research work on the link prediction for user following relationship in twitter social networks which covers the shortage of existing link predictors in complex network. Social networks like twitter are interest-driven online community and the posted tweets can explicitly reveal users' personal tastes. While existing link prediction methods in complex network are unable to use text information. In order to make up the shortage, we propose a new local indice of link prediction which can both utilize users' interests and network topology. Experiment results on twitter and sina weibo show that the new algorithm can outperform other

traditional link predictors in AUC, precision and reall.

Keywords: Online social network; Complex network; Group network; Paticipation behavior; Interest diffution; Twitter social network; text information; link prediction.

目 录

1 引言

1.1 在线社交网络的研究背景与意义

社交网络，也叫社会网络，是由多社交主体互动构成的社会关系结构，是一种基于节点之间相互连接的社会组织形式，其包含两大要素：节点（主体，通常指个人或组织）和边（社会关系）。社会弱关系到强关系等各种关系通过将各类节点连接起来，构成我们熟知的各类社交网络，比如朋友网络、同学网络、商业伙伴网络或者科学家合作网络等。社交网络作为人类自发形成的社会结构，其蕴含着潜在的人类行为模式、规则、机制，以社交网络角度去深入分析人类行为、探究背后的机理，对于我们认识自己，利用这些潜在的规律创造商业价值有着巨大的科学与实践意义。由于精确获取传统的社交网络数据存在难度，因此我们转向在线社交网络作为切入点。

在线社交网络，是人类真实的线下社交行为在线上的一种反映，是一类能够帮助用户寻找好友，维持好友关系，建立在线社交圈子，是一种支持用户在平台上发布信息，分享心情、兴趣爱好、活动状态的在线应用服务。比如我们熟知的facebook、人人网、twitter、新浪微博、QQ等在线社交媒体，这些平台允许用户创建、发布、阅读、转发、分享、评论相关内容，创建群、邀请好友、加入兴趣组、关注好友，群体讨论、投票等等。这些行为使得在线社交网络成为线下社交网络的一个镜像，这种自组织构建的网络，同样蕴含着人类的行为机理。在线社交网络是非常重要的人类行为数据提供源，数据一般包括：用户互动与内容分享[1,2],用户的感觉[3]，观点与情感表达[4]等文本、多媒体信息以及用户朋友关系或者关注关系等拓扑结构的信息。

对于在线社交网络的研究，有助于我们更加深入理解社交网络的性质特点、

群体及个体行为层面的行为、信息传播机制、重要节点有效挖掘的规律等。能够更好得控制谣言传播、净化水军，提高有利信息传播的有效性，对用户流失进行预警，提高内容推荐、好友推荐、广告投放的精确性，提升用户体验与商业收益，这不仅能够产生丰富的学术成果，而且对于商业应用，前景巨大。

本文主要关注两类在线社交网络，一类是QQ群在线社交网络，QQ群网络有别于之前接触到的人人网或者Facebook网络，该网络明确定义了群结构，用户可以建群、加群、退群，群规模伴随着生长、消亡过程。而传统社交网络的群落结构一般是通过社群划分算法来区分，由于受限于算法的准确度等问题，这会对研究群组层面的群体行为造成噪音偏差。另外之前的研究由于受到数据收集困难的影响，造成这方面的社交网络研究非常少，使得我们对于该类群组网络的用户群体行为了解的非常少。对群演化机制缺乏理解导致现在使用的群推荐算法等常常会失效，达不到理想的效果。因此对这类群组社交网络上群体行为的研究，将有助于加深我们对群的演化生长、用户行为特点等的认知，进一步帮助我们设计更加准确有效的群推荐算法，提高用户体验。精准的群推荐算法能够帮助用户寻找到其需要的适用的群，提高信息在该社交网络上的有效传播，因此这方面的研究非常有现实意义。另一类在线社交网络为twitter及新浪微博用户关注关系网络。微博这类社交网络上用户的行为偏向于兴趣驱动，这与Facebook、QQ等社交网络存在差异。Facebook类别网络一般来说是熟人社交网络，用户的好友一般为同学、家人、公司同事等，线下关系对用户行为的影响相对来说更明显。而微博上用户的关注对象范围更广，可以关注其他城市的用户或者明星、教授，甚至是美国总统，只要用户感兴趣都可以关注，此类网络上的关注边通常是单向关注。此外，用户在微博上可以发布原创内容或者转

发其感兴趣的信息。这些文本信息体现着用户个人的兴趣爱好。在这类社交网络上，为用户推荐其潜在的感兴趣的好友对于整个网络生态健康发展尤为重要。好的推荐，可以帮助用户建立自己的结构稳定且活跃度高的社交圈子，寻找到其确实喜欢的好友，这可以提高用户对在线社区的粘性，保持高用户活跃性。经典的复杂网络链路预测算法在预测未来的用户关注关系时，通常只利用了网络拓扑信息，很难将文本数据加以利用。因此，本文提出一种既可以利用文本信息同时又兼顾拓扑结构的新链路预测算法来提升好友推荐的精准度。这项工作对于推动学科发展以及提升商业价值具有很重要的作用。

1.2 在线社交网络的研究进展

社交网络挖掘是一个非常热的研究课题。目前，国内外学术界对社交网络进行了相当多的研究。虽然不同科研领域的学者对于社交网络问题，在解决思路与问题关注点上会有些不同，但是由于社交网络问题的复杂性，目前呈现出非常明显的学科交叉趋势，利用多领域知识协同分析与解决相关问题。 我们常常可以看到不同领域学者进行科研合作，多种不同领域方法混合使用。当前针对社交网络的研究，主要集中在数据挖掘、机器学习、复杂网络等科研领域，同时还包括了人工智能、优化、图理论、移动计算、数据库等其他领域。研究着眼点大致有：社交网络结构分析与建模、特殊结构识别、网络结构反推、社区识别、子图模式识别、动态网络演化、单一网络及耦合网络的信息传播与控制、重要节点识别、信息安全与隐私保护、链路预测、个性化推荐、精准广告投放、信用分析、情感分析、观点挖掘、用户聚类及分类问题、用户画像构建、用户流失预警、大规模数据存储与计算等。数据挖掘技术一般用于信息检索、统计模型构建，其能够为从大规模数据中提取诸如趋势、模式、规则等有用信

息提供非常广泛且有用的技术与知识。机器学习则是通过计算机优化学习，找出最接近于真实数据的分布函数，一般在特征抽取、分类、聚类、预测等问题上有很好的表现。复杂网络主要针对网络性质、结构演化、信息传播、控制、重要节点识别、链路预测等问题。一般来说，能够通过关系构建的网络都可尝试应用网络科学的理论与方法来进行研究。

目前，比较热门的社交网络研究问题有很多。由于我们本文关注的是群组社交网络结构特点及演化问题和微博社交网络链路预测问题，首先我们将回顾这两方面的相关工作：

社会交往是我们作为人类必不可少的部分，针对社交网络结构的研究与理解对于帮助我们更好地认识自己，挖掘价值具有非常重要的意义。早在1930s，相关的研究就已经出现，六度分离假设[5]，指的是任意两个人之间可以通过较少的熟人进行相互联系。这一假说最近已经在Facebook网络上得到了验证[6]。近年来，随着在线社交网络的不断发展，关于人类行为的社交网络数据更加容易被获得，这大大加速了针对社交网络的研究。很多学者对社交网络的社团结构进行了研究，社团结构指的是网络模块内部节点之间的连边密度要远远大于这些节点与该模块之外的节点的连边密度。James等人研究了社交关系连通性与社团结构的关系，提出了social cohesion的概念[7]；Amanda等人则基于Facebook数据采用社团结构等指标来研究理解线下关系的发展[8]；其他学者将注意力放到了社团结构的探测与识别算法研究上，如Jure等人比较了多种在线社交网络社团识别算法的效果及关系[9]。在复杂网络界，Girvan和Newman提出一种使用介数中心性来寻找社团边界[10]。随后，他们又提出基于模块指标的网络划分算法[11]。不过最近研究表明随机网络具有较多的模块结构，这会导致在一定模块规模范

围内的社团结构很难被探测到[12]。从这可以看到使用社团探测算法去识别社团结构还是存在一定的误差。此外，在线社交网络演化问题也得到相关学者的关注。Alan等人研究了Flickr社交网络的生长问题，特别是链路形成过程的研究，发现那些拥有较多连边关系的用户更倾向于创建新的连边[13]。Newman则对科研合作网络的演化显示出浓厚兴趣，研究发现科学家合作概率与他们共同的熟人数量、前期合作的人数、前期合作的工作数等具有强烈的关联[14]。Barabasi也对科研合作网的演化进行过研究，提出模型模拟网络的随时间演化过程[15]。可以看到，随着Facebook等在线社交网络数据获取便捷性的提高，我们对其研究越来越深入广泛。但是从中也可以看到，由于缺乏明确定义的社群网络数据，我们对这类网络的性质特点及社群演化规律还知之甚少。如果直接使用社团划分算法来定义群结构，由于算法本身的局限性，会导致结论不可靠。因此针对社群网络的研究，需要有明确定义群的数据的支持。

链路预测是本文中关注的另一个社交网络研究点。网络中的链路预测是指如何通过已知的网络结构等信息预测网络中尚未产生连边的两个节点之间产生连接的可能性大小[16]。复杂网络领域提出了一系列的算法，最简单的链路预测算法框架是基于相似性。假设每一个节点对(x,y),计算得到其相似性值为S_{xy}，所有网络中当前未观测到的链路按照相似性值进行排序，排在越前面，也就是相似性越高的链路，越有可能被预测为将来会存在的边。尽管这类算法框架的思想很简单，但确实具有非常不错的效果。因此整个算法关键就是如何去设计相似性值的计算函数。如果两个节点之间具有共同的一些特点，那么节点之间一般具有较高相似性。但是这些特点一般都是隐藏非显性的，因此复杂网络中的经典链路预测算法一般采用结构相似性。这些算法大致可以划分成以下类别:局

域相似性、全局相似性、无参以及带参数函数相似性、基于节点或者基于边的相似性等等，具体的讨论可参见文献[16]。基于局部信息的相似性指标，从不同角度以不同的方式可产生多种相似性衡量指标，如Salton[17]，该算法采用类似于余弦相似性的计算公式；此外还有Hub Promoted Index[18]，Leicht-Holme-Newman(LHN1)[19]等。基于路径相似性指标，如局部路径指标[20]、全局路径指标Katz指标[21]等。基于随机游走的相似性指标，如平均通勤时间[22]、有重启的随机游走[23]等。基于最大似然估计的链路预测[24]。在机器学习、数据挖掘中，链路预测同样非常重要。文献[25]将链路预测看成是监督学习问题，将原始的排序算法看成二分类问题，通过抽取用户连边的多个有效特征，再使用分类算法进行连边与否判断。文献[26]使用类AdaBoost的迁移学习算法对网络中的边进行预测。本文主要针对复杂网络中经典链路预测算法无法有效利用文本信息的缺陷，提出新的链路预测算法指数。

除此之外，社交网络的研究热点还包括以下方面：

1、非社群网络的结构分析，对这些类别网络的拓扑结构、度分布，静动态结构分析与演化都是热点。复杂网络领域，对网络结构的研究日趋成熟。Watts[27]提出小世界模型，通过实证发现大多数社交网络平均路径在6步以内，且具有较大的簇系数。Barabasi与Albert[28],Dorogovtsev和Mendes[29,30]以统计机制视角对增长性质的网络结构做出开创性的工作等等。在机器学习与数据挖掘领域，同样存在很多工作。基于连接的结构分析[31],静态动态网络分析[32,33,34]。Friedman等人提出probabilistic relational modes[35]，Kersting等人提出Bayesian logic program(BLPs)[36],Jensen等人提出relational probability trees（RPTs）[37]来应对用户的关系数据。

2、重要节点识别,早期一般会比较偏重图论的方法，来识别出哪些是具有大影响力的节点，哪些是重要的边。比如文献[38]使用中心性度量值来衡量其对团簇形成的影响力。文献[39]则利用参数化中心性指标来研究网络结构并对节点的连接性进行排序。近年来，复杂网络兴起，对于网络的相关问题解决有着天然的优势。针对重要节点识别，复杂网络界学者们提出了一系列的方法：Kitsak等人[40]提出使用K-壳分解法来确定网络中节点的重要性位置，其核心思想为将外围的节点一层层去除，落于内部的节点往往具有更高的影响力，该方法是一种粗粒度的基于节点度的重要性排序方法；此外还有基于路径的排序算法[41],该算法采用无环路由策略，使用拓扑排序将网络表示为从源节点到目标节点的有向无环图，并且同时考虑最短路径及网络流在传输过程中的非最短路径;基于信息指标[42],该算法通过路径中传递的信息量来衡量节点的重要性，其核心思想是假定信息在网络路径中传播是会存在噪音的，长路径往往具有更大的噪音，将路径长度的倒数定义为路径上信息传递量，节点对(V_i,V_j)间的传递信息总量即为两个节点之间所有路径上传递信息量的总和;随机游走介数中心性算法[43]，在该随机游走过程中，计数较多的一般是那些短路径，即该过程相当于给短路径上的节点赋予更大的权重值；基于特征向量的排序算法[44,45,46,47]等，其中值得一提的是leaderRank算法，该算法由pagerank演化而来。在pagerank算法中，每一个节点具有等价随机跳转概率。然而实际情况是热门网页相对于冷门网页的被选择概率要更高，而且pagerank算法中参数c的选择一般是通过实验获得，不同的网络上的参数c存在差别。而leaderRank算法可以很好地应对这两个问题，其核心思想是通过添加一个背景节点作为中继节点，该节点与网络中所有节点都是双向连通，然后通过类似于pagerank随机游走就能得到一个无参的算法，

实验结果表明该算法相对于pagerank算法，能更好识别重要节点。

3、社团划分同样是研究的热点。传统的方法有分级聚类[48]。复杂网络界提出基于模块度的社团检测算法（CNM）[49]，其核心思想是模块内的边的密度要远远高于模块内节点与其他模块之间的边的密度，该算法在运行时间上要远远低于现有算法，特别是在应对大规模网络。关于多片网络社团检测算法[50]，作者提出一个通用网络质量函数来帮助找到多片网络中的社团结构等。

4、个性化推荐问题。由于社交网络上的用户数呈现指数增长，其内容甚至近乎海量，如何设计算法帮助用户获取其最需要的内容以及最想关注的对象，成为科学家热衷的课题。由该问题发展出个性化推荐领域，推荐算法旨在根据用户的兴趣特点和历史行为记录，向用户推荐感兴趣的信息，这使得用户可以节省大量浏览搜索时间，提升用户体验，避免用户产生信息过载问题。对个性化推荐系统研究，各领域专家都投入了相当大的关注，创造了一系列的推荐算法：基于相似性算法，有user-based、item-based的协同过滤算法[51,52,53]。复杂网络界针对信息分配提出了Diffusion-based methods[54,55,56]等。机器学习与数据挖掘领域一般注重对特征的抽取，比如用户特征、对象特征，用户与对象之间关系特征（偏好）以及结构特征。此外还有使用最近邻协同过滤方法，矩阵分解提取偏好特征方法 [57,58]等，特别是矩阵分解算法近年来得到了非常广泛的关注，由于其具有很好的扩展性，在原始矩阵分解算法上可以加入多种考量因素，比如用户偏置效应，有些用户对商品总是会打出比别人高的分，或者有些用户评分尺度很宽松，同样有些商品总是可以得到更高的评分，这样就产生了带偏置的矩阵分解推荐算法。此外，还衍生出融合隐式行为（点击、收藏、检索）等的矩阵分解算法，也有考虑时间因素（用户兴趣随着时间变化）。

其他的一些机器学习推荐算法基本上是对原有算法进行融合，在简单结构信息或者内容信息基础上增加社交信息、用户个人特征数据等，然后使用学习算法进行学习[59]。另外，在推荐领域，除了关注算法外，另外一些问题同样得到了研究者很浓厚的兴趣：数据稀疏性[60]、可扩展性[61]、冷启动[62,63]、兼顾准确性与多样性[64]、短期与长期兴趣[65,66]等。

5、数据降维，由于社交网络的结构比较复杂，用户关注关系维度，用户特征维度，商品条目维度等较高，如果直接将原始数据作为特征输入，会导致算法过拟合或者表现过差，计算复杂度较高，因此数据降维对于实现好的学习算法或达到好的结果具有非常重要的意义。机器学习领域对社交网络数据的降维做出了非常多的工作。Horak Z等人[67]使用形式概念分析与矩阵分解混合算法对社交网络数据进行降维。D.B.Skillicorn[68]使用Sigular value decomposition（SVD）、Semidiscrete decomposition(SDD)、Independent component analysis（ICA）等对社交图网进行矩阵分解将高维数据投影到低秩空间进行压缩提取主要网络主要信息。还有使用Principal Component Analysis(PCA), MulTidimensional Scaling(MDS)[69,70],Generative Topographic Mapping(GTM)[71,72], Self-Organizing Maps(SOM)[73], SMACOF[74]等其他降维压缩方法。

6、信息传播，针对社交网络上信息传播模式的研究对于理解信息扩散机制、谣言控制等具有非常重要的作用。文献[75]采用疾病传播模型SI与SIR模拟研究信息传播,发现具有越多的好友的用户会在接收信息上而不是传播信息上产生影响，同时发现整个网络中的用户数不会对信息传播速度产生作用。文献[76]基于网络用户相似性对信息在网络上的传播进行模拟来研究信息传播的影响要素。Pedro等人研究了社交网络上的谣言传播，提出两个衡量指标，一是传播因

子，衡量的是节点与其他节点之间进行信息交换需要经过的平均最大邻居数，二是传播速度，衡量信息扩散快慢。他们将所提出的信息扩散模型应用到无标度网络及小世界网络，发现好友关系的数量会强烈影响谣言的扩散[77]。Eytan等人则研究了社交网络在信息传播中所起的作用，通过实验发现尽管强关系在信息传播中对于个人来说具有更大的影响力，但是那些弱连接则扮演着扩大传播新鲜信息的角色[78]。

此外，Nozomi Nori[79]关注用户打标签行为表现出来的兴趣，通过迁移学习，来预测其在其他社交网络上的兴趣表现，达到聚合多源数据的目的。社交网络的用户分类问题[80,81]，隐私保护问题[82,83]，用户流失预测[84,85,86]等也都是比较热门的社交网络研究点。

由此可见，关于社交网络的挖掘，已经有很多成果。不仅在学术界产生影响，而且目前很多已经被成功应用于商业领域。因此，针对该领域研究课题的探索，极具科学与现实意义。目前，虽然国内外学术界对社交网络做出很多的研究，但是对于一些有价值的社交网络的实证研究与机制探索依然欠缺，对于实际问题的处理虽然已经提出了相关的解决思路，但在效率与准确度之上依然有提升的空间。

本文主要从两个角度出发，一个是通过研究群组社交网络的特点性质，加深对这类社交网络的认知，从而一定程度上为设计更加准确的群推荐算法做理论准备，以缓解目前群推荐算法存在的推荐效果差的问题。另一个则是从算法适用性出发，针对当前复杂网络界链路预测算法存在不能有效利用文本信息的问题，提出新的算法，以更好地服务于微博社交网络的链路预测及用户推荐应用。

1.3 本文主要工作及研究方法

本文主要是通过研究在线社交网络来解决现实中存在的算法准确度及算法适用性问题。具体来说包括两部分：

1.由于缺乏对群组社交网络及用户加群行为的理解，现行群推荐算法效果差强人意，精准度不高，给用户带来不好的体验。本文核心思想是通过分析QQ群组网络的群结构特点，不同类别用户的群体行为以及群群之间的关系及差异来加深对该类社交网络的认知，并试图找出导致群推荐效果差的原因。此外我们针对实证发现，提出相应的群生长机制，来揭示用户加群的规律。该工作能够帮助我们理解群组规模演化及用户加群的机制，理解不同类别用户行为，进而帮助设计推荐准确性更高的群推荐算法。该研究主要工作包括：

（1）基于用户与群的从属关系构建三类网络，分别是用户与群的超图网络、群群网络以及用户用户网络。这样做是为了便于研究用户与群、群与群、用户与用户之间的关系。

（2）研究用户年龄与加群行为关系。群中用户年龄分布、年龄差异，用户加群数与年龄关系，邻居用户年龄与用户邻居数的关系，加群兴趣偏好转移与年龄的关系，群规模演化与加群兴趣偏好转移的关系等。

（3）研究用户性别与加群行为关系。男性用户与女性用户在加群行为上的共通性与差异性，区分性别后再次计算用户年龄与加群行为的关系。

（4）基于实证发现，提出类渗流的兴趣扩散机制对用户加群行为进行

建模，试图解释群组规模演化的背后规律。

2.准确的链路预测算法，可以被用于社交网络上用户画像构建，潜在好友的推荐，精准广告投放，不但可以提高用户体验，而且可以增强用户粘性，提升商业价值。经典的复杂网络链路预测算法一般使用网络拓扑结构信息，没有利用文本信息。在用户加入twitter或者新浪微博初期，没有发表一定数量的微博，这些基于结构信息的算法较为适用，但当该类社交网络上，用户发布较多微博内容后，这些算法由于存在不能有效利用文本信息的缺陷，就显得不那么适用。这是因为微博是一类兴趣驱动社交平台，用户所发表的微博内容可以显性表现出个人的偏好，如果不加以利用，而只使用结构上的信息，将会导致算法准确度上大打折扣。因此，本文提出一种同时利用文本内容信息及结构信息的链路预测算法。该研究工作主要包括：

（1）回顾经典复杂网络链路预测算法。介绍算法的核心思想以及在twitter及新浪微博数据上的结果表现。

（2）对微博文本数据进行预处理。包括分词、停顿词删除、主题划分、词向量构建、用户兴趣向量修正等。

（3）实证分析微博数据上存在关注关系的用户对以及无关注关系用户对之间在兴趣表现上的差异，并分析拓扑结构在信息传播中的作用及影响。

（4）提出新的链路预测指标，并与经典算法进行性能比较，而且对所提出的算法进行讨论分析。

1.4 论文的组织结构

本文的具体结构安排如下：

第一章：引言。介绍了本论文的研究背景、研究意义，国内外发展现状，研究方法以及论文的结构安排。

第二章：对QQ群社交网络性质特点的分析与研究。分别从群网络结构，年龄、性别因素对用户群体行为造成的影响，群群之间关系及差异，以及群生长模型等方面对群组社交网络进行详细研究，发现了群生长的两种不同机制，揭示了用户加群机制，并分析出一系列在群推荐算法设计中需要注意的关键点。

第三章：基于twitter及新浪微博社交网络数据研究用户关注行为预测算法。首先回顾复杂网络中经典的链路预测算法，描述现有算法的不足之处。然后实证分析twitter及新浪微博这类社交网络上存在关注关系的用户对及无关注关系用户对在兴趣表现上的差异。此外，分析拓扑结构在信息传播中的作用及影响。最后，基于实证发现，提出可以利用文本信息及网络结构信息的新链路预测算法，并与经典算法在twitter及新浪微博数据上进行实验结果比较。

第四章：总结与展望。总结本论文所做的主要工作，并指出今后进一步研究的方向。

2 QQ群在线社交网络结构分析及建模

群组社交网络是一类有别于Facebook、twitter等在线社交网络的网络结构。该社交网络上明确定义了群结构，用户具有建群、加群、退群等行为。这类网络无需通过社团结构探测算法对网络中的群（社团）结构进行隐性识别。在此之前，由于Facebook、twitter等网络数据相对容易获取，针对此类社交网络已有非常广泛而深入的研究，关注个体用户社会关系的研究也层出不穷。但是相比较而言，明确定义了群结构的社交网络的数据获取较难，因此对该类别网络的研究还很少，这也导致目前我们对这类网络上用户的群体行为（如加群行为）还知之甚少。正因为对群组社交网络这样的复杂系统缺乏了解，在此类系统上设计群推荐算法往往难以取得好的效果。现行的群推荐算法往往出现推荐准确性差的问题，给用户带来不好的体验。本章内容以QQ群数据为基础，深入分析这类社交网络的性质特点、用户群体行为，试图找出该类网络上用户加群的机制，并分析影响群推荐效果的原因，为下一步设计更加准确的群推荐算法提供必要的实证及理论基础。

2.1 数据

2.1.1 数据集

腾讯QQ是由腾讯有限公司于1999年开发的在线即时聊天工具。到现在，其已经拥有将近7亿活跃用户，是中国最大的在线聊天平台。用户可通过PC或者智能移动设备在该平台上发送消息及文件、发布微博、语音视频聊天等等。群功能是QQ为方便多用户沟通而推出的举措。QQ用户在群内发布消息能够立刻被群中其他成员所接收到，类似于广播的功能，并且如果需要，

群内任意两个用户都可以开展双人私密对话。一般来说，每一个QQ用户都具有创群能力，能够创建不多于6个群。当群创建完毕，其他用户可通过检索群名或者群号搜索该创建的群。当用户需要加群时，可向群管理员提交申请，审核通过后即加入该群。不同级别的群的用户规模上限是不同的，分为100,200,500,1000几档。群的形成有几种可能性，一类是由于大家都有共同的喜好，比如喜爱看电影，那么创建电影群。另外是为了方便线下活动，如同学群。后者原因创建的群，群中成员往往是线下熟人。

本章所使用的QQ数据集包含58,523,079个群以及274,335,183个用户，其中48,676,355个群同时具有ID、成员列表、创建时间等信息。由于普通用户加群数量上限最大为2000，因此我们将加群数大于2000的用户进行了剔除，这样的用户总共有34个，且基本属于企业营销号、腾讯内部账号以及机器人账号几个类别。另外，由于一些QQ用户未提供年龄信息，如0岁，以及一些比较异常的年龄，如1岁，在分析年龄相关的指标时，我们只保留10到70岁之间的用户进行分析，这批用户数为244,521,321。此外，数据中具有性别信息的用户数为273,204,518，其中42.5%(116,135,972)为女性用户。数据中群最早创建时间为2005年9月22日，最迟创建时间为2011年3月25日。

2.1.2 网络构建

基于获得的QQ群网络数据，我们构建了三种类型的网络，以全面分析QQ群组社交网络的性质特点。

1. 用户-群超图网络——超图[87,88]是一种由节点和超边构成的图结构，其中每一条超边连接两个及更多节点。如图2.1（a）所示，这里超

图描述的是用户与群之间的关系，节点表示个体用户，超边表示群。举例来说，用户B是群G_1和G_2的成员，因此其通过G_1与用户A相连，同理C和D通过超边G_2相连。本章中，超图使用H来表示。

2. 群网络——如图2.1(b)所示，在我们的数据环境下，群网络表示的是带权重的网络，其中节点表示群，如果两个群至少有一个共同的成员用户，则这两个群存在连边，连边权重值为群与群之间的共同用户数。比如，图2.1(a)中群G_3和G_4有3个共同用户，那么图2.1(b)中G_3和G_4的连边的权重为3。本章中，我们使用字母G表征该群网络。

3. 用户网络——为了研究社会群组的行为，在这里我们使用的用户网络不是通常意义上理解的QQ朋友关系网络，而是通过共同加入的群来确定，如果两个用户同时加入一个群，那么这两个用户之间存在连边，两个用户共同加入的群数作为这条连边的权重。因此，群中所有成员都是全连通的。如图2.1(a)所示，用户C和D同时是*G2*，*G3*以及*G4*的成员。因此图2.1(c)中用户C和D之间的连边的权重为3。本章使用U来表征用户网络。

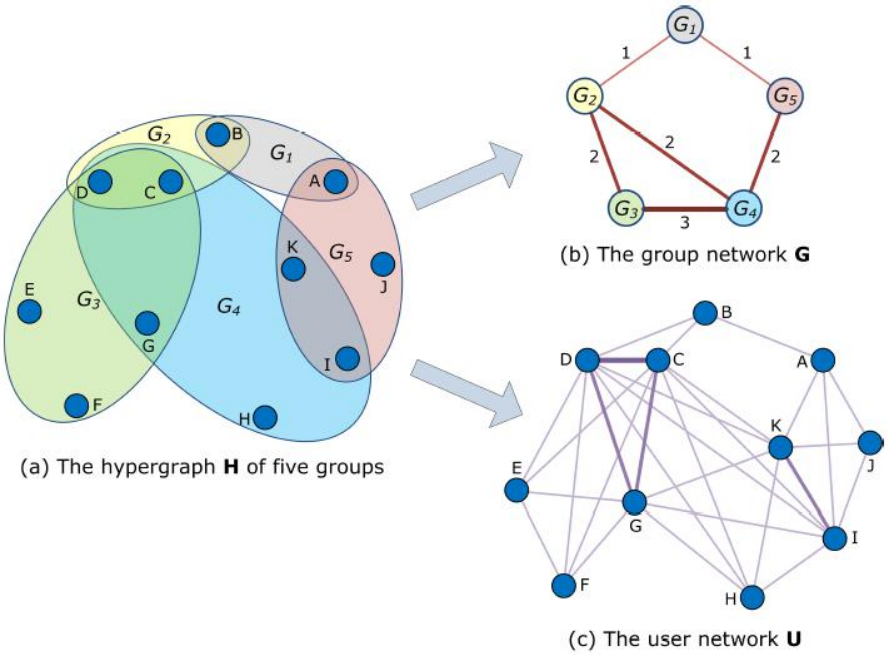

图 2.1 三个网络示意图 (a)用户-群超图网络 H,(b)群网络 G,(c)用户网络 U。该示意图数据由 5 个群(椭圆),11 个用户构成。图(b)和图(c) 中边的粗细表征边权重。

Figure 2.1 Schematic diagram showing (a) the user-group hypergraph ***H***, (b) the group network ***G***, and (c) the user network ***U***. The data is composed of five groups denoted by the colored ellipses in (a) and eleven users. The thickness of edges in (b) and (c) is proportional to the weight on the edges.

2.2 用户–群超图 H 的结构性质

2.2.1 群规模相关性质

社会群组网络中群规模的分布是非常有意思且很重要的一个特征。在本章中,群规模等价于用户-群超图中的超边的秩 s^H。如图 2.2(a)所示,群规模分布 $P(s^H)$在[0,50]范围内表现出缓慢平滑衰减。当群规模超过 50 后,该衰减出现加快的趋势,且该分布在 100,200,500,1000 处出现非连续现象,该现象是由于群规模上限造成的。并且我们发现该断层部分可以使用两个

power-law分布函数来包裹，这两个分布函数的幂指数分别为-3.5和-5.0，如图2.2(a)中虚线所示。这些幂指数相对于其他社交网络上的分布幂指数显得更小，这表明要维持一个具有大量成员的群相对于个人维持大量的朋友来说要更加困难。该结果也同时暗示群规模分布存在更明显异质现象，该现象可能的原因是在大规模群组中要维持紧密关系相对来说比较困难，该因素限制了群规模的增长。另外，如图2.2（b）显示，我们将断层的三部分分别拿出，进行重整化，可以发现不同规模的群的生长机制是一致的，与群规模大小是无关的，也就是说存在某种普适的群生长规律。

此外，直觉上来说，创建时间越早的群有更长时间集聚用户，有更大可能性拥有大规模用户成员。为了研究群规模大小与群创建时间关系，我们计算了同一时间创建的群的平均规模。如图2.2（c）所示，可以看到，平均群规模大小与创建时间几乎不相关，这与我们的猜测截然相反。这表示绝大多数群创建后在较长时间范围内几乎就不再发生明显的规模增长。可能的原因是因为群一旦创建完成，该创群信息会在创建者社交圈内迅速传开，对其感兴趣的用户会很快加入其中，如同学群。另外，还有可能是群保持着动态平衡，即伴随着成员退群及新用户加群的过程。在群推荐算法设计中，需要考虑该信息，更多的去推荐有动态增长、保持动态平衡的群，而较低权重放在几乎一层不变的群。举例来说，高中同学群，一经创建完毕，几乎所有人都已经加入，因此即使这类群表现出来的兴趣与目标用户的兴趣一致，但当将该类群推荐给非该圈子中的目标用户时，目标用户加群的可能性也几乎为0。进一步，我们计算了同一时间创建的群的群规模大小标准差，同样与群创建时间相对独立。此外，我们将接近群规模上

限的群排除在外，即群规模在90-100,180-200,450-500之间的群，发现剩下的群的平均规模分布与原始的数据表现出一致的结果，见图2.2（c）中蓝紫色曲线。该现象与其他社交网络呈现出缓慢增长的结果存在明显差异。

2.2.2 用户加群数相关性质

除群规模大小外，个体用户加群数同样是社交网络中非常重要的指标。在超图环境下，用户加群数分布可以用超度k^H来表示。图2.2(d)显示加群数分布P(k^H)的尾部可以使用幂指数-3.82的幂函数来拟合。尽管该幂指数与其他社交网络相比来说小得多，该power-law衰减表面加入大规模数量群的用户是存在的。不像之前研究发现的社交网络指标会与用户性别有关，我们在该社交网络中发现男性与女性具有相似的分布P(k^H)。

另外，文章使用k^H_{max}来表示群中成员加群数最大的值。我们观测到k^H_{max}和群规模大小呈现出强相关性，如图2.3(a)所示。$<k^H_{max}>$与s^H关系可以使用幂指数为0.54的power-law函数进行拟合。$<k^H_{max}>$随着群规模大小的增长而增长，该现象可以理解为那些活跃的用户一般更多的出现在有大量成员的群。然而，如图2.3(b)显示，群成员平均k^H并没有随着群规模大小增长而增长，这表明不同规模大小的群中活跃成员与非活跃成员的组成结构是相似的。在群推荐算法设计中，可以更多的偏向活跃用户，可以根据用户目前的加群情况来辨别用户的活跃的程度，因为该类别用户的加群偏好更强，有更大可能性加入其感兴趣的群。而非活跃用户，即使遇到其感兴趣的群，但由于本身没有加群意向，不仅使得推荐效果欠佳，而且给用户造成负担。

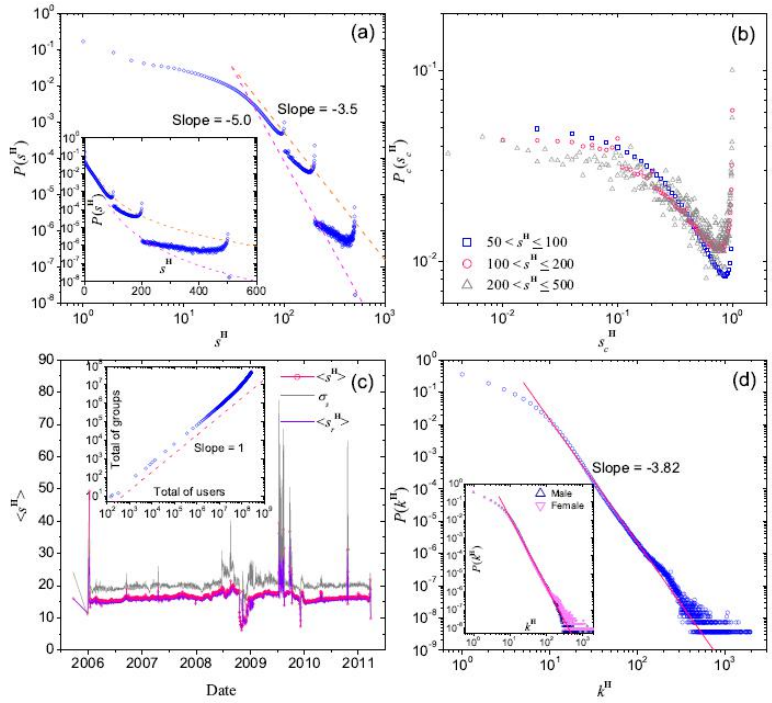

图 2.2 超图 *H* 相关统计指标。(a) P(s^H),群规模大小 s^H,内嵌小图是对其进行 semilog 转换后的结果。图中两条虚线的幂指数分别为−3.5（橘色）和−5.0（紫红色）。（b）经过重整化后的非连续部分，其中 $s^Hc = s^H/\langle s^H\rangle$,$\langle s^H\rangle$为每一部分 s^H 的平均值，Pc(s^H)是相应的重整化后的概率分布。（c）给定具体时间点创建的群规模大小的平均值（粉色）与标准差（灰色）。内嵌图为群总数与群覆盖用户总数之间的关系。（d）P(k^H)表示个体用户加群数分布。内嵌图表示男性与女性的加群数分布。粉色线是幂指数为−3.82 的 power−law 函数。

Figure 2.2 Statistics for the hypergraph *H*. (a) P(s^H), the distribution of group size s^H, with the distribution in semi-log scale shown in the inset. The two dashed lines in show the range of the tail exponent of P(s^H), namely −3.5 (orange) and −5.0 (magenta). (b) The data collapse of the different broken parts on P(s^H) after re-scaling, in which $s^Hc = s^H/\langle s^H\rangle$, here $\langle s^H\rangle$ is the average value of s^H in each section, and Pc(s^H) is the corresponding re-scaled probability. (c) The average (pink) and the standard deviation (grey) of group size given specific date of establishment, and the inset shows the scaling relationship between total of groups and total of users at each date. (d) The distribution P(k^H) of the number of joined group by individual users. P(k^H) for

male and female users are shown in the inset. The pink lines correspond to power-law fits with exponent −3.82.

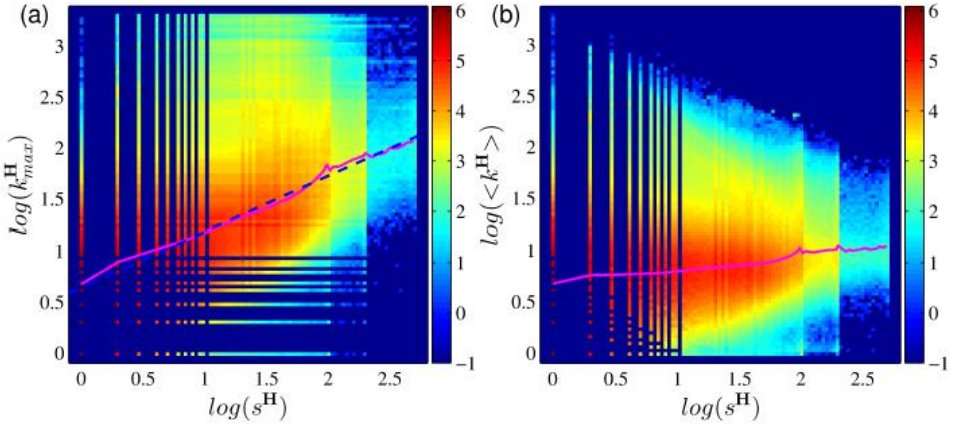

图2.3 群规模和群成员加群数关系热度图。颜色尺度对应于群规模大小出现次数的对数变换值。(a)群中个体成员加群最大值，(b)群中成员加群数平均值。图中粉色曲线是x轴对应的y的平均值。(a)中蓝色虚线可以被斜率0.54的power函数进行拟合。

Figure 2.3 The heat maps showing the correlation between group size and the number of joined group of members. The color scale corresponds to the log-frequency of occurrence between the size of a group and (a) the largest number of group joined by an individual member in the group, and (b) the average number of group joined by the members in the group. The pink lines show the curves on their means along vertical values, and the blue dashed line in (a) shows the fitting power function with slope 0.54.

2.3 群网络 *G* 的结构性质

2.3.1 群网络度相关性质

在本部分，群权重网络描述的是群与群之间非直接的互动行为。如前面网络构建部分所述，当两个群之间存在共同成员用户，则这两个群存在连边，边的权重为两个群共同的用户数。

如图2.4(a)所示，群度k^G分布P(k^G)在小于120范围内呈现幂指数为-0.8的power-law分布，在大于120范围内，呈现幂指数为-2.23的power-law分布。同样地，图2.4(b)显示加权度分布K^G也呈现出两段式power-law分布，加权度K^G分布指的是目标群与所有其他相连的群的共同用户总数，在小于160范围内遵循幂指数为-0.81的power-law分布，在大于160范围内则表现为幂指数-2.33的power-law函数分布。该分布结果表明5800多万个群中绝大多数群度都在100左右范围内，也就是说大部分的群仅仅与相当少量的群存在共同的用户，这反映出群网络上成员集中性表现出局域特性。另一方面，边权重也遵循两阶段power-law分布，见图2.4(c)。这进一步表明两群之间共同兴趣用户绝大多数限制在100左右的量级。这暗示了在群推荐算法设计中，需要考虑群与群共同用户数，用户不可能加入多个具有相同兴趣或者需求的群，因此在推荐的时候需要考虑群的重合性问题。

容易知道，群度与两个因素是相关的，一是群中用户数，一是群中成员所加其他群的总数。图2.5(a)展现了群度k^G与群规模之间的关系s^H，其中群规模与对应规模的群的平均群度<k^G>的关系使用粉色曲线表示。结果显示群度是随着群规模的增长而变大的，这可以理解为更大规模的群中其具有更活跃的用户，也就是说群中出现成员加入不同多群的情况会比较常见。在图2.5(b)中，相似的统计现象也出现在群度与群中成员最大加群数k^H_{max}关系上。现象出现的原因与图2.5(a)中相似，大群往往拥有更多的活跃用户，使得个体成员加群数出现较大值的可能性增大。此外k^G平均值存在较大的过度现象，如图2.5(b)所示，从快速增长到缓慢生长，这表示在群度小于100范围内，群度与k^H_{max}相关性更强。该结果显示，在群处于群度小的时

期，活跃性用户在促进群度增长过程中起着更加重要的角色。

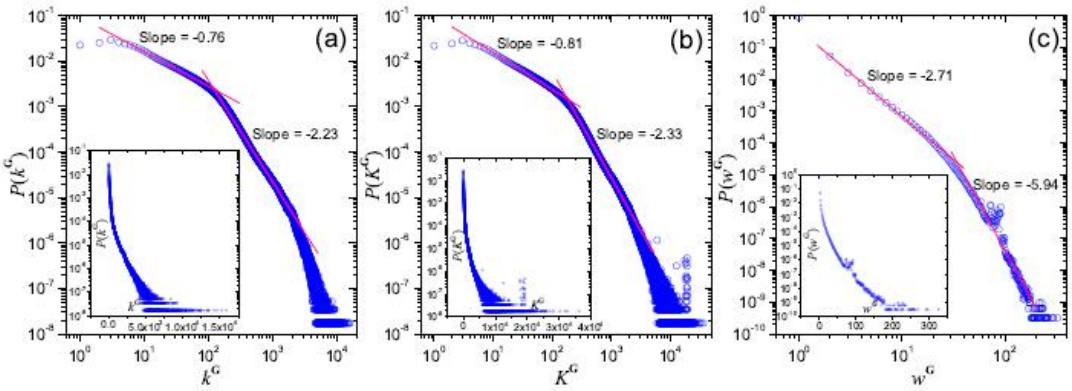

图2.4 群网络G的性质。(a)群度分布P(k^G)，(b)加权度K^G分布P(K^G)，(c)边权重分布P(w^G)。其中内嵌图为对应数据的semi-log尺度形式。

Figure 2.4 Properties of the group network G. The figures show (a) the distribution P(k^G) of group degree, (b) the distribution P(K^G) of weighted group degree K^G of G, and (c) the distribution P(w^G) of edge weight. The insets show the same curves in semi-log scale.

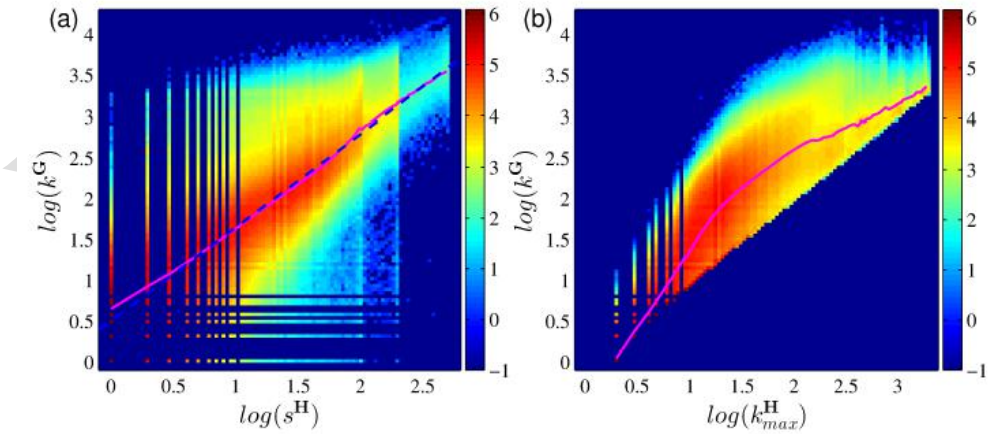

图2.5 k^G and s^H相关关系与k^G and $k^H max$相关关系的热度图。颜色尺度对应数据点出现的频次。粉色线表示相应x轴坐标的y的平均值,(a)中蓝色虚线满足斜率为1.14的power函数分布。

Figure 2.5 The heat maps which show the correlation (a) between k^G and s^H, and (b) between k^G and $k^H max$. The color scale corresponds to the log-frequency of

occurrence. The pink lines show the curves on their means along vertical values, and the blue dashed line in (a) shows the fitting power function with slope 1.14.

2.3.2 群网络小世界现象

此外，我们发现尽管QQ群网络非常稀疏，但是却表现出很强的小世界现象。这与Facebook朋友关系网络的性质是一致的[89,90]。与Facebook相比，其平均度与平均加权度都要稍小于QQ群网络，分别为108.8和133.6。相对于群网络的规模来说，这样的值是非常小的。为了更确切地展现小世界现象，我们随机选择20,000个群对计算他们的网络距离，发现平均距离为3.7，这类似于Facebook网络最近被观察到的四度分离现象[89]。此外，我们也计算了局部簇系数C^G,公式如下：

$$C^G = \frac{2n_T}{k^G(k^G - 1)}, \tag{2-1}$$

对于10,000个随机选择的群，n_T表示的是目标群的邻居群的连边数。在图2.6中，我们展示了群值对(k^G,C^G)频次图。可以看到，C^G与k^G呈现负相关，大致满足幂指数为−0.62的power−law函数分布，与Facebook的情况[89]类似。而平均簇系数为0.35，相对于其他社交网络显得更高。

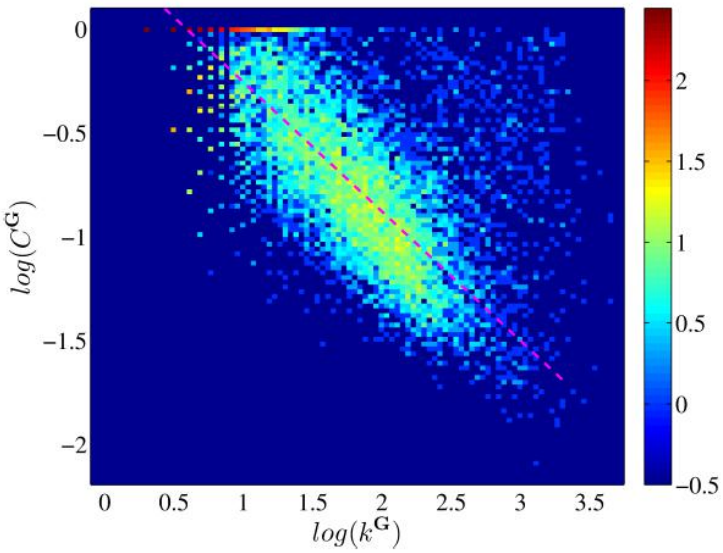

图2.6 局域簇系数C^G与群度k^G之间相关关系的热度图。颜色尺度对应于104个随机抽取的群的(C^G, k^G)值对出现频次。粉色虚线是x轴对应的y的平均值，满足斜率为−0.62的power函数分布。

Figure 2.6 The heat map which shows the correlation between local clustering coefficient C^G and the degree k^G in group network G. The color scale corresponds to the log-frequency of occurrence over 104 randomly sampled groups. The pink dashed line shows the fitting curve with slope −0.62 on the means along vertical values.

2.4 用户网络U的结构性质

我们对加权用户网络做了类似的分析。图2.7展示了用户网络度分布，其尾部满足幂指数为−3.22的power−law函数，整个网络的平均度为135.3。此外，我们检验了性别对用户加群行为的影响，见图2.7底部内嵌图，发现性别因素没有对用户网络度分布产生明显影响。通过对10,000个随机选择的用户对，计算平均网络距离，该平均值接近于4.17，这和Facebook朋友网络中的距离指标是相似的。这些结果表明用户网络同样存在稀疏性和小世界现象，也就是说用户所加的群中的成员很可能来自网络距离较远的用户。在群推荐算法中，需要加以考虑该信息。

通过比较度分布P(k^U)和边权重w^U分布图P(w^U)，我们观察到后者可被衰减函数$P(w^U)=10^{-5.45}[\log(w^U)]^{-7.96}$很好地拟合，该衰减函数慢于power-law衰减函数。这暗示了度大的用户更偏好与其他用户分享群，导致边权重更大。在群推荐中，可以给与度大的用户更多的关注，这符合这些人的行为意图。

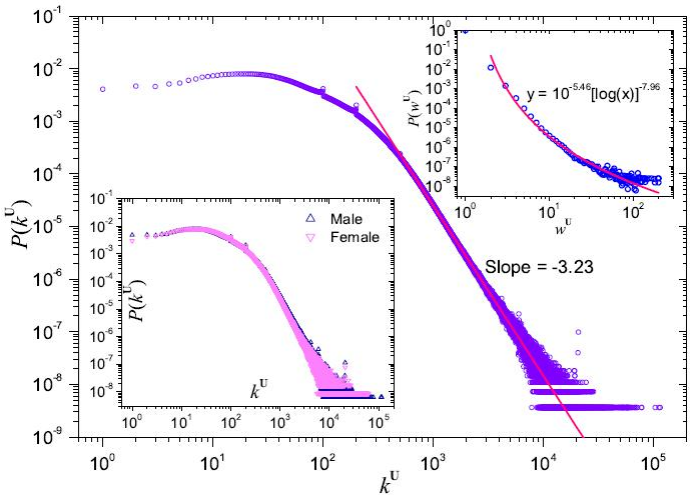

图2.7 用户网络度分布P(k^U)。其中底部内嵌图为男性女性度分布。顶部内嵌图为用户网络边权重w^U分布图P(w^U)。

Figure 2.7 Degree distribution P(k^U) in the user network U. The bottom inset shows the same distribution over male and female users respectively. The top inset shows the distribution P(w^U) of edge weight w^U.

2.5 用户加群行为与年龄、性别相关的性质

文献[91]对社交网络中用户偏好随着年龄变化的关系进行了研究。在该部分，我们同样对QQ群网络数据中用户年龄与其他所加群之间的关系进行了深入的挖掘。

2.5.1 群成员年龄分布性质

本章使用使用a来表示年龄变量，图2.8(a)表示的是个体用户年龄的分

布P(a)和群中成员平均年龄分布$P_G(<a>)$。可以看到，QQ群中成员基本是20岁左右的年轻人。如图2.8(b)所示，我们将四个年龄段的用户分别作了加群数分布，图中显示该分布与年龄有一定的相关关系。40-44年龄段用户加群数分布P(k^H)的衰减在较小范围k^H部分要更快，在较大k^H内稍显平缓。我们进一步发现用户加群数的平均值分布呈现双峰分布形式，分别在15岁和大于65岁年龄阶段出现峰值，这两个时间点对应青少年和老年阶段。当在40岁左右,用户加群数是最少的。这些结果表明相对于40岁左右的中年人，青少年和退休人群在群组社交网络中是相当活跃的。特别的，我们检查了那些年龄较大的人所加入的群，发现基本都是娱乐群，这暗示他们更偏好休闲活动以及社会交往来打发闲暇时间。对于40岁左右的用户来说，他们更倾向于事业和家庭，而较少活跃在QQ群中。因此考虑到年龄的重要性，在群推荐算法设计中，需要更多考虑向青少年群体以及退休老年群体推荐相关潜在兴趣群，同时如果向40岁左右中年人推荐群的时候，更多的偏向与目标用户相关的事业群及家庭孩子成长相关的群。

同时,我们计算每一个群中成员年龄的相对标准差,定义为$c_{va}=\sigma_a/<a>$，其中σ_a表示年龄标准差，$<a>$表示群中成员年龄平均值。图2.8(c)表示的是c_{va}的分布，显示出绝大部分群中成员年龄标准差小于0.1，这表明群成员通常都是年龄相仿的。这可以理解为相似年龄的用户通常具有相似的兴趣，以至于更大可能加入到相似的群组织中。因此在群推荐算法中，可以计算所推荐群中的成员年龄与目标用户年龄的相似性，当然也存在另一种情况，就是群中成员年龄相差很大，这个群很可能是兴趣表现趋同性极强的群，吸引了多样的用户。

另一方面,不同成员平均年龄的群在c_{va}行为表现是存在差异。如图2.8(c)所示，对于群成员平均年龄为14岁的群来说，其c_{va}首先达到一个峰值，这个年龄段对应于青少年时期，另外在平均年龄为33的群，也存在峰值现象，这段时期对应事业成长期的人群。对那些老年人用户所加的群，其c_{va}平均值相对较低。一般来说，c_{va}指标可以被看成是一种衡量群中用户多样性的指标。以上的发现可能意味着青少年和30岁左右的正在事业发展期的人群会显得更加开放，愿意接触更多不同领域不同年龄段的人，一方面与用户的性格有关，另一方面与用户所处的环境相关，青少年时期会接触很多新鲜的事物，比如游戏、学习新的技能等。做事业的会接触各种商业上的合作者、业务伙伴等。不过，在20岁左右，我们发现c_{va}的平均值并不大，这反映出这个年龄段的用户偏好于寻找相似年龄的好友，如大学同学或者男女朋友。在群推荐中，需要考虑到不同年龄段用户的偏好。

在以上平均群规模<s^H>分布中我们观察到的双峰现象，也同时发生在群网络中平均度分布上。如图2.8(d)以及内嵌图显示，s^H和k^6的第一个峰值都出现在年龄19左右，第二个波峰出现在28岁左右。这些结果都表明了用户对群偏好随着年龄的增长，呈现出一种非单调的变化。

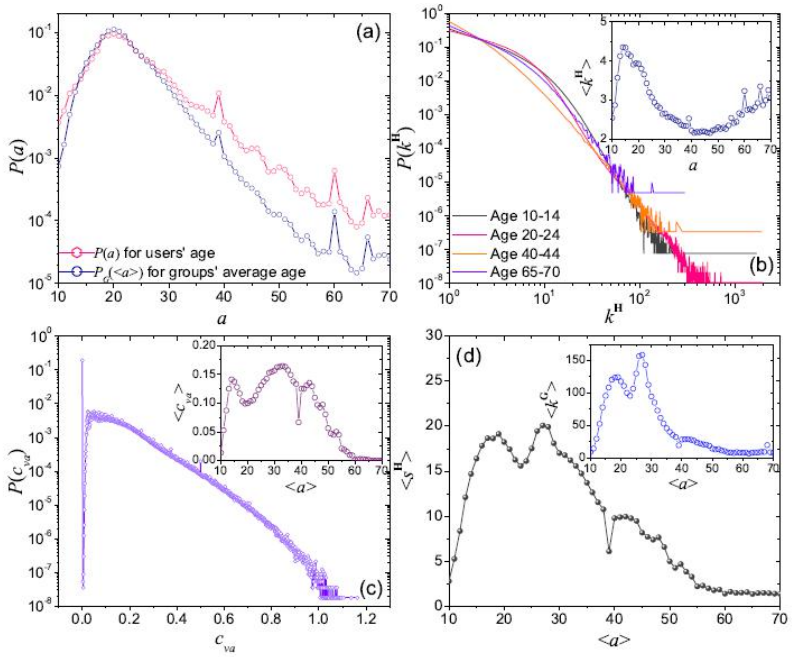

图2.8 群成员与群相关统计指标之间的关系。(a)个体用户年龄分布及群中成员平均年龄分布。(b)不同平均年龄的用户加群数分布。其中内嵌图为不同年龄用户对应所加群数平均值的分布。(c)群中成员年龄相对标准差c_{va}分布$P(c_{va})$以及群成员平均年龄与对应的相对标准差平均值的关系分布。(d)群平均年龄$\langle a\rangle$对应的群的平均群规模大小$\langle s^H\rangle$的分布图。内嵌图为群平均年龄$\langle a\rangle$对应的群的平均群度$\langle k^G\rangle$的分布图。

Figure 2.8 The relation between member age and group characteristics. (a) The distribution of age over individual users and average member age over individual groups. (b) The distribution of the number of joined groups by different age groups. Inset: the average value of of joined group over users at different ages. (c) The distribution $P(c_{va})$ of the coefficient of variation c_{va} for users' age in each group, and the average value $\langle c_{va}\rangle$ as the function of the average age of groups is shown in the inset. (d) The dependence of average group size $\langle s^H\rangle$ on average group member age $\langle a\rangle$. Inset: The dependence of average group degree $\langle k^G\rangle$ on average group member age $\langle a\rangle$.

2.5.2 用户加群偏好与年龄的关系

为了更加全面认识用户加群行为受年龄的影响，我们进行了以下的分析：(1)针对特定的群成员平均年龄，找出该平均年龄对应的群，计算群各项统计指标的平均值；(2)同时以2D空间图的方式展现变量对之间的关系，这可以构成统计指标随着年龄增长发生的演化轨迹。如图2.9(a)，展示了用户网络平均度和群成员年龄相对标准差随年龄增长发生的变化轨迹，结果表明青少年的社交圈子不大，但是圈子中邻居好友的多样性很强，然后慢慢地随着年龄的增长，社交圈子多样性开始降低，但圈子规模逐渐增长。这可能与他们学业相关，因为上初中、高中、大学接触的人越来越多，使得圈子规模变大，但是由于接触的都是相似年龄的用户，因此在好友多样性方面呈现降低趋势。进一步随着青年人步入社会，开展工作，好友多样性又呈现出上升趋势，这是因为在工作中会接触到形形色色的人，同事、合作伙伴等，进一步由于顾及工作、家庭等原因，使得其他的娱乐、之前好友交往等活动有所下降，导致退出所加的群，比如游戏群、旅游群等，缩小了社交圈子。这和现实中的情况是一致的，工作之后我们所活跃的圈子呈现缩小的趋势。这些结果和之前研究发现的成熟期前后存在过渡现象是保持一致的。

以上的处理是针对所有用户的。在图2.9(b)中，我们考虑性别因素，分别对男性与女性用户做了相似的轨迹分析。可以看到，女性用户随着年龄的增长，社交圈子规模的扩增速度要快于男性用户，而且存在更早的转折点，在16岁，路径演化趋势就发生了转变，而男性用户是在20岁左右。通常来说，女性用户比男性用户更早成熟，这也体现在了路径转折点上。另外，相比于男性用户，女性在成熟年龄期间，社交圈子更小，这与女性

用户特点有关，女性一般会偏向于活跃在熟悉紧密的圈子，以及由于成立家庭和更多为家庭事务所操劳进而限制了她们的社交圈子的扩大。

进一步，我们研究了群规模大小 s^H 和用户加群数 k^H 随着年龄的变化的情况。如图 2.9(c)所示，该路径明显的区分点，在年龄为 15 岁左右。该区分点将整个行为阶段分为了成熟期前后两个阶段。在成熟期前，用户偏好加入更多的群，而且群大小一般较小，然而到了成熟期及后期，用户显示出较弱的加群偏好，倾向于加入少量群，但群规模一般较大。图 2.9(d)展示了分别对应于男性和女性的相似变量路径，可以看到，女性用户更早地发生转变趋势，进入成熟期，这与图 2.9(b)中观测到的现象是一致的，并且女性用户在成熟期，所加的群都比较少，这与之前观察到她们的社交圈子规模减少的现象也是类似的。

除此之外，我们还研究了 QQ 用户的邻居好友多样性和邻居好友的度与年龄增长的关系，年龄多样性衡量指标使用 $\langle c_{va}\rangle_f$ 表示，邻居数多样性使用 $\langle c_{vd}\rangle_f$ 表示，尖括号代表其为平均值类型。如图 2.9(e)所示，15–23 岁之间的用户，也就是处于成熟前期用户，通常具有成员年龄相似的朋友网络，这与他们在校学习接触到的人有关。然而，在这段时期内，他们的邻居好友的度的多样性随着年龄增长在逐渐增大。当经历完过渡期，25–35 岁之间的用户具有更宽范围年龄好友的朋友圈，但这些好友的度基本相近。对于年纪更大的用户来说，他们在邻居好友多样性及好友的度两个指标上都比较中等。图 2.9(f)则是分别对男性和女性进行的相似变量路径分析。

以上结果表明用户在 20 岁之前活跃在不同的社交圈子中。当他们开始接收大学教育后，活动范围开始有所下降，逐步演化为和大学同龄人进行

社会交往。从图2.9中可以观察到，年龄25是一个普遍的分界点。过了25岁用户基本进入了成熟期，其社交圈子逐步趋向稳定，这段时期对应于用户从学生时代向工作时代的演化。在成熟时期，用户更偏好于加入多样性强的社群。另外，我们同时也注意到女性用户相对于男性用户要更早地进入成熟期，这与她们的社会角色可能存在关系。

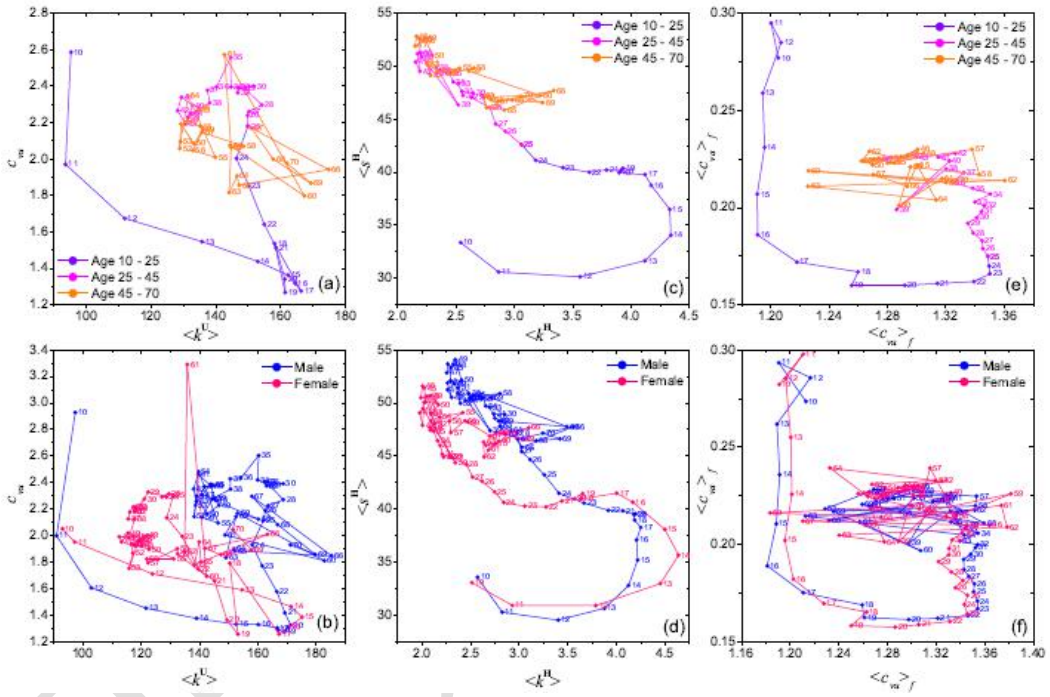

图2.9 平均指标随年龄变化轨迹图。(a)X轴表示用户网络中每一个年龄对应的用户平均度，Y轴表示每一个年龄对应的用户度的相对标准差。(b)图(a)中变量路径在性别上的差异。(c)X轴表示个人用户加群平均数，Y轴表示所加入群的群规模平均数。(d)图(c)中路径在性别上的差异。(e)X轴表示用户网络中用户邻居度的相对标准差，Y轴表示用户网络中用户邻居年龄的相对标准差。(f)图(e)中路径在性别上的差异。

Figure 2.9 The paths of changes of two averaged variables along with age. (a) X-axis: the averaged value $\langle k^U \rangle$ of the degree k^U of user network U for users in each

age, Y-axis: the coefficient of variation c_{vu} of k^U for each age. (b) gender differences on the age trail in panel (a). (c) Horizontal axis: the average number of joined groups by individual users; Vertical axis: the average value size of the joined groups. (d) The same path in (c) by averaging only male and female users respectively. (e) Horizontal axis: the average coefficient of variation $\langle c_{vu} \rangle_f$ among the degree of neighbors of a user in the user network U, Vertical axis: the average coefficient of variation $\langle c_{va} \rangle_f$ among the age of neighbors of a user in the user network U. (f) The same path in (e) by averaging only male and female users respectively. The labels close to each data point corresponds to the value of age, and the different colors in (a), (c) and (d) respectively show the data points in three different age stages.

2.6 社群生长模型

通过前述实证内容，我们发现QQ群大致可以分为两种类型，一种群是自创建之后短时间内群规模大小就增长到稳定水平，之后不再发生明显的增长变化，这类群一般是同学群、同事群等，与用户实际线下社交关系紧密联系；另一类群则是随着建群时间增长，群规模会有一个明显的逐渐增加的过程，这类群一般是基于成员共同爱好自发产生并逐步吸引相似兴趣的用户，产生集聚效应，此类群中成员的社交关系与线下关系不一定对应。我们将目标用户与其好友处于同一组织机构，或者具有相同的爱好，都视其拥有共同兴趣。另外从实证分析观察到群网络及用户网络都存在小世界现象，那么也就是说朋友关系网络上用户即使相隔较远也会由于存在相似兴趣而加入同一个群。由此本章提出了一种类渗流过程的兴趣扩散模型，基于QQ朋友关系网络，对QQ群网络中社会群组结构的产生及生长机制进行研究。

2.6.1 模型变量描述

在本模型中，除了考虑加群行为中的兴趣因素，还需要注意其他结群相关的因素：(1)用户的精力是有限的，正常的用户不可能同时与数量庞大的直接邻居朋友产生结群行为，当邻居朋友数量很多时，用户只会与其中的某些好

友产生结群行为；(2)用户加群意向存在差异，从实证数据分析可以看到，用户加群数分布呈现异质性，大部分用户只会加入少量群，同时也存在加入大量群的用户；(3)群规模的限制，群成员数由于存在规模上限不可能无限制增加。

我们为模型中每个用户都定义了一个N维的二值随机兴趣向量$H=(h_1, h_2,\cdots, h_N)^T$，$h_i=0$ 或 1，i=1,2,…N。其中 H 的每个维度都代表了一类兴趣，也就是说维度值为 1 时表示用户具有该维度表征的兴趣，0 则代表用户没有这方面的兴趣。在本部分模型实验中，N 设置为定值 10。此外，每一个用户都被赋予一个兴趣构建概率 $P_v\in(0, \alpha]$，$0<\alpha\leqslant1$，用于相应用户的随机兴趣向量的构建。由于用户的兴趣范围有大有小，存在差异，因此我们在设置用户兴趣构建概率时，需要保证用户兴趣构建概率呈现出该特点。兴趣广泛的用户其向量中会有多个维度的值为 1，这种用户的兴趣构建概率 P_v 值一般会很接近 α。有一种情况还需要考虑到，就是如果用户兴趣向量中所有维度的值全为 0，当发生这种情况时，我们将随机选取该类用户的兴趣向量某一维赋值为 1，即避免出现在兴趣扩展过程中，由于用户不存在任何兴趣而存在扩展死角。

模型规则中,每个用户都能够基于那些维度值为 1 的兴趣点进行群的创建。也就是说，如果一个用户有 m 个兴趣点则可创建 m 个群。同样的，模型需要考虑到不同用户存在加群倾向差异性的事实，加大量群和加少量群的用户同时存在，因此模型引入加群偏好概率 $P_{join}\in(0, \beta]$，$0<\beta\leqslant1$。P_{join} 的值越大表示用户偏好于更强的加群倾向，当遇到感兴趣的群时，偏好加入。这里规定，模型开始阶段每一个用户即被赋予相应的概率值 P_{join}。在模型中，社群规模演化过程是沿着 QQ 朋友关系网络路径进行的，但如图 2.10 所示，朋友关系网的度分布呈现出 3 段式幂律分布特点，异质性非常明显，网络中大部分用户只有少量的邻居朋友，而少部分用户则拥有数量庞大的直接朋友。由于受限于用户精力等原因，其不可能同时与大量直接邻居发生结群行为，而且 Dunbars 等人研究表明用户最多可以维持少量的紧密好友[92],因此在模型的社群生长过程中我们加入对这层信息的考量，在与直接好友发生结群行为时，伴随着互

动好友人数衰减计算，以维持结群行为发生在一定数量的直接好友中的假设，保证社群生长的平滑性。如果不引入该衰减函数，将会出现模型最终结果中群大小分布发生断层现象，也就是一部分处于较小范围，另一部分处于大规模数值范围，中间存在明显分界线。因此模型规则设置为：当用户与少量相邻朋友数发生结群行为时，用户能够与所有邻居发生结群行为；随着邻居数的增大，由于时间和精力的限制，用户仅能与一定数量的朋友发生结群行为。该衰减函数定义为“漂移幂律”形式[93]：

$$f_d = c(x + K)^{-a}, \tag{2-2}$$

其中K和a是可调参数，K是对幂律分布偏移量的测度，K越小，函数越接近幂律分布，当K为0时就变为幂律分布。c是常数。x是相邻朋友数。在本章中，我们对相关变量做如下设置：c = 100，K = 39，a = 1.4。

对用户网络边权重及群网络边权重的实证分析发现很多用户都会加入相同的一些群，我们猜测这可能是由于某些用户具有很强的信息分享精神及开放的态度，倾向于向他的好友推广其加入的群以使得更多的用户加入他所在的群，这种情形一般出现在具有很紧密关系的线下朋友之间。一群用户同时加入多个相同的群，会导致群群之间的用户重叠度变大，即使得群网络中边权重增大。根据以上讨论，在模型中，每一个用户都被赋予了协同加群概率 $P_c \in (0, \gamma]$，$0 < \gamma \leqslant 1$，用以描述用户协同加群的倾向。基于此，当用户 u_1 决定是否加入由用户 u_2 创建的群时，其加群概率 P_{join} 修正为：

$$P'_{join} = \begin{cases} P_{join} + P_c, & P_{join} + P_c \leq 1 \\ 1, & P_{join} + P_c > 1 \end{cases}, \tag{2-3}$$

其中，P_{join} 是用户 u_1 的原始加群概率，P_c 是用户 u_2 的协同加群概率，而 P'_{join} 是 u_1 的修正加群概率。P_c 越大，则表示 u_2 的邻居加入同样由 u_2 创建的群的概率越大。

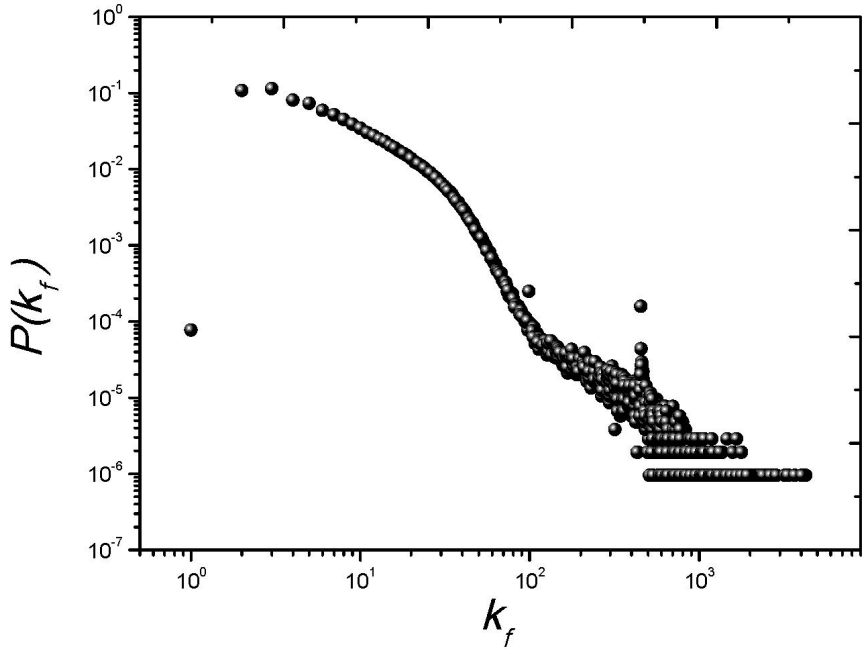

图2.10 QQ朋友关系网络实际数据的度分布

Figure 2.10 The degree distribution of real QQ friend-network

2.6.2 模型模拟过程

将以上讨论的参数变量融入到模型，并基于QQ用户朋友关系网络的实际数据进行该类渗流机制兴趣扩散模型的实验，通过实验模拟用户结群行为及社群生长演化过程，验证我们所提出的假设，并挖掘出社群生长的基本机制。该QQ朋友关系数据共包含1,052,129个用户节点和8,022,535条朋友关系边，数据集来自于同一个城市的用户，另外，我们也计算了其平均簇系数和平均距离，分别为0.609和4.167，这显示出该网络极强的小世界现象。

图2.11用示意图的方式描述了以目标用户u_1创建的群G在其朋友关系网上的演化过程。图中正方形小方块表示模型中QQ用户节点，节点之间的连边表示用户之间的朋友关系。图中每个用户节点的兴趣向量都只有两个维度的值为1，其余为0，使用不同图案的矩形框表示不同的兴趣。浅色底椭圆表示群G，该群是用户u_1基于其右侧部分的第i维兴趣创建的群。图中的虚线圈表示群G扩张的边界。用户u_1初始阶段基于第i维兴趣创建群G，这一设定规定

了加入该群中的成员用户必须表现出第 i 维兴趣，也就是兴趣向量第 i 维度的值为 1，该维度兴趣也可以称为 G 的群兴趣，此时群 G 的用户集为 $S=\{u_1\}$。基于之前讨论的变量因素控制，群 G 的演化迭代步骤可以表述如下：

(1)扩散之前首先考虑邻居数衰减函数 f_d，群 G 首先向用户 u_1 的直接邻居进行扩散，每一个邻居首先需要确定是否参与结群行为。过程为首先对每一个邻居产生随机概率 $P_d \in [0, 1)$，若 $P_d < f_d$，则将该邻居加入初始用户集 S_1，也就是能够产生结群行为的邻居集合，本示意图显示通过处理得到候选用户集 $S_1=\{u_3,u_4,u_7,u_{11}\}$。

(2)从候选用户集 S_1 中，我们进一步选取具有第 i 维兴趣的用户构建候选结群用户集 S_2，换句话说，S_2 中的用户都必须表现出第 i 维兴趣。

(3)S_2 候选用户集中，我们对每一个用户逐个修正其加群概率，再进行加群判别，具体步骤为产生随机数 $P_r \in [0, 1)$，当满足 $P_r < P'_{join}$，那么对应用户将选择加入群 G。举例来说这里用户 u_{11} 尽管表现出第 i 维兴趣，不过依然以一定的概率没有加入群 G。

(4)完成目标用户 u_1 一轮的直接邻居用户结群行为后，此时群 G 当前成员用户集为 $S=\{u_1,u_3,u_4,u_7\}$，下一轮将基于该批新加群的用户进一步扩张群 G。如图 2.11 中用户 u_3，下一步需要考虑其直接邻居 u_2、u_5，同理对 u_4、u_7 的直接邻居进行加群判断。

在群扩张过程中有几点注意的地方，一是当前群扩张过程不考虑之前群扩张过程中已经进行过加群判断处理的用户，但这部分用户仍参与扩散邻居数衰减函数的处理。而是重复以上步骤，直至达到扩展边界为止。这里的扩展边界指：(a)群规模达到上限 2000；(b)没有进一步可扩展的候选朋友集。三是模型结果计算中，如果多个群同时具有相同的群兴趣以及同批群成员，那么认定这批群为重复群，只需要保留唯一一个，进行去重处理，因为同一批用户基于同样的兴趣建立多个群是不符合事实的。

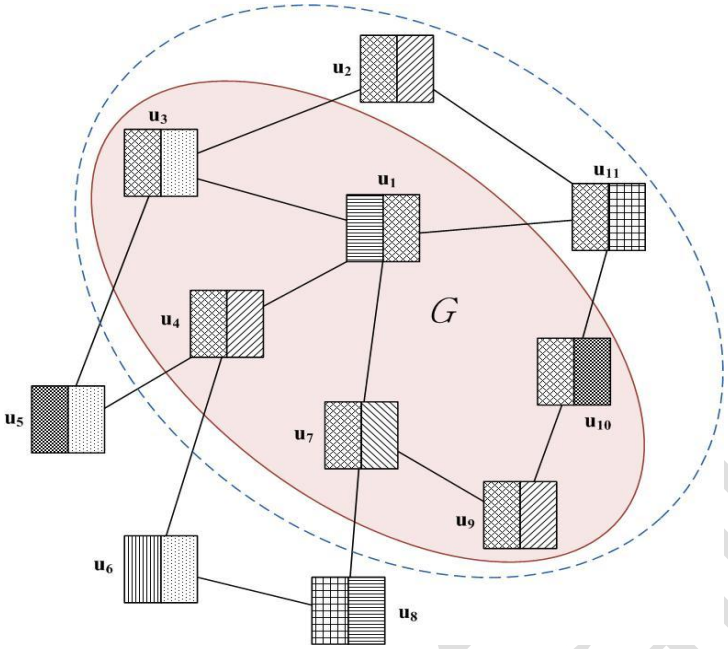

图 2.11 运用兴趣扩散模型基于朋友关系网的群生长过程示意图

Figure 2.11 Illustration of group growing process based on friend-network with the interest diffusion model.

2.6.3 模型结果及分析

通过在实际QQ朋友关系数据上进行实验得到模型结果，即模型中用户加群关系数据，以用户–群的形式保存。采用2.1部分中构建网络的形式，我们对模型结果数据也进行了类似的构建，并计算了相关统计性质，而且为了验证模型的准确性，将模型结果统计指标与之前实际的群关联网络的统计特性进行了差异比较。本章主要衡量四种分布性质：（1）用户加群数分布$p(k_H)$，其中用户加群数使用符号k_H来表征，表示一个用户所加的群总数；（2）节点度分布$p(k_G)$，其中节点度使用符号K_G表示，也就是与该节点（即群）相连接的群数量；（3）加权节点度分布$p(k_{GW})$，其中加权节点度使用符号k_{GW}表示，反映一个群与其他所有群的共同成员总数；（4）边权重分布$p(w_G)$，其中边权重使用符号w_G表示，即两个群之间的共同成员数目。如图2.12所示，大坐标系内散点表示模型结果，小坐标系内散点表示实

证结果。（a）为用户加群数的分布p(k_H)；（b）为群网络的度分布p(k_G)；（c）为群网络的加权度分布p(k_{GW})；（d）为群网络的边权重分布p(w_G)。该实验中模型参数α、β、γ分别为0.19、0.195、1。可以看到模型得到的用户加群数分布以及群网络在节点度、加权节点度以及边权重方面呈现出的幂律分布与实际数据是一致的。在群网络度和加权度分布的前面部分，模型数据中小度群的比例与实证数据相比偏高，该偏差引起的原因可能是模型没有考虑活跃用户的退群行为以及群的人为解散，当群成员中活跃用户数较少时，由于群中信息交流等行为几乎消失殆尽，维持一个群的必要性会降低，大量用户往往会退群或者造成群解散的结果。总体来说，模型所得到的结果与实证数据基本吻合，表明所提出的基于渗流机制的兴趣驱动的社群生长机制可以有效解释实际的社群生长的基本原理。

进一步，我们对主要参数对模型结果的影响进行了研究，分别是：用户兴趣构建概率上限α（如图2.13，空心圆、五角星以及三角形分别表示α = 0.15，α = 0.19以及α = 0.25下的模型结果。（a）用户加群数的分布p(k_H)；（b）群网络的度分布p(k_G)；（c）群网络的加权度分布p(k_{GW})；（d）群网络的边权重分布p(w_G)。该实验中模型参数β、γ分别为0.195、1。用户加群概率上限β（如图2.14，空心圆、五角星以及三角形分别表示β=0.05，β=0.195以及β=0.5下的模型结果。（a）用户加群数的分布p(k_H)；（b）群网络的度分布p(k_G)；（c）群网络的加权度分布p(k_{GW})；（d）群网络的边权重分布p(w_G)。该实验中模型参数α、γ分别为0.19、1。以及用户协同加群概率上限γ（如图2.15，空心圆、五角星以及三角形分别表示γ=0.8，γ=0.9以及γ=1下的模型结果。（a）用户加群数的分布p(k_H)；（b）群网络的度分布p(k_G)；（c）群网络的加权度分布p(k_{GW})；（d）群网络的边权重分布p(w_G)。该实验中模型参数α、β分别为0.19、0.195。

我们发现，α、β、γ的概率上限都会对用户加群数分布产生影响，随着该上限值的提高，加群数分布头部出现降低趋势，变量α造成的分布右偏现象最为明显，这是因为随着α的提高，用户平均兴趣范围也随之增加，直接导致用户能对更多的群产生结群行为。此外，α、β、γ对群权重网络的节点度分布与加群节点度分布产生一定影响，随着上限值的提高，大度节点以及大的加权度节点明显增多。除此之外，我们也注意到β对边权重分布结果的影响以及范围要弱于α和γ产生的影响。随着α和γ上限值的增加，网络中权重值大的边的比例明显增加。

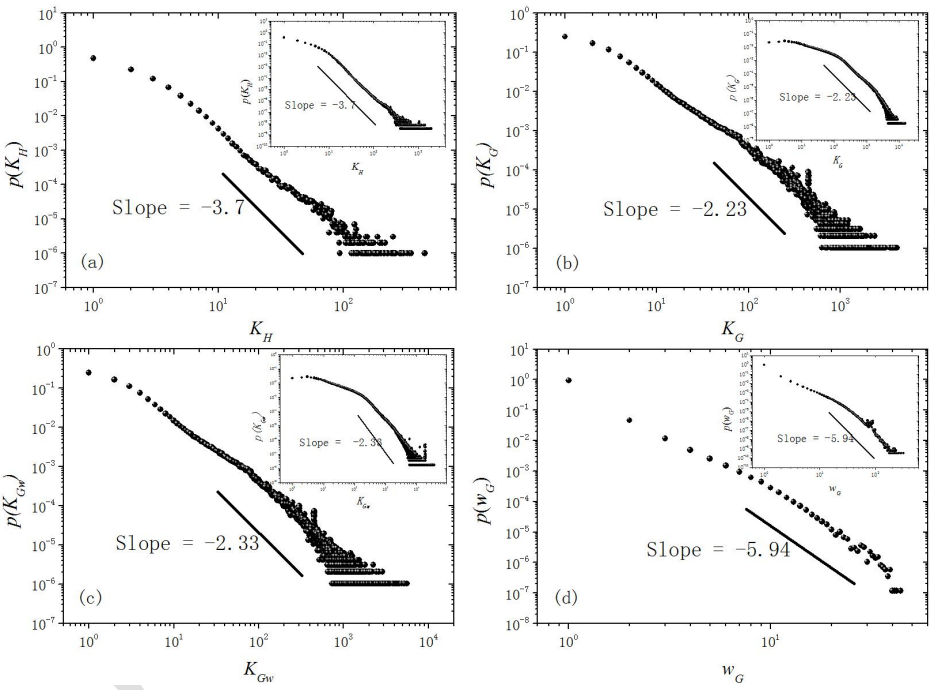

图2.12 模型结果与实证结果在四种统计指标上的对比

Figure 2.12 The comparison between model results and the empirical statistics on four metrics

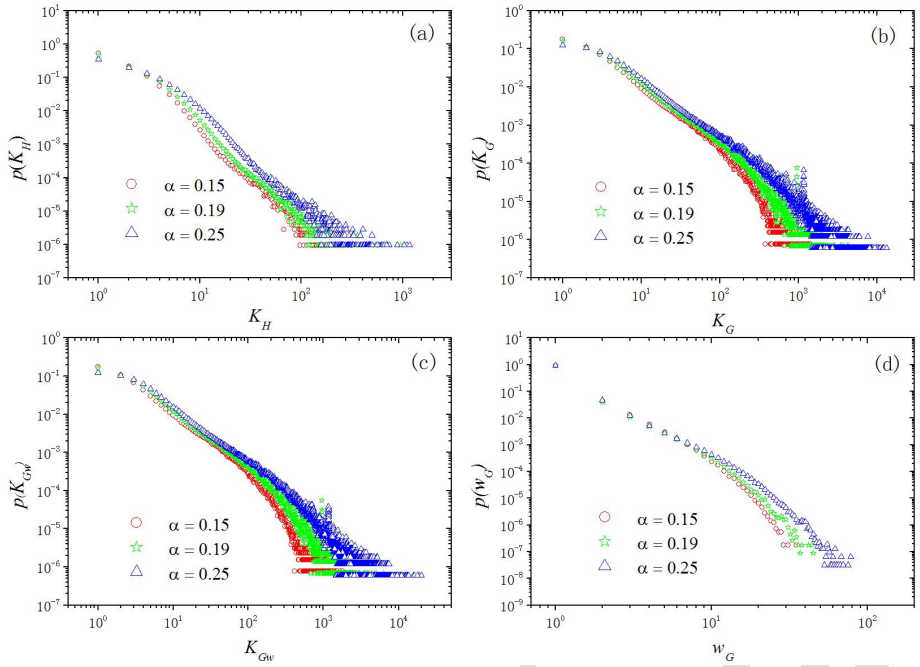

图 2.13 不同兴趣构建概率上限下的模型结果

Figure 2.13 Performance of different values of the limit of interest vector construction probability α

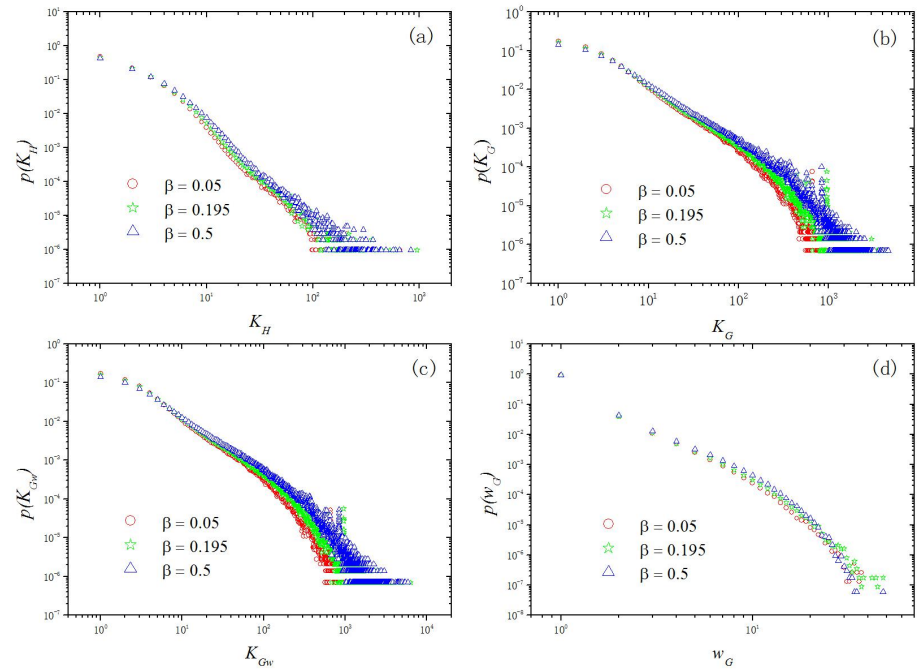

图 2.14 不同加群概率上限下的模型结果

Figure 2.14 Performance of different values of the limit of Group-join probability β

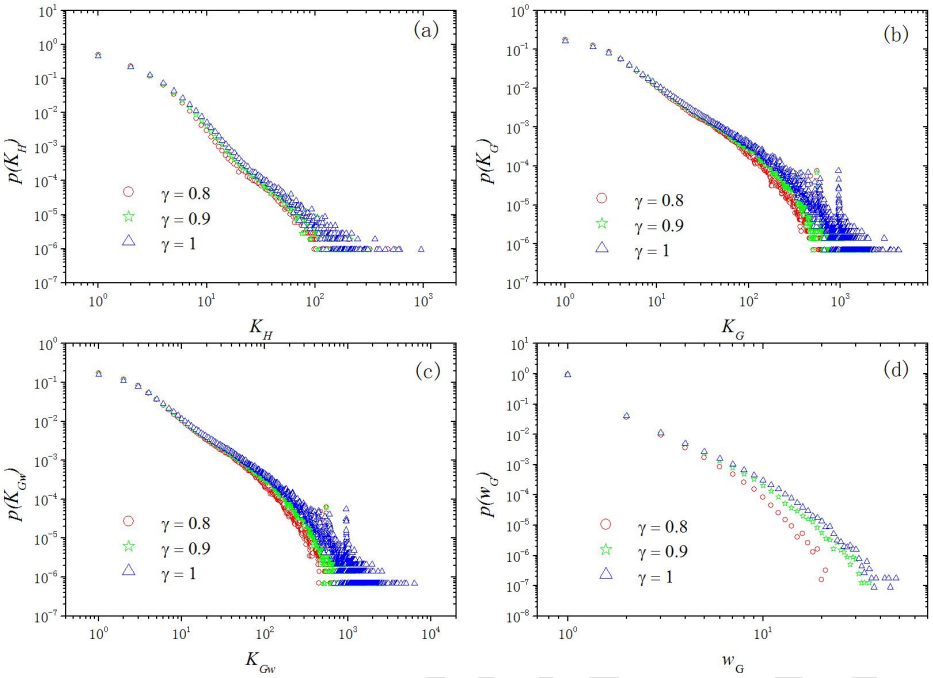

图 2.15 不同协同加群概率上限下的模型结果

Figure 2.15 Performance of different values of the limit of Collaborative Group-join probability γ

2.7 本章小结

本研究试图对明确定义群结构的社交网络加深认知与理解，以缓解当前群推荐算法存在的精度差等问题，为进一步设计更好的群推荐算法提供一定的理论基础。本章从实证分析及模型模拟两个角度对以 QQ 群网络为代表的社群网络进行了深入细致的研究，具体从群规模分布、群规模演化与群创建时间的关系、用户加群数分布、群中用户最大加群数和群成员加群总数与群规模的关系、用户权重网络度分布、群度与群中成员用户数以及群中成员所加群的关系、群权重网络中加权度分布及边权重分布、群度与群中成员最大加群数关系、网络小世界性质、局域簇系数与群度关系、网络中及群组中用户年龄分布、用户加群行为与年龄和性别之间的关系、用户兴趣偏好变化与年龄的关系以及社群生长机制探索等多角度研究了该在线

社群网络。从中我们发现了一些在群推荐算法设计中需要特别注意的点，这些方面可能会对群推荐效果产生明显影响：

（1）根据群规模分布发现，网络中要维持大规模群组是困难的，这暗示在做推荐过程中，如果将规模较大但活跃低的群推荐给用户，很容易造成用户流失。因此在推荐这类群的时候，需要更多将群活跃度因素纳入算法体系中。

（2）对群规模与创群时间关系的研究，发现群大致存在两种形式，一种是自创建以后短时间内即达到稳定状态，之后很少会发生明显变化，如同学群。一种是随着建群后时间的增长，群规模逐渐扩大。举例来说，大学同学群，一经创建完毕，几乎班级中所有同学都已经加入，因此即使这类群表现出来的兴趣与目标用户的兴趣一致，但当将该类群推荐给非该圈子中的目标用户时，成功加入群的可能性几乎为0，这类群的推荐需要更多基于线下社交关系。而后一种类的群，更容易被感兴趣的目标用户接纳。

（3）研究发现群中成员类型组成是相似的，即群中一般都会包含活跃用户和非活跃用户，非活跃用户对于群的维持与增长所起的效果并不明显。在群推荐算法设计中，可以更多的偏向活跃用户，根据用户目前的加群数及群多样性来辨别用户的活跃的程度，因为活跃用户的加群偏好更强，有更大可能性加入其感兴趣的群。而非活跃用户，即使遇到其感兴趣的群，但本身没有加群意向，不但使得推荐效果欠佳，而且给用户造成负担。

（4）通过对群网络中边权重分布分析，进一步表明两群之间共同兴趣用户绝大多数限制在100左右的量级。这暗示了在群推荐算法设计中，需要考虑群与群共同用户数，用户不可能加入多个具有相同兴趣或者需求的群，

因此在推荐的时候需要考虑群的重合性问题。

（5）我们发现尽管QQ群网络非常稀疏，但是却表现出很强的小世界现象。这表明群中用户并不局限在局域范围内，而是可能存在来自网络距离较远的用户。

（6）对用户权重网络的边权重分布的研究，表明度大的用户更偏好于与其他用户分享群。在群推荐中，可以给与度大的用户更多的关注，在为度大用户的好友推荐群的时候，可以更多偏重于该度大用户已经加入的群，这符合这些人的行为意图。

（7）考虑到年龄的重要性，在群推荐算法设计中，需要更多考虑向青少年群体以及退休老年群体推荐相关潜在兴趣群，同时如果向40岁左右中年人推荐群的时候，更多的偏向于与目标用户相关的事业群及家庭孩子成长相关的群。

（8）在群推荐算法中，可以计算所推荐群中的成员年龄与目标用户年龄的相似性，20岁左右的用户倾向和相似年龄用户交往。当然也存在另一种情况，就是群中成员年龄相差很大，这样的群很可能是兴趣表现趋同性极强的群，吸引了多样的用户，在推荐时，更多考量目标用户与该群的兴趣相似性。

（9）研究结果显示在20岁之前用户普遍活跃在不同的社交圈子中。而当他们开始接受大学教育后，活动范围开始有所下降，逐步演化为和大学同龄人进行社会交往。年龄25是一个普遍的分界点。过了25岁基本进入了成熟期，用户社交网路逐步趋向稳定，这对应于用户从学生时代向工作时代的演化。在成熟时期，用户更偏好于加入多样性强的社群。另外，我们

同时也注意到女性用户相对于男性用户要更早地进入成熟期，这与她们的社会角色可能存在关系。

（10）我们提出基于好友共同兴趣的类渗流的扩散过程来揭示群组社交网络上社群的生长机制。模型结果与实证结果相符，暗示这种基于共同兴趣的社群生长机制在实际的社群生长中扮演着重要角色。这一机制的揭示，对于社群涌现和生长的预测，也有着相对重要的意义，例如可以通过对共同兴趣的判断来进行社群推荐或者预知潜在的社群关系等。

以上结论显示，对于社群网络的分析与研究，确实能够帮助我们加深对其认知，而且可以为设计更优的推荐算法提供可靠思路。

3 微博用户关注网络的性质分析及关注关系预测

链路预测在很多应用领域都取得了成功，如在生物网络、社交网络[94-99]等，因此目前已经成为非常热门的研究点，吸引了很多学者的注意。链路预测问题主要是在基于当前能够观测到的链路信息基础上，预测未来最可能存在的边。目前来看，一些链路预测算法确实已经带来了实质效果。在生物领域，通过使用有效的链路预测算法，研究者在研究蛋白和蛋白互作时不需要盲目地进行所有蛋白和其他蛋白两两互作研究，而是可以只针对那些最可能产生关系的蛋白对进行研究，这大大节省了时间精力和财力[100]。另外，在电子商务领域，链路预测算法被证实能够用于协同过滤推荐，从而得到比原始的协同过滤算法更好的推荐效果[101]。

本篇文章主要关注链路预测算法在社交网络中的应用。更确切地讲，是研究twitter 和 weibo 网络上的用户关注预测。社交网络上的用户关注边预测是非常重要的研究点，一个好的链路预测算法能够被用于精准的潜在好友推荐，用户社交圈子构建，以此来提高用户在社交平台上的活跃性及粘性。众所周知，保持用户高活跃度是社交网络平台可持续发展和成功的必要前提。目前，学术界已经研究设计出很多的链路预测算法。在机器学习领域，学者采用监督学习框架来应对社交关系的预测[102-104]。不过这类算法要面临非平衡训练集的问题[105]。在网络科学领域，也已经有相当多利用节点及边信息来做链路预测的算法[16]。一般来说，两个节点，或者说两个用户如果有很多共同的好友，那么这两个节点相互产生连接的可能性会更高，这也是复杂网络中基于局域相似性指标的核心思想。比如common neighbor[106]算法通过计算两个节点之间共同的邻居节点数，来对可能存在的边按照该邻居数来降序排列，排在最前面的节点对越可能在未来进行连边。

除此之外，还有 Katz、Hitting time 等考虑网络全局信息的算法。然而，这一系列的算法并不是专门针对 twitter 这类社交网络设计的，而是针对其他诸如科学家合作网、生物蛋白质网、食物链网及美国航空网等设计的。因此直接应用这类现有算法到社交网络的实际应用，会存在算法适用性问题，其效果可能并不是非常理想。首先，不像文献[16]中研究的网络，只有网络拓扑结构的信息，在线社交网络同时具有文本信息和网络结构信息。在目前的复杂网络算法中，还没有能够有效使用文本信息的算法。然而像 twitter 这类网络上用户发表的微博信息能够显性表现出用户的兴趣偏好。因此，我们试图提出一种新的算法，该算法可以在低计算复杂度情况下，同时利用文本内容数据和社交网络信息来做链路预测。

通过对 twitter 和 weibo 数据的实证分析发现，存在关注关系的用户之间通常具有很强的相似兴趣。并且相关信息传播结构对潜在被关注用户对目标用户的可见性大大增强。因此，很自然的，我们使用目标用户的好友的兴趣特征作为目标用户的协同兴趣来构建完整的目标用户兴趣。此外，有研究[107]证明用户普遍偏好在社交平台上发布带有自己口味的内容，这使得我们利用微博内容去挖掘用户兴趣是可行的。基于以上的实证发现，以及受到相关工作[94,107,108]的启发，我们提出了一种叫做基于主题兴趣相似性的最大化偏好(Maximum Preference on Interest Similarity(MPIS))的新算法，通过利用文本信息和网络结构信息来进行更加准确的链路预测。本章中，我们使用 AUC、precision、recall 三种衡量指标来验证算法的有效性，结果表明所提出的新算法远远好于现有的基准算法。

3.1 数据

3.1.1 数据集

在本章中，我们使用两个数据集来验证算法的有效性，分别为新浪微博和twitter数据。新浪微博是中国大陆地区使用用户最多规模最大的微博产品，类似于twitter，用户可以发布长度为140 的文本信息、图片信息等，同时用户可以转发、评论微博内容。另外用户可以关注感兴趣的其他用户，当关注之后，即可收听到该被关注用户的微博。每一个数据集都包含大量的用户关注边以及用户发表的微博内容。新浪微博数据是通过从新浪微博在线社交平台采用滚雪球抽样[112]的方式爬取得到的数据。我们随机选取五个用户作为种子，首先爬取这些用户的所有关注对象，然后再爬取这些关注对象的关注对象，以此类推，就像滚雪球一样，这类似于广度优先搜索。最终，我们爬取到20,000个用户和他们之间的2887761条有向关注边。对于每个用户，其微博数量必须大于等于30 条，这样做是为了保证每个用户拥有足够的微博内容信息，以便提取该用户的兴趣特征，该数据包含7115502条微博内容信息，平均每个人的微博数为356。Twitter数据集包含也两部分信息，一部分是由Haewoon Kwak提供的twitter社交网络结构数据[108]，另一个是由斯坦福网络分析项目提供的twitter微博内容数据[113]。通过与微博数据相似的抽样方式，我们获得了11,016个用户和1,202,751条关注边。同时，共有4,415,250条微博被收集到，平均每人发表401条。在这里，我们的条件更加严格，只保留了微博数不小于100的用户。

对于新浪微博和twitter数据，每一个数据集都被划分为两部分：训练集E^T和测试集E^P。训练集包含了90%的已知边，在这里我们排除了关注边小于5条的用户。这样处理后，微博数据的测试集有19,817个用户和289,792条边，twitter数据的测试集有10,604个用户和120,827条边。对于每一个用户u_i，测试集中会包含一定数量的被关注者E_j，边使用E_{ij}表示。

3.1.2 数据预处理

Twitter 和新浪微博都允许用户发表 140 字段文本信息，称为 tweets。我们认为微博内容，不管是原创的，还是从好友那转发而来的，涵盖了大量的主题信息，在一定程度上可以反映用户的兴趣爱好。而要准确获得用户的兴趣表现是非常具有挑战性的。通常的做法是使用 tf-idf[114]来构建用户的兴趣向量，但是这样的做法存在问题，一方面忽略了词与词在结构顺序上的关联性，另一方面所得到的兴趣向量的维数非常大，维度之间存在很多冗余的信息。比如“班级”和“学校”同属于教育这个主题，但却表现为两个维度。目前，尽管有一些比较成熟的降维方法，如 SVD[115]和 PCA[116]可以被用于该 tf-idf 兴趣矩阵降维，但这些算法由于人为定义保留前 K 个特征会造成信息的损失。另一种降维方式是主题模型，比如 LDA[117]，但是基于相关研究[118]，我们发现 LDA 不适合应用在短长度且噪音大的数据，比如 twitter 这类。因此，我们提出另一种方式来获得用户的兴趣向量。在这里，我们使用到了 word2vec 工具来进行词聚类以保证主题的成功抽取。每一个词聚类都可以看成是一个主题。Word2vec 是 google 公司推出的可以处理文本方面工作的工具[119]。给与足够的数据，word2vec 可以使得相关的词或者词组聚合到一块。整个过程表述如下：首先，我们使用 IKAnalyzer 对文本形式的微博内容进行分词。IKAnalyzer 是一款开源的，基于 java 语言开发的轻量级的中文分词语言包，它是以 Lucene 为应用主体，结合词典分词和文法分析算法的中文词组组件，可以很好地应对中英文语句。我们去除停顿词和微博中出现的短链接，还有用户名信息（如“@李开复”）等。然后使用 word2vec 软件对分词后的微博内容进行词聚类，这里聚类数选取为 100，从结果上看聚类效果以满足实验

需求。有了每个词的类别后，我们就可以获得每个用户的 100 维的主题词向量，将属于相应聚类的词频数作为该类别相应维度上的权值 w_i，并进行归一化处理，得到用户的主题兴趣特征向量，见公式(3-1)。用户在这 100 维上的值大小，我们就认为是用户的每一相应类别主题的兴趣大小。

$$t_{ij} = \frac{Freq_{ij}}{\sum_{j}^{n} Freq_{ij}} \quad , \qquad (3\text{-}1)$$

其中 t_{ij} 表示用户 u_i 在 j 类主题上的偏好，$Freq_{ij}$ 对应于用户 u_i 在第 j 维的权值。n 为词类别数，即用户特征向量维度数。通过归一化得到每一维的权重值。

3.1.3 基准预测算法

大多数经典链路预测算法都是基于相似性的，通过节点与边信息来计算相关链路预测分数值，对缺失边进行打分预测。文献[16]详细地对现有的经典链路预测算法进行了比较分析。我们从中选择了些算法作为与新算法进行比较的基准，分别属于局域相似性指标和全局相似性指标。这些算法分别是：Common neighbor、Salton、Jaccard、Sorenson、Hub promoted index（HPI）、Hub depressed index(HDI)、Leicht-Holme-Newman Index(LHN1)、Adamic-Adar Index(AA)、Resource Allocation Index(RA)以及 Katz Index。由于新浪微博和 twitter 都是有向网络，我们使用入度、出度来测试基准链路预测算法的结果表现，发现使用出度来计算基准预测算法的表现要更好些，因此本章中所有结果都是基于出度来计算的。比如 Common neighbor 算法计算的是两个用户共同关注对象的数量。其他算法，诸如 Jaccard、Sorenson 也都是基于出度用户的数量的，但使用了不同的标准化方法。Adamic-Adar 更偏重度小的节点用户的重要性。Resource Allocation

则是从网络科学中的资源分配动力学演化而来。除了这些基于局域相似性的算法外，我们也采用了基于路径性息并属于全局相似性指标的 katz 算法，以使得比较算法更具有说服力。

3.1.4 评价指标

在本章中，我们使用三种指标来衡量算法的有效性，分别为 Area under curve(AUC)[120]、准确率和召回率[121]。

AUC 指标是衡量能够成功区分目标用户感兴趣的对象和不感兴趣的对象的能力。该指标可以被理解为随机选择一条缺失边，其得分要高于那些本身就不存在的边的概率。给定 n 条测试边，如果其中有 n'条边的得分要高于本身不存在的边，有 n''条边的得分是等于本身不存在的边的话，那么 AUC 计算公式可以定义如下：

$$AUC = \frac{n' + 0.5n''}{n}, \tag{3-2}$$

如果随机选择一条边作为缺失边，将得到 AUC 值为 0.5，这等同于纯粹随机选择。因此，AUC 值大于 0.5 的程度是衡量一个算法比随机选择好的程度。

在链路预测算法中，包含缺失边和不存在边的所有观察不到的边按照得分值进行降序排列。除了 AUC 指标，我们还关注该排序列表头部边的预测准确性和召回度。在这里，我们采用通常的做法，只考虑排在前 top-L 部分的边的效果。对于目标用户 u_i，准确率 $P_i(L)$和召回率 $R_i(L)$可以定义为：

$$P_i(L) = \frac{d_i(L)}{L}, R_i(L) = \frac{d_i(L)}{D_i}, \tag{3-3}$$

其中 $d_i(L)$是排序列表前 L 部分中用户 u_i 的缺失边的数量。D_i是用户 u_i 测试集中总

的缺失边数量。

3.1.5 用户主题兴趣相似性指标

在本章中，我们采用 Kullback－Leibler divergence(简写为 KL divergence)来衡量用户之间主题兴趣相似度。KL 距离是一种非对称性指标来衡量分布 P 和 Q 之间的差异性，使用符号 $D_{KL}(P\|Q)$。在这里，每一个用户的口味都可以表征为一个 n 维兴趣向量,其中每一维的值都表示该用户对相应类别主题的偏好，因此这个向量也可以理解为是用户兴趣分布。如果两个兴趣是完全一致，那么该距离值为 0。也就是说值越小，表明用户兴趣越相似。我们分别计算了存在连边和不存在连边两种类别用户对之间的 KL 距离。对于目标用户 u_i，使用符号 $D_{KL}(P_i\|Q_j)$来表示 u_i 和 u_j 之间的 KL 距离，其中 Pi 表示 ui 的主题兴趣分布，Qj 表示 uj 的主题兴趣分布。那么归一化后的 KL 距离公式为：

$$< D_i^{link} >= \frac{\sum_{j \in S_{link}}^{N} D_{KL}(P_i \,||\, Q_j)}{N}, < D_i^{Nolink} >= \frac{\sum_{j \in S_{Nolink}}^{M} D_{KL}(P_i \,||\, Q_j)}{M}$$

$$D_i' = \frac{< D_i^{link} >}{< D_i^{link} > + < D_i^{Nolink} >}, D_i'' = \frac{< D_i^{Nolink} >}{< D_i^{link} > + < D_i^{Nolink} >},$$

（3-4）

其中 S_{link}表示与目标用户存在连边的用户结合，S_{Nolink}表示的是与目标用户没有连边的用户集合。N、M分别是 S_{link}和 S_{Nolink}的大小。D'和 D''分别表示有连边用户对的 KL 距离和没有连边的用户对 KL 距离。

3.2 算法

像 twitter 这类社交网络可以被自然地表示成有向网络 *G(U,E)*，其中 *U* 是用户集合，*E* 是用户之间有向关注边的集合。该有向网络可以使用一个邻接矩阵 A 表示，如果元素 $a_{ij}=1$ 表示用户 u_i 关注用户 u_j，否则为 0。在该种情况下，u_i 称为用户 u_j 的关注者，u_j 称为 u_i 的被关注者。用户 u_j 发表、转发等信息都可以被用户 u_i 浏览到。在这样的社交网络中，链路预测任务就是基于现在的网络结构和文本内容对未来可能存在的边进行预测。

3.2.1 用户主题兴趣相似度

复杂网络中经典的链路预测算法一般上都仅仅使用网络拓扑结构信息，然后对未观测到的边进行打分，按分值降序排列后推荐排在最前面的链路作为最可能在未来存在的连边。但是这些算法仅仅利用了结构信息，像微博内容、用户个人属性以及用户之间的互动等都没有被很好地利用起来。然而，众所周知，twitter 和微博这类社交网络是兴趣驱动的社交平台，这意味着那些在经典算法中未被使用的信息可能可以提供更加丰富的用户特征信息。此外，微博这类社交网络和传统的熟人社交网络，如 Facebook 等是不同的。在 Facebook 网络中，线下社交关系所起的作用是很大的，用户一般偏向于关注同学、朋友、同事或者亲人，在自己的小社交圈子中分享生活细节。相对来说，微博这类社交网络，用户的社交圈子会更广，用户在平台上可以关注娱乐明星、大学教授，甚至是美国总统，只要用户感兴趣想关注即可，这使得用户的圈子中的成员陌生人会比较多。进一步，我们使用 KL 距离[109]来定量的衡量具有关注边的用户对之间和没有关注边的用户对之间的主题兴趣相似度，见图 3.1，可以看到存在关注边的用户对之间的主题兴趣相似度值与没有关注边的用户对之间的主题兴趣相似度值之间存在明显

的分界线。在 twitter 社交网络数据中，存在关注边的用户对之间的 KL 距离值几乎都小于没有关注边的用户对之间的 KL 距离值，仅仅 7.8%的用户对的 KL 距离值落在 0.5 之上。这表明用户普遍倾向于关注具有相似兴趣的用户，在微博网络中同样存在相似的现象，从侧面暗示了仅仅依靠结构信息很难完全获取用户的兴趣特征。

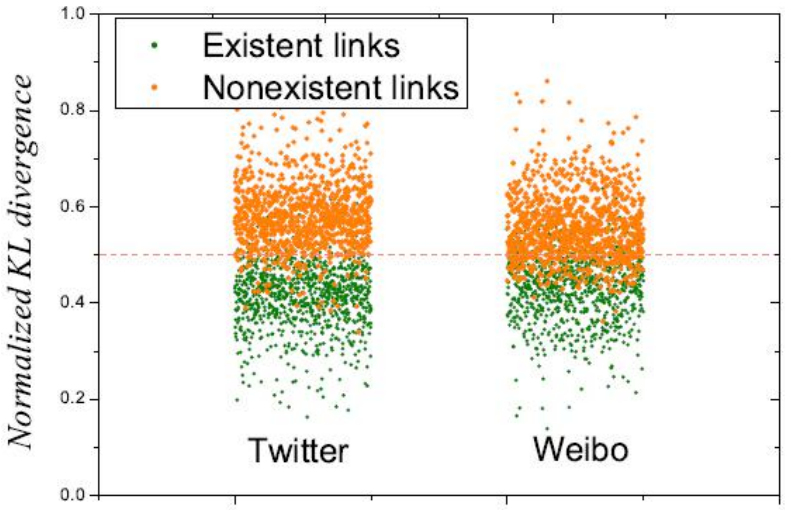

图3.1 twitter和新浪微博数据上存在关注边的用户对和没有关注边的用户对的标准化KL距离。绿色点表示有关注边的用户对之间的标准化KL距离。黄色点表示没有关注边的用户对之间的标准化KL距离。粉色虚线是分界线。

Figure 3.1 The Normalized KL divergence of user pairs having links and non-links on twitter and weibo. Green dots represent the normalized KL divergence of user pairs which have links between them. Yellow dots denote the normalized KL divergence of user pairs which have no direct connection. Pink dash line is the saparatrix.

3.2.2 主题兴趣向量构建及修正

正是由于兴趣的重要性，首先，我们提出一种新的能够基于微博内容获取更加准确和完整用户口味的方式。衡量用户之间兴趣相似度的传统做法是直接计算

用户兴趣向量的内积或者 cosine 值。然而，我们认为，直接使用全部的用户兴趣向量是不能准确反映用户真实口味的，而是会造成噪音干扰。通过实证分析若干 twitter 和微博中用户的微博内容，我们发现尽管用户给自己打上诸如“软件工程师”、“机器学习研究员”或者“体育迷”等，但他们所发表的微博确并不都是关于这些方面的，而是会存在像“今天晚饭不错”这类不能显性表示其真实兴趣的微博内容。如果直接使用完整的兴趣向量来计算会把这类噪音也包含进来，造成信息的误导。因此简单地使用兴趣向量不能很好地反映用户真实完整的情况。进一步深入研究分析后，我们发现用户的关注对象的信息可以被作为该用户的协同信息，这是因为用户虽然不怎么发关于自己兴趣领域的微博，但通常其关注对象一般是经常发该兴趣领域的内容。这样做可以弥补之前所提到的缺陷。此外，我们发现像“晚饭不错”这样的内容并不是经常出现，而是偶尔才会存在，这可能是一种瞬时的用户状态表达。我们认为只有那些兴趣向量中权重值比较大的维度表示的兴趣才是用户最真实而且长期的兴趣口味，算法更应该关注的是这些类别的兴趣。基于以上的讨论，我们提出一种称为兴趣向量修正方式(Interest Vector Rectifing，简写为 IVR)来构建更加准确完整的用户主题兴趣向量。图 3.2 和图 3.3 展示了对目标用户 u_1 的 IVR 过程，其中图 3.3 是图 3.2 中社交网络的一个实例。用户 u_1 是目标用户，当前的链路预测任务是修正目标用户的主题兴趣向量以获得目标用户更加准确完整的口味。从图 3.2 可以看到，用户 u_4、u_5、u_6 以及 u_7 是用户 u_1 的被关注者，而用户 u_2 和 u_3 是用户 u_1 的关注者。每一个用户最开始的时候都有自己的一个原始主题兴趣向量。如图 3.3(b)所示，通过将目标用户及其关注对象中的兴趣维度按照维度值降序排列后，分别选择每一个用户的前 K 个兴趣维度，融合后用于修正目标用户的主题兴趣向量。在本示意图中，K 取 4。

每一个用户的前 K 个维度权值都需要进行归一化处理，如图 3.3(c)所示。那么目标用户 u_1 的修正主题兴趣向量可通过将相关用户的相应维度值进行求和得到，如图 3.3(d)所示，如公式(3-5)所示。

$$\omega_{1j} = \sum_{j}^{n} \omega_{ij}, \quad (3-5)$$

其中 w_{ij} 是用户 u_i 对第 j 个维度的兴趣权重，n 是目标用户和其关注者的数量总和。

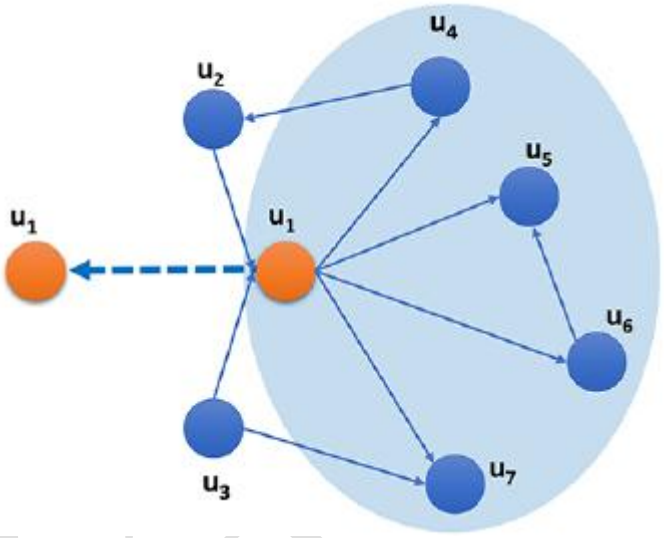

图3.2 目标用户u_1的社交网络示意图。蓝色圈表示用户。蓝色线表示关注边，箭头表示用户关注方向。虚线表示对u_1进行修正主题兴趣向量后的新向量表示。

Figure 3.2 Network illustration of target user u_1. Blue circles represent users. The arrow lines mean following relationships. The direction of arrow line represents the following direction. The dash arrow line represents rectifying the interest vector for u_1.

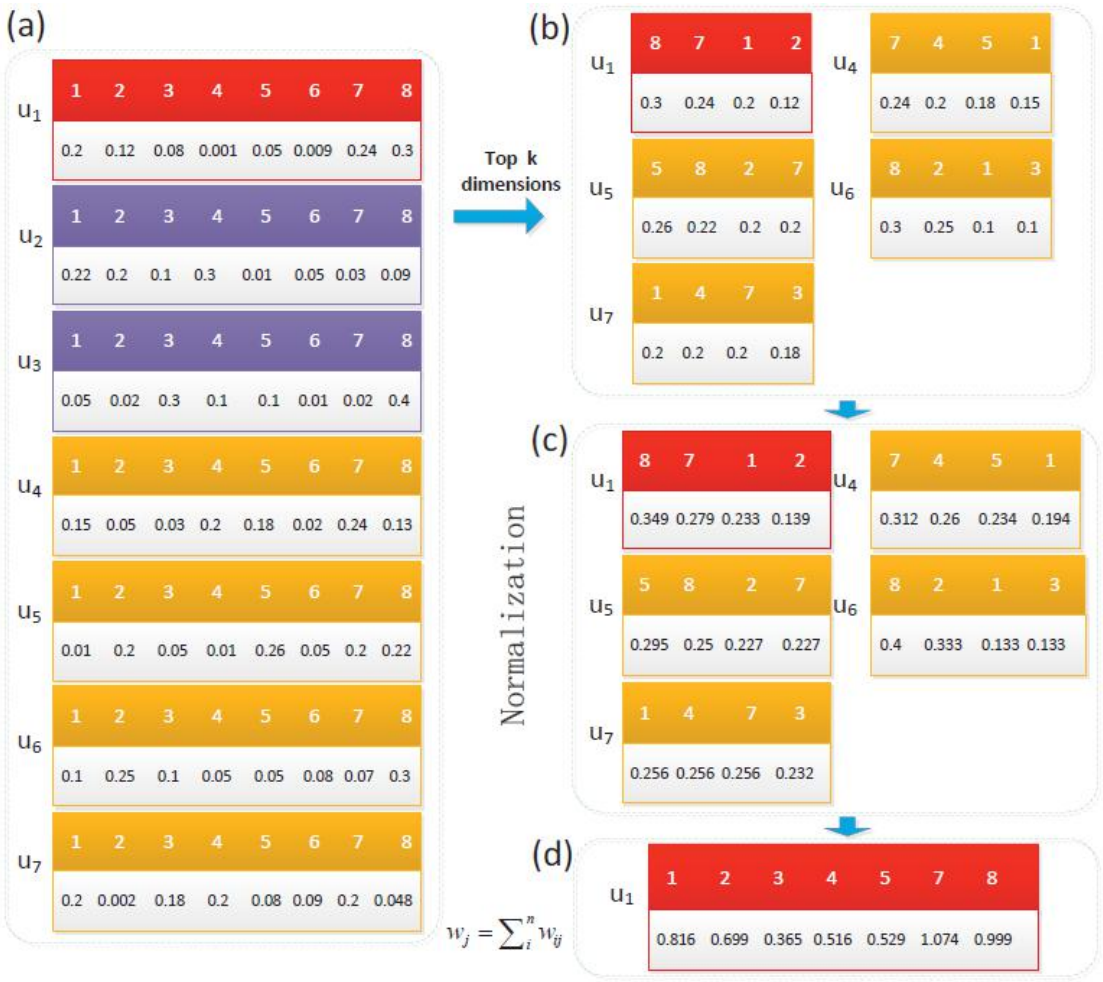

图 3.3 基于图3.2社交网络对目标用户u_1进行主题兴趣向量修正过程。红色框表示用户u_1，黄色框表示用户u_1关注对象，紫色框为关注u_1的用户。(a)用户原始兴趣向量；（b）目标用户及其关注对象的前K个兴趣维度；（c）（b）中用户兴趣向量进行标准化后的向量结果；（d）通过对（c）中相应兴趣维度权值进行计算后得到的用户u_1的修正主题兴趣向量。

Figure 3.3 Process of Interest Vector Rectifying for target user u_1 based on social graph of (Figure 3.2). Red tables represent target user $u1$, yellow ones for followees of u_1 and purple for others. (a)Original interest vectors of users. (b)Top K dimensions of interest vectors for target user and its followees. (c)Normalization for the top K dimensions of interest vectors in (b). (d)Rectified interest vector for target user u_1 by sum the weights of corresponding dimensions in (c).

3.2.3 桥接节点的作用

此外，我们发现桥接节点在信息传播中扮演着非常重要的角色。给定目标用户 u_i，及其该目标用户潜在的关注者 u_j，我们定义 u_i 目前的关注者中同时直接关注 u_j 的节点为 u_i 与 u_j 之间的桥接节点。如图 3.4 所示，要预测用户 u_1 是否会关注用户 u_7，黄色圈表示的用户 u_4 和 u_5 被定义为桥接节点。为了确定这些桥接节点的重要性，我们设计了一个能量随机游走实验来确定出他们在信息传播中所发挥的作用。首先，基于训练集数据，我们构建了信息流网络。如果一条关注边为用户 u_i 关注用户 u_j，那么信息流动方向刚好是相反的，是从用户 u_j 流向用户 u_i。随机从测试集中抽取一条缺失边，为 E_{ij}，对应的两个用户 u_i 和 u_j，其中 u_i 定义为信息流动目标节点，u_j 为信息流动初始节点，然后在该信息流动网路上跑随机游走实验。首先初始节点被赋予一个单位能量，其他节点没有能量，整个扩散过程可以被描述为通过转移矩阵概率 P，其中元素为 $P_{ji}=a_{ji}/k_j$，表示的是随机游走者从节点 u_j 在下一步跳到节点 u_i 的概率，$a_{ji}=1$ 表示 u_j 是 u_i 的被关注对象，否则为 0，k_j 是 u_j 在信息传播网络中的出度。从初始节点出发，我们使用 $\pi_{ji}(t)$ 来表示能量经过 t 时间步后传递到 u_i 的概率值，如公式（3-6）所示：

$$\vec{\pi_j}(t) = P^T \vec{\pi_j}(t-1), \tag{3-6}$$

其中 $\vec{\pi_j}(t)$ 表示的是一个 N x 1 的节点向量，向量中的元素是当前 t 时间下每一个节点所得到的能量值。字母 j 表示的是初始能量节点是第 j 个节点，T 表示转置。经过 t 时间步后，该信息传播网络中的节点都会被赋予一定的能量[110]。在我们的实验中，能量一旦传递到目标节点用户，那么该信息将会被目标节点完全吸收，而不会从他再次传出这是为了避免受其自身能量传递的影响，如果目标节点扩散

出的能量，再次通过其他节点传递回来，会造成能量的重复计算。并且，我们设置能量传递阶数最多为 3 步，这是因为有研究表明 99.9%的信息传递都局限在 3 步以内[111]。这里，我们随机选择测试集中 1000 个用户对计算从源节点到目标节点通过相应桥接节点传递的总能量比例。该比例称为有效信息传播能力(Effective Information Dissemination Capability，简称 EIDC)。如图 3.5 所示，新浪微博和 twitter 中绝大多数节点对的 EIDC 都大于 0.5，比例分为是 86.6%和 85.7%。在新浪微博上，从源节点通过桥接节点传递到目标节点的平均能量值为 0.663，在 twitter 上为 0.668。这证明了信息从源节点传递到目标节点过程中，桥接节点的重要性。

3.2.4 基于主题兴趣相似性的最大化偏好算法

基于以上的讨论，本章提出一种叫做基于主题兴趣相似性的最大化偏好(Maximum Preference on Interest Similarity(MPIS))的新算法来利用微博内容和结构信息去应对社交网络中的链路预测问题。给定一个链路预测任务预测边 E_{ij} 的未来存在性，对于目标用户 u_i，我们可以通过 IVR 过程得到目标用户的修正主题兴趣向量。其中该向量的每一维权重值表示的是用户 u_i 对相应类别兴趣的偏好，该值也可以被理解为用户会以多大概率基于该兴趣去关注其他用户。这里需要注意潜在被关注者 u_j 的兴趣向量是不需要被修正的，只需要提取前 K 个兴趣维度进行权值的归一化操作。这是因为当用户考虑关注其他人的时候，仅仅会注意其本人的微博内容，而不会去关注该用户关注的邻居的内容。那么我们就可以使用用户 u_i 的修正后的兴趣向量和 u_j 的前 K 个兴趣维度的并集来作为新的兴趣向量总的兴趣维 m 维。因此，接下来就是分别将用户 u_i 的修正兴趣向量和 u_j 的

前 K 个兴趣向量重整为 m 维度的向量空间。符号 $\vec{A}$ 和 $\vec{B}$ 分别用来表示重构后的 u_i 修正兴趣向量和 u_j 前 K 个维度的兴趣向量。通过计算他们之间的内积可以得到用户 u_i 和 u_j 之间的主题兴趣相似度。此外，我们将用户之前的重叠兴趣维度的数量加以考虑。如果向量 $\vec{A}$ 和 $\vec{B}$ 中某个对应维度的权值都是大于 0 的，那么重合度加 1。该指标可以反映用户的重要兴趣的重合率，用以弥补向量内积的缺陷。因此内积的计算仅仅展现了相似性的强度，而没有考虑兴趣重合广度。除了该兴趣相似性之外，我们将桥接节点的贡献也融入到算法中。两个节点之间桥接节点越多，潜在被关注者的信息更可能传递到目标用户。换句话说，桥接节点扮演着信息可见性扩大的作用。在微博或者 twitter 这类社交网络中，如果两个用户相隔很远，那么即使他们之间的相似性非常高，仍然不太可能产生连边。因此，最终用户 u_i 关注 u_j 的概率计算公式为：

$$P_{ij} = \sum_{k \in S_{if}} (I_{kj} \times (\vec{A} \bullet \vec{B})) + \frac{I_{\vec{A}} \bullet I_{\vec{B}}}{K} \quad , \qquad (3\text{-}7)$$

其中 S_{if} 值得是 u_i 的关注对象结合。I 是指示函数。当 u_k 关注 u_j 时，I_{kj} 为 1，否则为 0。$I_{\vec{A}}$ 是一个二值向量。如果向量中维度权值大于 0，则该二值向量中对应的维度设置为 1，否则为 0。$I_{\vec{A}} \bullet I_{\vec{B}}$ 是用户 u_i 和 u_j 之间兴趣重合度。K 是 top-K 的值。

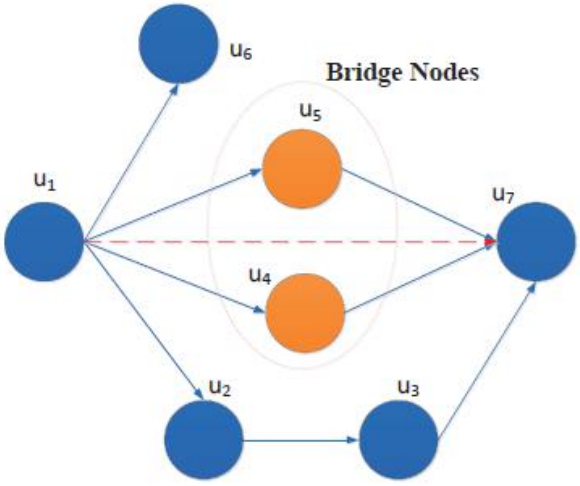

图 3.4 桥接节点示意图。圆圈表示用户，粉色虚线表示链路预测任务，在这里表示预测用户u_1是否会关注用户u_7。在椭圆中的黄色圈表示的是桥接节点。

Figure 3.4 Illustration of Bridge Nodes. Circles represent users and pink dash line is a link to be predicted between user u_1 and u_7. Yellow circles in ellipse with yellow dash line are bridge nodes and blue one are ordinary nodes.

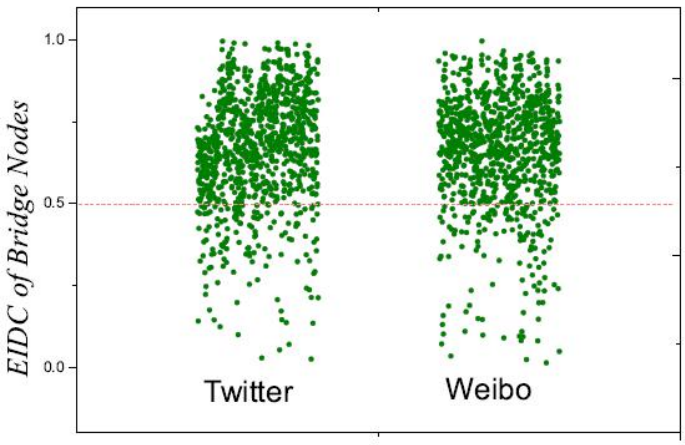

图 3.5 桥接节点的有效信息扩散能力。绿色点表示从初始节点通过桥接节点传递给目标节点的能量比例。粉色线表示分割线。

Figure 3.5 Effective Information Dissemination Capability of Bridge Nodes. Green dots represent ratio of source node's information spreading through bridge nodes and other nodes. Pink dash line is the saparatrix.

3.3 结果与分析

3.3.1 算法效果比较

我们比较了所提出的算法与基准算法在 twitter 和新浪微博数据上的结果表现。评价指标选用了 AUC、准确率和召回率三种。其中 AUC 指标是通过 100 次独立实验平均后的结果。表 3.1 和 3.2 列出了所提出的算法的 AUC 结果，在 twitter 和新浪微博数据上，分别比最好的基准算法还要高出 15%和 9.1%。进一步，我们比较了各个算法在稳定性上的表现，通过计算结果的标准差，发现所提出的算法的标准差要远小于其他算法，证明我们的算法波动性比较小。表 3.3 和 3.5 以及图 3.6 展示了各类算法在不同的推荐列表长度上的推荐准确度。我们的算法要远远好于其他算法，特别是在第一位的推荐精度上，在新浪微博上，准确率相较于最好的基准算法提升了 103.7%，在 twitter 上提升了 12.4%。对于召回率的表现，MPIS 也是远远好于最好的基准算法，仅仅在 twitter 数据上前两位的推荐精度上和 Katz 相一致，见表 3.4 和 3.5 以及图 3.7。除了高准确率，我们的算法计算复杂度远远小于 katz 算法，这是我们出色的地方。

表 3.1 新浪微博数据上各个算法的 AUC 结果和稳定性结果

Table 3.1 AUC and Stability of all algorithms on weibo data

Algorithm	Mean of AUC	Standard deviation
MPIS	0.9207	0.000621
Common Neighbor(CN)	0.7923	0.000761
Jaccard(JAC)	0.7927	0.000735
Salton(SAL)	0.8010	0.000760
Sorenson(SRE)	0.7996	0.000758
HPI	0.7895	0.000770
HDI	0.7822	0.000709
LHN-I(LHN)	0.7552	0.000693
Adamic-Adar(AA)	0.7914	0.000769

Resource Allocation(RA)	0.7694	0.000764
Katz(10^{-2})	0.6789	0.002845

表 3.2 twitter 数据上各个算法的 AUC 结果和稳定性结果

Table 3.2 AUC and Stability of all algorithms on twitter data

Algorithm	Mean of AUC	Standard deviation
MPIS	0.8991	0.000958
Common Neighbor(CN)	0.8240	0.001057
Jaccard(JAC)	0.7786	0.001137
Salton(SAL)	0.8066	0.001133
Sorenson(SRE)	0.8155	0.001090
HPI	0.8133	0.001162
HDI	0.7616	0.001128
LHN-I(LHN)	0.7142	0.001192
Adamic-Adar(AA)	0.8244	0.001064
Resource Allocation(RA)	0.7929	0.001094
Katz(10^{-2})	0.6747	0.001686

表 3.3 weibo 数据上各个算法的准确率结果

Table 3.3 Precision of all algorithms on weibo data

Length	MPIS	AA	CN	HDI	HPI	JAC	LHN-I	RA	SAL	SRE	Katz
1	0.203	0.071	0.072	0.086	0.023	0.104	0.012	0.058	0.109	0.104	0.046
2	0.179	0.062	0.063	0.074	0.026	0.089	0.013	0.050	0.094	0.089	0.041
5	0.142	0.048	0.049	0.057	0.027	0.069	0.015	0.039	0.073	0.068	0.032
10	0.114	0.039	0.040	0.045	0.025	0.054	0.014	0.032	0.058	0.054	0.026
15	0.099	0.034	0.035	0.039	0.024	0.046	0.014	0.027	0.049	0.046	0.023
20	0.089	0.031	0.031	0.035	0.022	0.041	0.013	0.025	0.044	0.041	0.021
30	0.075	0.026	0.027	0.029	0.020	0.035	0.012	0.021	0.037	0.034	0.019
40	0.066	0.023	0.024	0.026	0.018	0.030	0.011	0.019	0.033	0.030	0.017
50	0.060	0.021	0.022	0.023	0.017	0.028	0.011	0.017	0.029	0.027	0.016
60	0.055	0.020	0.020	0.022	0.016	0.025	0.010	0.016	0.027	0.025	0.015
70	0.051	0.019	0.019	0.020	0.015	0.024	0.010	0.015	0.025	0.023	0.014

80	0.048	0.018	0.018	0.019	0.015	0.022	0.009	0.014	0.023	0.022	0.014
90	0.045	0.017	0.017	0.018	0.014	0.021	0.009	0.013	0.022	0.021	0.013
100	0.042	0.016	0.017	0.017	0.013	0.020	0.009	0.013	0.021	0.020	0.013

表 3.4 weibo 数据上各个算法的召回率结果

Table 3.4 Recall of all algorithms on weibo data

Length	MPIS	AA	CN	HDI	HPI	JAC	LHN-I	RA	SAL	SRE	Katz
1	0.020	0.009	0.009	0.009	0.005	0.011	0.002	0.008	0.012	0.012	0.004
2	0.034	0.015	0.015	0.015	0.010	0.018	0.005	0.013	0.019	0.020	0.006
5	0.064	0.027	0.027	0.027	0.021	0.033	0.010	0.023	0.036	0.037	0.012
10	0.100	0.041	0.041	0.041	0.035	0.049	0.018	0.035	0.054	0.055	0.019
15	0.127	0.052	0.052	0.051	0.045	0.061	0.024	0.044	0.068	0.068	0.025
20	0.150	0.061	0.060	0.060	0.054	0.071	0.030	0.051	0.079	0.079	0.031
30	0.187	0.075	0.075	0.074	0.068	0.087	0.039	0.063	0.097	0.096	0.042
40	0.217	0.087	0.087	0.085	0.080	0.100	0.047	0.072	0.111	0.110	0.052
50	0.243	0.097	0.097	0.095	0.090	0.112	0.054	0.080	0.123	0.122	0.060
60	0.265	0.106	0.106	0.103	0.099	0.121	0.061	0.088	0.134	0.132	0.068
70	0.284	0.114	0.114	0.111	0.107	0.130	0.066	0.094	0.143	0.141	0.077
80	0.302	0.121	0.122	0.118	0.115	0.137	0.072	0.100	0.151	0.149	0.084
90	0.318	0.128	0.129	0.124	0.122	0.145	0.077	0.106	0.159	0.156	0.092
100	0.333	0.135	0.135	0.130	0.128	0.151	0.082	0.111	0.166	0.163	0.098

表 3.5 twitter 数据上各个算法的准确率结果

Table 3.5 Precision of all algorithms on twitter data

Length	MPIS	AA	CN	HDI	HPI	JAC	LHN-I	RA	SAL	SRE	Katz
1	0.200	0.114	0.134	0.112	0.018	0.129	0.008	0.100	0.133	0.152	0.178
2	0.173	0.100	0.116	0.095	0.021	0.110	0.009	0.087	0.115	0.129	0.150
5	0.135	0.078	0.089	0.070	0.026	0.083	0.010	0.067	0.089	0.097	0.109
10	0.105	0.062	0.070	0.054	0.026	0.064	0.010	0.052	0.069	0.074	0.081
15	0.089	0.053	0.060	0.046	0.025	0.054	0.010	0.045	0.058	0.062	0.066
20	0.079	0.048	0.054	0.040	0.024	0.047	0.010	0.039	0.051	0.055	0.058
30	0.065	0.040	0.045	0.033	0.022	0.039	0.009	0.033	0.042	0.045	0.046
40	0.057	0.035	0.039	0.029	0.021	0.034	0.009	0.028	0.037	0.039	0.040

50	0.050	0.031	0.035	0.026	0.019	0.030	0.009	0.025	0.033	0.034	0.035
60	0.046	0.029	0.032	0.023	0.018	0.027	0.008	0.023	0.030	0.031	0.031
70	0.042	0.027	0.030	0.022	0.017	0.025	0.008	0.021	0.027	0.029	0.029
80	0.039	0.025	0.028	0.020	0.017	0.023	0.008	0.020	0.025	0.027	0.026
90	0.036	0.023	0.026	0.019	0.016	0.022	0.008	0.019	0.024	0.025	0.024
100	0.034	0.022	0.025	0.018	0.015	0.021	0.007	0.018	0.022	0.024	0.023

表 3.6 twitter 数据上各个算法的召回率结果

Table 3.6 Recall of all algorithms on twitter data

Length	MPIS	AA	CN	HDI	HPI	JAC	LHN-I	RA	SAL	SRE	Katz
1	0.033	0.015	0.017	0.016	0.005	0.019	0.003	0.014	0.020	0.023	0.034
2	0.054	0.028	0.029	0.027	0.012	0.031	0.006	0.025	0.032	0.038	0.055
5	0.098	0.053	0.056	0.047	0.030	0.055	0.013	0.046	0.058	0.067	0.096
10	0.146	0.082	0.087	0.068	0.055	0.080	0.023	0.069	0.087	0.098	0.137
15	0.180	0.103	0.110	0.085	0.074	0.099	0.031	0.085	0.108	0.110	0.166
20	0.207	0.120	0.129	0.097	0.090	0.114	0.039	0.098	0.124	0.136	0.188
30	0.250	0.147	0.158	0.118	0.117	0.137	0.053	0.118	0.150	0.162	0.223
40	0.282	0.168	0.180	0.134	0.139	0.154	0.065	0.134	0.169	0.182	0.248
50	0.309	0.186	0.199	0.147	0.157	0.169	0.076	0.147	0.185	0.198	0.270
60	0.332	0.201	0.216	0.158	0.173	0.181	0.085	0.159	0.198	0.212	0.288
70	0.351	0.214	0.230	0.168	0.187	0.192	0.094	0.169	0.210	0.224	0.303
80	0.369	0.226	0.243	0.177	0.200	0.201	0.102	0.178	0.220	0.235	0.317
90	0.384	0.237	0.255	0.185	0.212	0.210	0.109	0.186	0.230	0.245	0.329
100	0.398	0.247	0.266	0.192	0.223	0.218	0.116	0.194	0.238	0.254	0.340

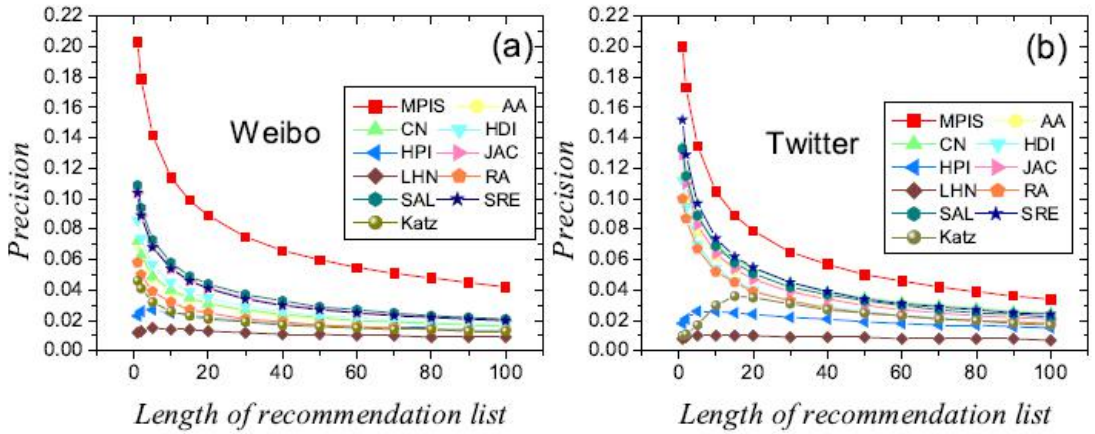

图3.6 准确性与推荐列表长度之间的关系。红色、黄色、浅绿、浅蓝、海蓝、粉色、灰色、金色、黑色、深黄分别表示MPIS、AA、CN、HDI、HPI、Jaccard、LHN-I、RA、Salton、Sϕrenson 和 Katz的结果。

Figure 3.6 Relation between precision and length of recommendation list. The red, yellowish, aqua, pale blue, navy, pink, grey, golden, petrol, black, dark yellow represent the cases of MPIS, AA, CN, HDI, HPI, Jaccard, LHN-I, Resource Allocation, Salton, Sϕrenson and Katz, respectively.

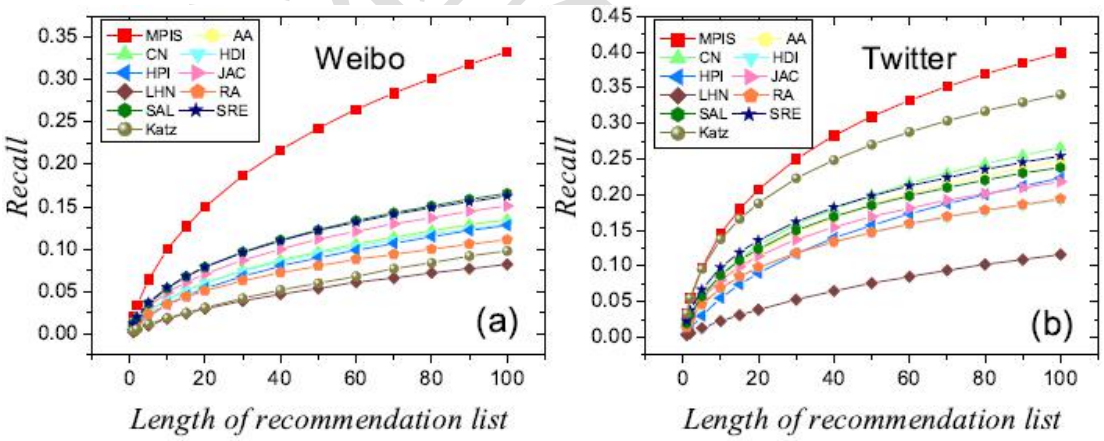

图3.7 召回率与推荐列表长度之间的关系。红色、黄色、浅绿、浅蓝、海蓝、粉色、灰色、金色、黑色、深黄分别表示MPIS、AA、CN、HDI、HPI、Jaccard、LHN-I、RA、Salton、Sϕrenson 和 Katz的结果。

Figure 3.7 Relation between recall and length of recommendation list. The red,yellowish, aqua, pale blue, navy, pink, grey, golden, petrol, black,dark yellow

represent the cases of TIS,AA, CN, HDI, HPI, Jaccard, LHN-I, Resource Allocation, Salton, Sørenson and Katz, respectively.

3.3.2 不同 IVR 过程对结果的影响

进一步，我们研究了不同 IVR 类型对结果的影响。图 3.8 显示的是三种类别 IVR 的结果比较。我们使用“FS”表示使用目标用户及其关注对象的兴趣向量信息对目标用户兴趣向量进行修正。“FO”表示只使用目标用户关注对象的兴趣向量对目标用户兴趣向量进行修正。“SO”表示只使用目标用户自己的兴趣向量进行计算，不进行修正。可以看到“FS”和“FO”的结果在 Recall 和 AUC 指标上要远远好于“SO”的方式。“FS”在 AUC 指标上要略好于“FO”方式。但是这三类指标在准确性上表现是类似的。因此我们可以得出结论，“FS”这类 IVR 方式可以获得更佳准确的用户主题兴趣向量。

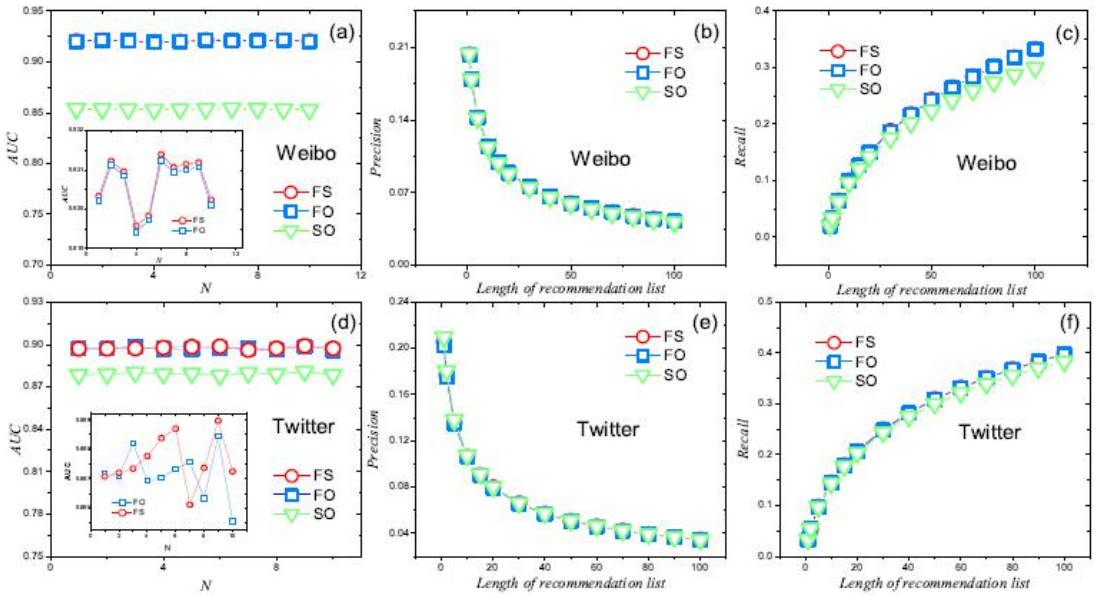

图3.8 IVR与结果(AUC、准确率、召回率)之间的关系。不同颜色表示不同的IVR类型。“FS”（红色）表示采用目标用户和其关注者的主题兴趣向量对目标用户修正兴趣向量。“FO”（蓝色）表示只使用目标用户的关注者的主题兴趣向量进

行修正目标用户的兴趣向量。“SO”（绿色）表示依然使用目标用户自身的兴趣向量，而不进行修正。（a）和（d）分别是在新浪微博和twitter上IVR与AUC的关系。N是实验进行的轮数。（b）和（e）是IVR与准确率之间的关系。（c）和（f）是IVR与召回率之间的关系。

Figure 3.8 Relation between IVR and evaluation metrics(AUC, Precision and Recall). Different colors represent different types of IVR. Symbol ’FS’(Red circle and line) means we employ the way using target user and its followees to rectify the target user’s interest vector. Symbol ’FO’(Blue circle and line) represents the way using target user’s followees to rectify its own interest vector. And Symbol ’SO’(Green circle and line) means not to rectify target user’s interest vector, using its original one. (a) and (d) are relations between IVR and AUC on weibo and twitter respectively. N is the number of experiments. Insets are comparisons of FS and FO. (b) and (e) are relations between IVR and precision. (c) and (f) are relations between IVR and recall.

3.3.3 不同 top K 对结果的影响

此外我们分析了不同的 K 值对结果的影响。如图 3.9 所示，小的 K 会导致大量信息的丢失，使得 AUC 效果比较差。不过随着 K 的增大，AUC 也随之增大，逐渐进入一个较平稳的阶段。但当 K 超过某一个特定值后，在新浪微博中是 80，在 twitter 中是 40，AUC 值开始降低。该现象表明原始的完整的兴趣向量是存在噪音的，通过抽取前 K 个兴趣维度可以降低噪音，获得更加准确和稳定的用户兴趣口味，可以确保得到高 AUC 表现。

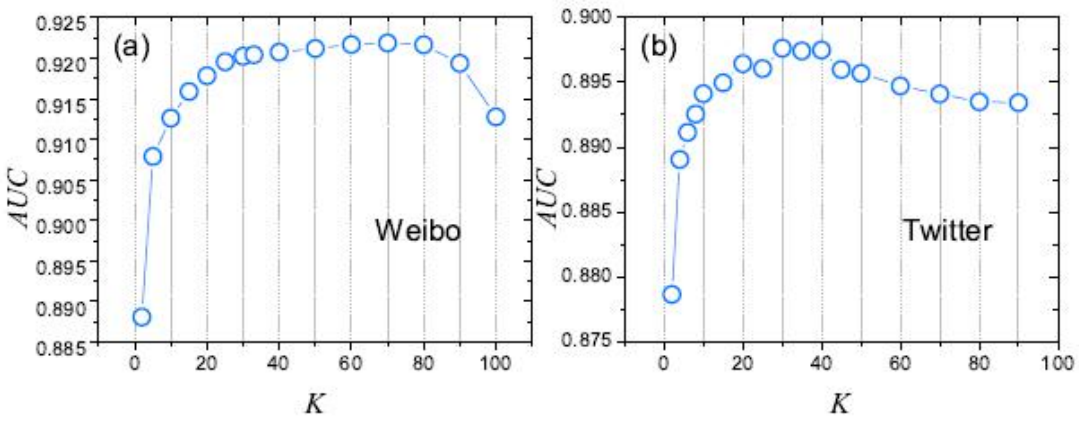

图 3.9 top K与AUC之间的关系

Figure 3.9 Relation between top K and AUC on weibo and twitter

3.4 本章小结

在本章中，针对复杂网络经典链路预测算法不能有效利用新浪微博和twitter这类社交网络上的文本信息来做边的预测，本文提出一种新的基于用户兴趣相似度的最大化偏好算法，能够同时使用蕴含用户口味的文本内容信息以及网络拓扑结构信息，来对未来可能会存在的边进行预测。首先，我们使用KL距离来实证衡量存在关注边的用户对之间和不存在关注边的用户对之间的兴趣相似性，发现存在关注边的用户对之间的兴趣相似性非常强，这表明用户关注其他人很大程度上是基于内容兴趣相似性。除了这点，我们又研究了网络结构对用户关注产生的影响，这是由于在微博这类社交网络上，当两个用户之间的网络距离很远，即使两个用户的兴趣非常相似，仍然不可能存在关注边，因为根本没有几乎注意到对方。通过设计能量随机游走实验，我们发现桥接节点在信息传播过程中扮演着非常重要的角色。因此，基于以上实证发现，我们将用户兴趣相似性的强度及广度，还有桥接节点作用融合到新的链路预测算法中。所提出的算法与基准算法在twitter和微博数据上都进行了比较，实验结果表明我们所提出的算法

在 AUC、precision 以及 recall 三种指标上都要远远好于经典的链路预测算法。此外，我们对 IVR 类型、top K 的影响都做了讨论。这表明我们所提出的算法能够很好地针对 twitter 这类社交网络进行边的预测。

小结

对社交网络的研究，可以帮助我们加深对人类行为的认知，挖掘人的兴趣偏好，探究人类行为背后的机理，所取得的成果不仅能够促进科学研究的发展，而且能够帮助在线服务平台设计更好的功能，创造更优的算法，进一步优化用户体验，提高用户的活跃度，塑造用户对于服务平台的信任，以此带来商业上的价值。

本文主要从两个角度对社交网络进行了研究。一方面是从目前学术界对社群网络的认识不足、工业界对社群网络上群推荐效果不够理想的问题切入，以 QQ 群网络数据为基础，对这类明确定义了群结构的社交网络进行了分析。研究方法采用复杂网络的相关知识，以实证分析和模型模拟两种方式，对该类网络进行了探究。另一方面是针对算法适用性问题，对 twitter 这类社交网络上链路预测算法进行了研究，由于复杂网络中经典算法不能有效利用文本信息，因此本文基于对 twitter 和新浪微博数据的相关实证分析，以信息检索及网络科学相关知识，通过实验方式，提出一种能够较好地同时利用网络拓扑结构和微博文本信息的链路预测算法，对算法进行了设计创新。

对于社群网络的研究，我们具体从群规模分析、群规模演化与创群时间的关系、用户加群数分布、群群之间关系、群与群中成员用户特征的关系、小世界性质、用户年龄及性别对群相关性质的影响以及社群生长机制探索等方面深入分析这类社交网络的性质特点。从中我们发现了一系列有趣规律，这些性质对于群推荐算法设计非常具有启发性。比如研究发现该类网络上的群按照群规模演化特点大致可分为两种，一种是自群创建以来短时间内群规模即达到稳定状态，之后较少发生明显变化的群，如同学群；另一种是随着建群时间的增长，群规模逐渐扩

大的群。这暗示在群推荐中，如果群属于前一种类型，那么需要计算用户与群中用户的社交关系紧密度,如果不是很紧密,即使用户与群表现出的兴趣非常相似，也最好较少推荐，因为这类群偏好线下紧密的社交关系，而非兴趣。缓慢演化的群一般来说是基于成员兴趣增长的，线下关系影响较小。此外，我们发现处于初始阶段的群，其规模的扩大与群中活跃用户非常相关，因此在群推荐过程中，为了构建更加良好的生态，可以更偏向于活跃用户推荐，不仅可以提高社群的生长速度，而且推荐的接受度也会较高，因为活跃用户本身具有较高的加群倾向。此外，我们发现年龄在加群行为中扮演着非常重要的角色，在青少年时期，该时期用户加群主要基于兴趣，用户普遍会加入多种不同兴趣类别的群，且群规模都不大，但群中用户年龄分布较广。随着年龄增长，用户加群数量会逐渐减少，加入的群,其成员年龄较为一致,也就是说该时期用户倾向于和相似年龄的用户交流，进一步到事业发展期，用户又重新偏向于那些成员年龄分布广的群，主要是工作后会接触各种类型的群体，而且此时用户的群数基本固定。也就是说对该时期的人推荐群，应该更偏向事业相关，而且推荐频次不能太高。当然，性别也会对加群行为产生巨大影响,女性用户普遍比男性用户更早进入相对固定的社交圈子时期。除此之外，还有很多发现，对于社群网络的研究确实加深了我们对其认识，能够帮助我们设计更加优良的算法。

对于 twitter 这类社交网络上的链路预测问题，我们首先从实证分析角度，研究这类网络上存在的性质特点，特别是文本内容上表现出来的特征。为了研究的便捷性及准确性，我们对用户的微博内容进行了分词处理，并使用主题模型进行降维，以向量空间的形式表征一个用户的主题兴趣向量。然后采用 KL 距离对存在关注边的用户对与不存在关注边的用户对分别进行了衡量,发现存在关注边

的用户对普遍表现出较低的 KL 距离值，这说明在这类社交网络上，用户普遍关注其感兴趣的用户。此外，我们从结构上分析了用户产生关注边的影响因素，发现桥接节点对于目标用户在对潜在关注对象的信息获取性方面起着至关重要的作用。这是因为在 twitter 这类社交网络上，如果用户所处的距离非常远，即使两者存在相似的兴趣特征，仍然不太可能产生关注边，除非是其他原因造成了影响。为了确定桥接节点所起的作用，我们特别设计了一种能量随机游走的实验，通过多轮独立实验，发现潜在受关注用户的信息绝大多数是通过桥接节点传递到目标用户的。因此，基于以上实证发现，我们将用户兴趣相似性的强度及广度，还有桥接节点作用融合到新的链路预测算法中，提出一种基于用户兴趣相似度的最大化偏好算法，能够同时使用蕴含用户口味的文本内容信息以及网络拓扑结构信息，对未来可能会存在的边进行预测。为了验证所提出算法的有效性，我们采用了 AUC、precision、recall 三种指标来衡量，并且与一系列经典算法进行了结果比较，实验结果表明我们所提出的算法要远远好于经典的链路预测算法，不仅表现在三个指标上，还体现在计算复杂度。此外，我们对 IVR 类型、top K 的影响都做了讨论。这说明所提出的算法能够很好地针对 twitter 这类社交网络进行用户关注关系的预测。

此外，在研究生期间，本人还对行人流特点、新闻时间发表时间特性、个性化推荐算法、引文网络中学术成就影响因素、PM2.5 的时空特性及预测、社交网络重要节点挖掘、网络结构恢复、排序学习等做了相关研究，积累了相关知识及经验。

基于现有的研究成果及知识储备，未来还有很多工作要做。一是将现有的社群网络实证发现转换成算法语言，对 QQ 群推荐算法进行优化设计。二是对于

twitter 这类社交网络的链路预测算法进行更深入的研究，目前我们在结构上使用的是一阶节点信息,未来需要往更多阶方向走,因为信息存在多阶传递的特点，结果或许会有进一步提升。

参考文献

1. Thompson J B. The media and modernity: A social theory of the media[M]. Stanford University Press, 1995.

2. Chelmis C, Prasanna V K. Social networking analysis: A state of the art and the effect of semantics[C]//Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on. IEEE, 2011: 531-536.

3. Kaplan A M, Haenlein M. Users of the world, unite! The challenges and opportunities of Social Media[J]. Business horizons, 2010, 53(1): 59-68.

4. Pang B, Lee L. Opinion mining and sentiment analysis[J]. Foundations and trends in information retrieval, 2008, 2(1-2): 1-135.

5.. Newman M E J, et al. The structure and dynamics of networks[M]. Princeton University Press, 2006.

6. Backstrom L, Boldi P, Rosa M, et al. Four degrees of separation[C]//Proceedings of the 4th Annual ACM Web Science Conference. ACM, 2012: 33-42.

7.Moody J, White D R. Structural cohesion and embeddedness: A hierarchical concept of social groups[J]. American Sociological Review, 2003: 103-127.

8.Traud A L, Kelsic E D, Mucha P J, et al. Comparing community structure to characteristics in online collegiate social networks[J].SIAM review, 2011, 53(3): 526-543.

9.Leskovec J, Lang K J, Mahoney M. Empirical comparison of algorithms for network community detection[C]//Proceedings of the 19th international conference on World wide web. ACM, 2010: 631-640.

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

11.Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. Physical review E, 2004, 69(2): 026113.

12.Fortunato S, Barthélemy M. Resolution limit in community detection[J]. Proceedings of the National Academy of Sciences, 2007, 104(1): 36-41.

13.Mislove A, Koppula H S, Gummadi K P, et al. Growth of the flickr social network[C]//Proceedings of the first workshop on Online social networks. ACM, 2008: 25-30.

14.Newman M E J. Clustering and preferential attachment in growing networks[J]. Physical Review E, 2001, 64(2): 025102.

15.Barabâsi A L, Jeong H, Néda Z, et al. Evolution of the social network of scientific collaborations[J]. Physica A: Statistical mechanics and its applications, 2002, 311(3): 590-614.

16.Lü L, Zhou T. Link prediction in complex networks: A survey[J]. Physica A: Statistical Mechanics and its Applications, 2011, 390(6): 1150-1170.

17.Chowdhury G. Introduction to modern information retrieval[M]. Facet publishing, 2010.

18.Ravasz E, Somera A L, Mongru D A, et al. Hierarchical organization of modularity in metabolic networks[J]. science, 2002, 297(5586): 1551-1555.

19.Leicht E A, Holme P, Newman M E J. Vertex similarity in networks[J]. Physical Review E, 2006, 73(2): 026120.

20. Lü L, Jin C H, Zhou T. Similarity index based on local paths for link prediction of complex networks[J]. Physical Review E, 2009, 80(4): 046122.

21. Katz L. A new status index derived from sociometric analysis[J]. Psychometrika, 1953, 18(1): 39-43.

22. Klein D J, Randić M. Resistance distance[J]. Journal of Mathematical Chemistry, 1993, 12(1): 81-95.

23. Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine[J]. Computer networks and ISDN systems, 1998, 30(1): 107-117.

24. Clauset A, Moore C, Newman M E J. Hierarchical structure and the prediction of missing links in networks[J]. Nature, 2008, 453(7191): 98-101.

25. Al Hasan M, Chaoji V, Salem S, et al. Link prediction using supervised learning[C]//SDM'06: Workshop on Link Analysis, Counter-terrorism and Security. 2006.

26. Ye J, Cheng H, Zhu Z, et al. Predicting positive and negative links in signed social networks by transfer learning[C]//Proceedings of the 22nd international conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2013: 1477-1488.

27. Watts D J. Small worlds: the dynamics of networks between order and randomness[M]. Princeton university press, 1999.

28. Albert R, Barabási A L. Statistical mechanics of complex networks[J]. Reviews of modern physics, 2002, 74(1): 47.

29. Dorogovtsev S N, Mendes J F F. Evolution of networks[J]. Advances in physics, 2002, 51(4):1079-1187.

30. Dorogovtesev S N, Mendes J F F. Evolution of Networks Oxford University Press[J]. 2003.

31. Aggarwal C C. An introduction to social network data analytics[M]. Springer US, 2011.

32. Papadopoulos S, Kompatsiaris Y, Vakali A, et al. Community detection in social media[J]. Data Mining and Knowledge Discovery, 2012, 24(3): 515-554.

33. Adedoyin-Olowe M, Gaber M M, Stahl F. TRCM: A methodology for temporal analysis of evolving concepts in twitter[C]//Artificial Intelligence and Soft Computing. Springer Berlin Heidelberg, 2013: 135-145.

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

35. Friedman N, Getoor L, Koller D, et al. Learning probabilistic relational models[C]//IJCAI. 1999, 99: 1300-1309.

36.Kersting K, De Raedt L, Kramer S. Interpreting Bayesian logic programs[C]//Proceedings of the AAAI-2000 workshop on learning statistical models from relational data. 2000: 29-35.

37. Neville J, Jensen D, Friedland L, et al. Learning relational probability trees[C]// Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003: 625-630.

38. Burt R S. Brokerage and closure: An introduction to social capital[M]. Oxford University Press, 2005.

39. Ghosh R, Lerman K. Parameterized centrality metric for network analysis[J]. Physical Review E, 2011, 83(6): 066118.

40. Kitsak M, Gallos L K, Havlin S, et al. Identification of influential spreaders in complex networks[J]. Nature Physics, 2010, 6(11): 888-893.

41. Dolev S, Elovici Y, Puzis R. Routing betweenness centrality[J]. Journal of the

ACM (JACM), 2010, 57(4): 25.

42. Stephenson K, Zelen M. Rethinking centrality: Methods and examples[J]. Social Networks, 1989, 11(1): 1-37.

43. Newman M E J. A measure of betweenness centrality based on random walks[J]. Social networks, 2005, 27(1): 39-54.

44. Bonacich P. Factoring and weighting approaches to status scores and clique identification[J]. Journal of Mathematical Sociology, 1972, 2(1): 113-120.

45. Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine[J]. Computer networks and ISDN systems, 1998, 30(1): 107-117.

46. Lü L, Zhang Y C, Yeung C H, et al. Leaders in social networks, the delicious case[J]. PloS one, 2011, 6(6): e21202.

47. Kleinberg J M. Authoritative sources in a hyperlinked environment[J]. Journal of the ACM (JACM), 1999, 46(5): 604-632.

48. Newman M. Networks: an introduction[M]. Oxford University Press, 2010.

49. Clauset A, Newman M E J, Moore C. Finding community structure in very large networks[J]. Physical review E, 2004, 70(6): 066111.

50. Mucha P J, Richardson T, Macon K, et al. Community structure in time-dependent, multiscale, and multiplex networks[J]. Science, 2010, 328(5980): 876-878.

51. Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithms[C]//Proceedings of the 10th international conference on World Wide Web. ACM, 2001: 285-295.

52. Balabanović M, Shoham Y. Fab: content-based, collaborative recommendation[J]. Communications of the ACM, 1997, 40(3): 66-72.

53. Wang J, De Vries A P, Reinders M J T. Unifying user-based and item-based collaborative filtering approaches by similarity fusion[C]//Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2006: 501-508.

54. Zhang Y C, Blattner M, Yu Y K. Heat conduction process on community networks as a recommendation model[J]. arXiv preprint arXiv:0803.2179, 2008.

55. Zhang Y C, Medo M, Ren J, et al. Recommendation model based on opinion diffusion[J]. EPL (Europhysics Letters), 2007, 80(6): 68003.

56. Zhou T, Ren J, Medo M, et al. Bipartite network projection and personal recommendation[J]. Physical Review E, 2007, 76(4): 046115.

57. Linden G, Smith B, York J. Amazon. com recommendations: Item-to-item collaborative filtering[J]. Internet Computing, IEEE, 2003, 7(1): 76-80.

58. Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems[J]. Computer, 2009, 42(8): 30-37.

59. Kobsa A. User modeling and user-adapted interaction[C]//Conference companion on Human factors in computing systems. ACM, 1994: 415-416.

60. Huang Z, Chen H, Zeng D. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering[J]. ACM Transactions on Information Systems (TOIS), 2004, 22(1): 116-142.

61. Sarwar B, Karypis G, Konstan J, et al. Incremental singular value decomposition algorithms for highly scalable recommender systems[C]//Fifth International Conference on Computer and Information Science. 2002: 27-28.

62. Schein A I, Popescul A, Ungar L H, et al. Methods and metrics for cold-start recommendations[C]//Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2002: 253-260.

63. Lam X N, Vu T, Le T D, et al. Addressing cold-start problem in recommendation systems[C]//Proceedings of the 2nd international conference on Ubiquitous information management and communication. ACM, 2008: 208-211.

64. Zhou T, Kuscsik Z, Liu J G, et al. Solving the apparent diversity-accuracy dilemma of recommender systems[J]. Proceedings of the National Academy of Sciences, 2010, 107(10): 4511-4515.

65. Min S H, Han I. Detection of the customer time-variant pattern for improving recommender systems[J]. Expert Systems with Applications, 2005, 28(2): 189-199.

66. Xiang L, Yuan Q, Zhao S, et al. Temporal recommendation on graphs via long-and short-term preference fusion[C]//Proceedings of the 16th ACM SIGKDD

international conference on Knowledge discovery and data mining. ACM, 2010: 723-732.

67. Horak Z, Kudelka M, Snasel V. Properties of Concept Lattice Reduction Based on Matrix Factorization[C]//Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on. IEEE, 2013: 333-338.

68. Skillicorn D. Social network analysis via matrix decompositions: al Qaeda[J]. Available f rom http://www. cs. queensu. ca/home/skill/alqaeda. pdf, Aug, 2004.

69. Kruskal J B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis[J]. Psychometrika, 1964, 29(1): 1-27.

70. Borg I, Groenen P J F. Modern multidimensional scaling: Theory and applications[M]. Springer, 2005.

71. Bishop C M, Svensén M, Williams C K I. GTM: A principled alternative to the self-organizing map[M]//Artificial Neural Networks—ICANN 96. Springer Berlin Heidelberg, 1996: 165-170.

72. Bishop C M, Svensén M, Williams C K I. GTM: The generative topographic mapping[J]. Neural computation, 1998, 10(1): 215-234.

73. Kohonen T. The self-organizing map[J]. Proceedings of the IEEE, 1990, 78(9): 1464-1480.

74. Seung-Hee B, Judy Q, Geoffrey F, "Scalable Dimension Reduction for Large Abstract Data Visualization[C]," Poster at IEEE Cluster 2011 Austin, Texas, Sept. 2011.

75. Wei Z, Yanqing Y, Hanlin T, et al. Information Diffusion Model Based on Social Network[C]//Proceedings of the 2012 International Conference of Modern Computer Science and Applications. Springer Berlin Heidelberg, 2013: 145-150.

76. Apolloni A, Channakeshava K, Durbeck L, et al. A study of information diffusion over a realistic social network model[C]//Computational Science and Engineering, 2009. CSE'09. International Conference on. IEEE, 2009, 4: 675-682.

77. Lind P G, da Silva L R, Andrade Jr J S, et al. Spreading gossip in social networks[J]. Physical Review E, 2007, 76(3): 036117.

78. Bakshy E, Rosenn I, Marlow C, et al. The role of social networks in information diffusion[C]//Proceedings of the 21st international conference on World Wide Web. ACM, 2012: 519-528.

79. Nori N, Bollegala D, Ishizuka M. Exploiting User Interest on Social Media for Aggregating Diverse Data and Predicting Interest[C]//ICWSM. 2011.

80. Wu Z, Chen C. User Classification and Relationship Detecting on Social Network Site[C]//Control, Automation and Systems Engineering (CASE), 2011 International Conference on. IEEE, 2011: 1-4.

81. Pennacchiotti M, Popescu A M. A Machine Learning Approach to Twitter User Classification[C]//ICWSM. 2011.

82. Young A L, Quan-Haase A. Information revelation and internet privacy concerns on social network sites: a case study of facebook[C]//Proceedings of the fourth international conference on Communities and technologies. ACM, 2009: 265-274.

83. Shakimov A, Varshavsky A, Cox L P, et al. Privacy, cost, and availability tradeoffs in decentralized OSNs[C]//Proceedings of the 2nd ACM workshop on Online social networks. ACM, 2009: 13-18.

84.Oentaryo R J, Lim E P, Lo D, et al. Collective churn prediction in social network[C]//Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012). IEEE Computer Society, 2012: 210-214.

85. Richter Y, Yom-Tov E, Slonim N. Predicting Customer Churn in Mobile Networks through Analysis of Social Groups[C]//SDM. 2010: 732-741.

86. Ngonmang B, Viennet E, Tchuente M. Churn prediction in a real online social network using local community analysis[C]//Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012). IEEE Computer Society, 2012: 282-288.

87. Berge C. Graphs and hypergraphs[M]. Amsterdam: North-Holland publishing company, 1973.

88.Berge C. Hypergraphs: combinatorics of finite sets[M]. Elsevier, 1984.

89. Backstrom L, Boldi P, Rosa M, et al. Four degrees of separation[C]//Proceedings of the 4th Annual ACM Web Science Conference. ACM, 2012: 33-42.

90. Ugander J, Karrer B, Backstrom L, et al. The anatomy of the facebook social graph[J]. arXiv preprint arXiv:1111.4503, 2011.

91. Palchykov V, Kaski K, Kertész J, et al. Sex differences in intimate relationships[J]. Scientific reports, 2012, 2.

92. Dunbar R. How many friends does one person need? Dunbar's number and other evolutionary quirks[M]. Faber & Faber, 2010.

93. Plate, Erich J. Methods of investigating urban wind fields—physical models[J]. Atmospheric Environment ,1999,33.24 : 3981-3989.

94. Adamic L A, Adar E. Friends and neighbors on the web[J]. Social networks, 2003, 25(3): 211-230.

95. Narang K, Lerman K, Kumaraguru P. Network flows and the link prediction problem[C]//Proceedings of the 7th Workshop on Social Network Mining and Analysis. ACM, 2013: 3.

96. Rowe M, Stankovic M, Alani H. Who will follow whom? exploiting semantics for link prediction in attention-information networks[M]//The Semantic Web–ISWC 2012. Springer Berlin Heidelberg, 2012: 476-491.

97. Scellato S, Noulas A, Mascolo C. Exploiting place features in link prediction on location-based social networks[C]//Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011: 1046-1054.

98. Brohee S, Van Helden J. Evaluation of clustering algorithms for protein-protein interaction networks[J]. BMC bioinformatics, 2006, 7(1): 488.

99. Yu H, Braun P, Yıldırım M A, et al. High-quality binary protein interaction map of the yeast interactome network[J]. Science, 2008, 322(5898): 104-110.

100. A. Clauset, C. Moore, and M.E.J. Newman. Hierarchical structure and the prediction of missing links in networks[J]. Nature, 2008, 453, 98.

101. Huang Z, Li X, Chen H. Link prediction approach to collaborative filtering[C]//Proceedings of the 5th ACM/IEEE-CS joint conference on Digital

libraries. ACM, 2005: 141-142.

102. 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.

103. Al Hasan M, Chaoji V, Salem S, et al. Link prediction using supervised learning[C]//SDM'06:Workshop on Link Analysis, Counter-terrorism and Security. 2006.

104. Backstrom L, Leskovec J. Supervised random walks: predicting and recommending links in social networks[C]//Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011: 635-644.

105. Barandela R, Valdovinos R M, Sánchez J S, et al. The imbalanced training sample problem: Under or over sampling?[M]//Structural, Syntactic, and Statistical Pattern Recognition. Springer Berlin Heidelberg, 2004: 806-814.

106. Newman M E J. Clustering and preferential attachment in growing networks[J]. Physical Review E, 2001, 64(2): 025102.

107. Weng J, Lim E P, Jiang J, et al. Twitterrank: finding topic-sensitive influential twitterers[C]//Proceedings of the third ACM international conference on Web search and data mining. ACM, 2010: 261-270.

108. Kwak H, Lee C, Park H, et al. What is Twitter, a social network or a news media?[C]//Proceedings of the 19th international conference on World wide web. ACM, 2010: 591-600.

109. Kullback, S. The kullback-leibler distance. American Statistician,1987, 41(4):340-340.

110. Liu W, Lü L. Link prediction based on local random walk[J]. EPL (Europhysics Letters), 2010, 89(5): 58007.

111. Goel S, Watts D J, Goldstein D G. The structure of online diffusion networks[C]//Proceedings of the 13th ACM conference on electronic commerce. ACM, 2012: 623-638.

112. Goodman L A. Snowball sampling[J]. The annals of mathematical statistics, 1961: 148-170.

113. Yang J, Leskovec J. Patterns of temporal variation in online media[C]//Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011: 177-186.

114. Ramos J. Using tf-idf to determine word relevance in document queries[C]//Proceedings of the first instructional conference on machine learning. 2003.

115. Zhang Z K, Yu L, Fang K, et al. Website-oriented recommendation based on heat spreading and tag-aware collaborative filtering[J]. Physica A: Statistical Mechanics and its Applications, 2014, 399: 82-88.

116. Wold S, Esbensen K, Geladi P. Principal component analysis[J]. Chemometrics and intelligent laboratory systems, 1987, 2(1): 37-52.

117. Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. the Journal of machine Learning research, 2003, 3: 993-1022.

118. Hong L, Davison B D. Empirical study of topic modeling in twitter[C]//Proceedings of the First Workshop on Social Media Analytics. ACM, 2010: 80-88.

119. Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

120. Hanley J A, McNeil B J. The meaning and use of the area under a receiver operating characteristic (ROC) curve[J]. Radiology, 1982, 143(1): 29-36.

121. Lü, L., Medo, M., Yeung, C. H., Zhang, Y. C., Zhang, Z. K., Zhou, T. Recommender systems[J]. Physics Reports, 2012, 519:1-49.

附录一：研究生期间取得的科研成果及获奖情况

1.作者简介：

尤志强（1990），男，浙江金华人，硕士研究生，主要从事复杂网络与数据挖掘。

1.研究生期间获得的科研成果及获奖情况：

(1) **尤志强**, 韩筱璞. 基于相关性的上海世界博览会行人流分析[J]. 上海理工大学学报, 2013, 35(4): 313-320.

(2) Zhang, Z. K., Yu, L., Fang, K., **You, Z. Q**., Liu, C., Liu, H., & Yan, X. Y. (2014). Website-oriented recommendation based on heat spreading and tag-aware collaborative filtering. *Physica A: Statistical Mechanics and its Applications*,*399*, 82-88.

(3) **尤志强**, 管远盼,韩筱璞, 吕琳媛.基于社交网络的社群生长模型[J]. 复杂系统与复杂性科学, 2015, 12(2):72-77.

(4) **尤志强**,朱燕燕,韩筱璞,吕琳媛.基于任务队列的新闻报道模型.电子科技大学学报,2014 年 11 月接收.

(5) **You Z Q**, Zhu Y Y, Han X P, Linyuan Lu, Modelling temporal patterns of news report,the 34th Chinese Control Conference and SICE Annual Conference 2015, Accepted.

(6) **You Z Q**, Han X P, Lü L, et al. Empirical studies on the network of social groups: the case of Tencent QQ[J]. arXiv preprint arXiv:1408.5558, 2014. Plos One, Under Review. (SCI,IF=3.534)

该篇论文同时被 CCCN2014 全国第十届复杂网络大会录用为**报告论文**

(7) **You Z Q**, Han X P, Modeling for academic competition:efficiency dominates scholars'outputs, *Physica A: Statistical Mechanics and its Applications*, Under Review.(SCI,IF=1.722)

(8) **You Z Q**, Zhou G, Zhang Z K, Shen Z S, Wang W X, A new indice of link

prediction for social netwok, In Preparation, Scientific Reports. (SCI,IF=5.078)

(9) Yuan-pan Guan, **Zhi-qiang You,** Xiao-pu Han. Detecting the growth of groups based on social network[J].Physica A: Statistical Mechanics and its Applications, Under Review.(SCI,IF=1.722)

(10) **主持** 浙江省大学生科技创新项目《基于社交网络的价值流失预警模型的研究》

(10) CCF-腾讯犀牛鸟基金项目《大规模在线社交网络中高影响力用户的挖掘》核心成员，团队并获得犀牛鸟优秀奖

(11) 2012-2013 年参与编写《杭州市互联网经济发展报告》，杭州市经济和信息化委员会主导，目前已由浙江大学出版社出版。

（12）首届中国互联网数据平台数据挖掘竞赛全国第一名

团队共三人，核心成员

（13）第十届华为杯全国研究生数学建模竞赛全国三等奖

团队共三人，担任队长