社交网络分析综述

乔婷婷1）陈潇雅2）吉梦瑶3）

（1，2，3）（北京理工大学计算机学院 北京 100081）

摘要：随着互联网的发展各种社交媒体应用风靡起来，在互联网出现的短短数十年内社交媒体应用已经拥有数十亿用户并对现实社会的方方面面产生着影响。为了更好的利用社交网络服务所产生的大量数据，使其产生价值，消除危害，所以产生了社交网络分析这门科学。本文学习研究了国内外关于社交网络分析的经典文献和前沿研究成果，整理了社交网络分析的一些基本概念，并就社交网络中影响力分析、推荐算法和信息传播三个方面进行了较为深入的探索，最后介绍了社交网络分析在构建用户画像、社交媒体服务、和公共安全三个领域的应用。

关键字：社交网络、影响力分析、推荐算法、信息传播

Overview of Social Network Analysis

Qiao TingTing1）Chen XiaoYa2）Ji MengYao3）

(1,2,3)(School of Computer Science, Beijing Institute of Technology, Beijing 100081, China)

Abstract：With the development of the Internet, various social media applications have become popular. In the short span of decades, social media applications have had billions of users and have an impact on all aspects of the real world. In order to make better use of the large amount of data generated by social network services to generate value and eliminate harm, the science of social network analysis has emerged. In paper, we studies the classic literature and frontier research results of social network analysis at home and abroad, sorts out some basic concepts of social network analysis, and carries out in-depth research on the three aspects of influence analysis, recommendation algorithm and information dissemination in social networks. Exploring, finally, introduced the application of social network analysis in the construction of user portraits, social media services, and public security

Keywords：Social network, Impact analysis, Recommended algorithm, Dissemination of information

# 研究背景

随着互联网的发展各种社交媒体应用风靡起来，在互联网出现的短短数十年内社交媒体应用已经拥有数十亿用户并对现实社会的方方面面产生着影响。

FaceBook和微信一般是为了沟通交流，是关于人与人之间的强关系网络，划分社交圈有助于朋友间相互推荐。领英则是基于职业需要的社交而形成的社交网络，有助于求职和商务合作交流。 微博，Twitter，豆瓣，微信公众号是关于关注与被关注的的弱关系网络，有助于消息和知识的传播。

在2016年的美国总统大选中，“推特治国”的总统特朗普就很好地将其作为宣传工具，打造了一个亲民的总统形象，赢得了选举；2018年11月21日，DG设计师嘉班纳被爆出在ins上的辱华言论，再加上其广告中的不当措辞，DG辱华事件爆发，微博等社交媒体讨论热烈，导致其时装秀取消，产品线上线下全面下架。

这些热点事件无一不是在社交网络服务上迅速发酵，并最终对现实社会产生影响。不论大事小情，这种线上影响线下的趋势越来越明显。

本文所说的社交网络指的是在线社交网络，在线社交网路具有迅捷性、蔓延性、平等性与自组织性等四大特点。为了利用好社交网络的特性，产生价值，消除危害，所以产生了社交网络分析这门科学。它是一种基于信息学、数学、社会学、管理学和心理学等科学的交叉科学。

根据社区可以对人进行精准化营销，推荐个性化的商品和服务，比如京东；

疾病传播也是由一个中心点向外扩散，切断网络中的关键节点就可以有效阻止传染病的传播；

识别互联网金融行业中的欺诈团伙，进行反欺诈预测；

# 基本概念

## 社交网络

社交网络是指“由许多节点构成的一种社会结构。节点通常是指个人或组织，而社交网络代表着各种社会关系。[1]”在互联网诞生前，社交网络分析是社会学和人类学重要的研究分支。社交网络分析是用来查看节点、链接之间的社会关系的分析方式。节点是网络中的个人参与者，链接则是参与者之间的关系。节点之间可以有很多种链接。一些学术研究已经显示，社会网络在很多层面运作，从家庭到国家层面都有，并扮演着关键作用，决定问题如何得到解决，组织如何运行，并在某种程度上决定个人能否成功实现目标。

而随着互联网的迅速发展，现在的社交网络分析几乎与在线社交网络分析（Online Social Network Analysis）等同，该门科学的发展是随着在线社交服务（Social Network Service, SNS）的出现而诞生。

在线社交服务的种类大致可分为四种：即时消息类应用（QQ、微信、WhatsApp、Skype 等），在线社交类应用（QQ空间、人人网、Facebook、Google+ 等），微博类应用（新浪微博、腾讯微博、Twitter 等），共享空间类应用（论坛、博客、视频分享、评价分享等）。

社交网络模型许多概念来自于图论，这里我们把人看作节点，把人与人之间的社交关系看作边，这样社交网络模型就是一个由节点（人）和边（社交关系）组成的图。图又可以分为无向图和有向图。

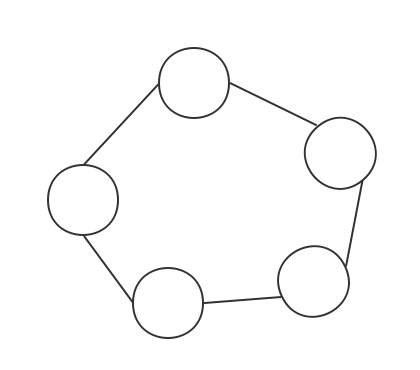


图 2-1-1 无向图

无向图仅表示节点和节点之间是否有关系，将这种关系映射到社交网络中，比如在P2P行业反欺诈建模中，可以通过申请者通讯录去获取其社会关系，例如，如果张三和李四的通讯录都有老赖王五，那么，张三和李四的贷款申请违约风险就会比较高。

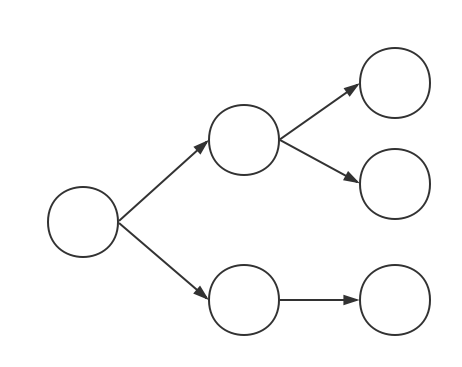


图2-1-2 有向图

有向图相比于无向图会携带方向信息，将其映射到社交网络中，最为人们广泛知道的是传销网络的上下线制度。这是一种发展团队十分迅速有效的手法，现在也被互联网公司广泛用于发展用户——好友邀请制度。

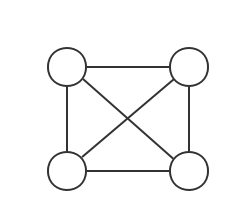


图2-1-3 完全子图

完全子图是一个网络图中任意两个节点都相连的子图。在社交网络中一个最大的完全子图就是一个派系，派系是指一群彼此紧密相连、有依赖性的行动者所构成的群体。

## 统计特性

由于社交网络模型的许多概念来自图论，因此，其统计特性也与图的一些统计概念相关。

度（Degree）：节点的度定义为与该节点相连的边的数目。网络平均度反应了网络的疏密程度，而通过度分布则可以刻画不同节点的重要性。

网络密度（Density）：网络密度可以用于刻画节点间相互连边的密集程度，定义为网络中实际存在边数与可容纳边数上限的比值，常用来测量社交网络中社交关系的密集程度及演化趋势。

聚类系数（Clustering Coefficient）：用于描述网络中与同一节点相连的节点间也互为相邻节点的程度。其用于刻画社交网络中一个人朋友们之间也互相是朋友的概率，反应了社交网络中的聚集性。

介数（Betweeness）：为图中某节点承载整个图所有最短路径的数量，通常用来评价节点的重要程度，比如在连接不同社群之间的中介节点的介数相对于其他节点来说会非常大，也体现了其在社交网络信息传递中的重要程度。

## 虚拟社区

虚拟社区基于子图局部性的定义：社区结构是复杂网络节点集合的若干子集，每个子集内部的节点之间的连接相对非常紧密，而不同子集节点之间的连边相对稀疏。

在社交网络中发现虚拟社区有助于理解网络拓扑结构特点，揭示复杂系统内在功能特性，理解社区内个体关系。为信息检索、信息推荐、信息传播控制和公共事件管控提供有力支撑。

同时值得关注的是社区并不是一成不变，而是在不断演化的，由于一些事件的发生，社交关系也在不断变化。

在线社交网络中存在着大量显性或者隐性的虚拟社区结构，这些虚拟社区结构并不是永恒不变的，随着事件变化，社区结构也在不断演变。分析动态的虚拟社区结构演化有助于理解整个社交网络的演化过程，所以有着重要的研究价值。

### 虚拟社区的涌现

虚拟社区涌现即在社交网络中虚拟社区从无到有的过程，其最重要的特征是网络聚集现象。

周期闭包：所谓周期闭包，是指网络节点倾向于和自己在网络中邻居的邻居建立连接关系而形成的结构，该机制是导致虚拟社区形成的主要因素。实验表明三元闭包的出现概率随着两个节点之间测地距离的增减呈指数递减。相反地，焦点闭包和测地距离无关，其生成原因是两个节点之间有共同的兴趣或参与共同的活动。

偏好连接：在很多真实网络中，新增加的边并不是随机连接的，而是倾向于和具有较大度数的连接。

### 虚拟社区的演化

在线社交网络虚拟社区演化过程非常复杂，影响因素很多。如何挖掘虚拟社区演化中的关键性因素成为社交网络研究中一个重要而有挑战性的课题， 用户个体的累积效应、结构多样性和结构平衡性三个基本因素对虚拟社区演化都存在影响。

# 关键技术

社交网络分析的相关应用领域非常宽泛，所以相关技术也非常多。本综述主要分析社交网络在影响力分析、推荐和信息传播方面的关键技术及其实现。

## 影响力分析方面

食物链中哪些物种对整个生态的影响最大?

“种子短信”发给哪些手机用户可以获得更多的转发?

全球经济体系中哪些国家或地区对于体系的健康发展至关重要?

当传染病来临的时候, 我们应该采取何种接种免疫策略来避免其大规模爆发?

高价雇佣微博大号做新产品的推广和营销真的有用吗? 如果有用, 又如何找到合适的达人?

为什么俄亥俄州克利夫兰市的几条烧断的高压线能够造成北美大停电事故, 导致数百亿美元的损失?

借助网络科学的发展, 对于这些问题, 如今我们已经有了一些定量化的描述和解决办法。实际上, 几乎所有的复杂系统(比如社会、生物、信息、技术、交通运输系统)都可以自然地表示为网络。其中, 节点代表系统的各种构成要素, 节点间的连边表示要素之间的联系。最复杂的人类社会系统就可以用一个社会网络刻画, 节点是人, 人与人之间的各种关系构成社会网络中的链接。

应用复杂网络的理论和方法能够帮助我们更好地理解这些复杂系统的特征，并对其进行更好地预测和控制。如上述的大停电事故归根到底是“网络相继攻击的脆弱性”问题。 我们能够事先对这个电力网络的结构有所了解, 并找到关键的地区采取预防措施, 就可能避免如此巨额的经济损失。这里最核心的问题就是如何识别这些重要的节点。

所谓的重要节点是指相比网络其他节点而言能够在更大程度上影响网络的结构与功能的一些特殊节点。这里的网络结构包括度分布、平均 距离、连通性、聚类系数、度相关性等, 网络功能涉及网络的抗毁性、传播、同步、控制等。重要节点一般数量非常少, 但其影响却可以快速地波及到网络中大部分节点。例如, 在对一个无标度网络的蓄意攻击中, 少量最重要节点被攻击就会导致整个网络瓦解；微博中最有影响力的几个用户所发的微博很快就能传遍整个网络；仅仅1%的公司却控制着40%的全球经济。可见重要节点对网络的结构和功能有着巨大的影响, 节点重要性的排序和重要节点的挖掘意义重大。

随着互联网的快速发展，社交网络节点影响力的研究成为热点。影响力可以是一种能够在网络连边传递的能量，个体影响力的大小为传递能量的程度。基于该思想，顾亦然等人提出了基于介数中心性和节点贡献度的个体对群体影响力算法。通过与介数中心性、k-shell 和 PageRank 算法的对比以及传染病传播模型的传播实验表明，该算法能够有效、准确的表征个体对群体的影响力；通过对个体对群体的影响力各区间段节点频数分布研究发现，其值呈现幂律分布，具有无标度现象，即少数节点具有较大影响力，普通节点虽然数量较大，但其影响力普遍较小。

为了深入研究及分析社交网络的节点重要性，学者们从网络结构等方面对节点影响力进行了广泛研究。其中，度中心性、PageRank、k-shell、介数中心性等算法刻画了节点在网络拓扑结构上的重要程度。度中心性(Degree Centrality ,记为 DC)刻画网络的局部特性，无法从全局刻画网络特征；PageRank 算法认为节点的重要性取决于网络中指向该节点的节点数量和质量，容易陷入悬挂节点；k-shell 是将网络一层层分解找出不同层次不同 影响力的节点，其度量重要性比较粗粒化。

### 度中心性

社会网络分析中, 节点的重要性也称为“中心性”, 其主要观点是节点的重要性等价于该节点与其他节点的连接使其具有的显著性。度中心性(degree centrality)认为一个节点的邻居数目越多, 影响力就越大, 这是网络中刻画节点重要性最简单的指标。

节点v的度数指与节点v相连的节点个数 ，即。一个节点有可能的最大连接数为 n-1，因此标准化后，节点v的度中心性为:

（1）

在有向网络中入度和出度有不同的意义(如社交网络中入度代表受欢迎程度, 出度代表合群程度), 一般会分别计算入度和出度的中心性。度中心性指标拥有简单、直观、计算复杂度低等特点. 在网络鲁棒性和脆弱性研究中, 针对无标度网络或指数网络, 如果攻击前一次性选择若干个攻击目标, 采用度中心性指标的攻击效果比介数中心性、接近中心性、特征向量中心性要好。度中心性指标的缺点是仅考虑了节点的最局部的信息, 是对节点最直接影响力的描述, 没有对节点周围的 环境(例如节点所处的网络位置、更高阶邻居等)进行更深入细致地探讨, 因而在很多情况下不够精确。

### 紧密度中心性

设结点，是图G 中的两个节点，如果是中心结点， 是非中心结点，则 比更快到达图G 中的剩余节点。这是紧密度中心性的直观表示。严格的描述，紧密度中心性是衡量节点与图中剩余其它的节点之间联系是否紧密的一个标准。而联系是否紧密是用到其它节点的平均路径长度来衡量的。将其形式化后表示如下：

其中，是节点的紧密度中心性，也就是其到其它节点的平均路径长度，表示节点 到节点的最短路径，n表示图中节点的个数。

### PageRank

PageRank 算法, 是谷歌搜索引擎的核心算法。传统的根据关键字密度判定网页重要程度的方法容易受到“恶意关键字”行为的诱导, 使搜索结果可信度低。 PageRank算法基于网页的链接结构给网页排序, 它认为万维网中一个页面的重 要性取决于指向它的其他页面的数量和质量, 如果一个页面被很多高质量页面指向, 则这个页面的质量也高。初始时刻, 赋予每个节点(网页)相同的 PR 值, 然后进行迭代, 每一步把每个节点当前的 PR 值平分给它所指向的所有节点。每个节点的新 PR 值为它所获得的 PR 值之和, 于是得到节点 在 t 时刻的 PR 值为

（2）

其中为节点 的出度。迭代直到每个节点的 PR 值都达到稳定时为止。公式(2)的缺陷在于 PR 值一旦到达某个出度为零的节点(称为悬挂节点 Dangling node), 就会永远停留在该节点处而无法传递出来, 从而不断吸收 PR 值。为解决这一问题, PageRank 算法在上述过程基础上引入一个随机跳转概率c。每一步, 不管一个节点是否为悬挂节点, 其 PR 值都将以 c 的概率均分给网络中所有节点, 以 1−c 的概率均分给它指向的节点。该过程实际上是考虑到了现实中网络用户除了通过超链接访问页面之外, 还可以通过直接输入网址的形式对网页进行访问的行为, 从而保证了即使是没有任何入度的网页也有机会被访问到。其实质是将有向网络变成强连通的, 使邻接矩阵成为不可约矩阵, 保证了特征值 1 的存在. 由此可得含参数 c 的 PageRank 算法:

 （3）

参数 c 的取值要视具体的情况而定。c 取值越大，收敛越快, c=0 时回到公式(2)。c 取值越大算法的有效性越低, c=1 时所有节点都有相同的 PR 值. 针对万维网的网页排序, 以前的研究显示, c=0.15 是一个比较好的参数。

PageRank 算法作为谷歌搜索引擎的核心算法, 它在商业应用上的极大成功激发了人们深入研究 PageRank 的热忱, 研究者们提出了一系列基于 PageRank 的改进算法。例如 Kim 和 Lee为了避免悬挂节点囤积 PR 值的问题, 将每一步到达悬挂节点的 PR 值平均分给网络中的 n 个节点, 即将概率转移矩阵中悬挂节点所在的列的 n 个元素修改为 1/n; PageRank 中从一个网页上的链接中挑选下一个访问目标时是等概率的, Zhang 等人认为这 n 个目标网页出度越大的越有可能被点击, 并提出 N-step PageRank 算法用以描述这一思想。2012 年 Brin 和 Page以相同的题目重新出版了当年提出 PageRank 算法的博士学位论文, 在文中他们对这十几年的网页排序算法进行了回顾, 并就如何用 PageRank 实现大规模搜索进行了深入讨论。另外, 作为有向网络节点排序最经典的算法, PageRank 及其改进算法广泛应用于其他领域, 如对期刊的排序、对社交网络上用户的排序、对风投公司(VC)的排序、对科学论文的排序以及科学家影响力的排序等。

### k-shell

度中心性仅考察节点最近邻居的数量, 认为度相同则重要性相同。然而, 近期的一些研究表明在刻画节点重要性的时候节点在网络中的位置也是至关重要的因素。在网络中, 如果一个节点处于网络的核心位置, 即使度较小, 往往也有较高影响力; 而处在边缘的大度节点影响力往往有限。基于此, Kitsak 等人提出用 k-shell分解法(k-shell decomposition)确定网络中节点的位置, 将外围的节点层层剥去, 处于内层 的节点拥有较高的影响力。

K-shell 分解方法给出了节点重要性的一种粗粒化的划分。其基本思想如下，假设边缘节点的 K-shell 值为1，然后往内一层层进入网络的核心，先去除网络中度值等于 1 的所有节点以及连边。 若剩下的节点里面，仍有度值等于 1 的节点，则重复上述操作，即去除这些节点和连边，直至所有节点的度值都大于 1，把这 些去除的节点的 K-shell 值记为 1，也就是说这些节点均处于 ks 值为 1 的层。然后依次去除度值小于或等于 k 的节点及连边（k 为整数，k≥2），直到所有的节点都有对应的 ks 值为止。图2-1给出一个 k-shell分解的示例。其中(a)为原网络，(b), (c), (d) 别表示 1-壳,2-壳和 3-壳。

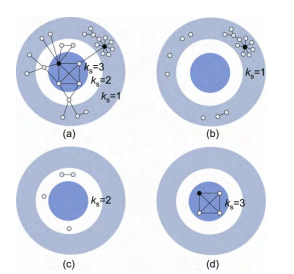


图2-1 一个可分解为3层壳的简单网络

整个函数最难实现的地方在于多次重新计算节点的度值并与关键的 key 值进行比较，重复删除节点与连边，然后循环比较寻找。其中，每一次的循环计算，整个社交网络图谱中的关系矩阵时刻在发生改变。因此，为了实现的简便，邵浩等人用（N+1）\*（N+1）的矩阵来存放图中所有节点与其连边的关系，其中矩阵的0号位置保存当前度值，另外确保了数组矩阵中节点的编号与位置相对应，均从1号位置开始计算。 这样，每一次删除节点与其连边的时候，只需要将矩阵对应位置的数值清 0 即可，再重新计算节点的度值。

图2-2中描述的是 K-shell 分解方法实现流程图，图中可以看出实现的难度在于多层循环的嵌套，导致逻辑实现的困难。但是最主要的地方在于每一次计算图中节点的度值，并进行多次的比较，然后挑选出符合条件的节点，然后删除该节点与其连边，重新计算度值。

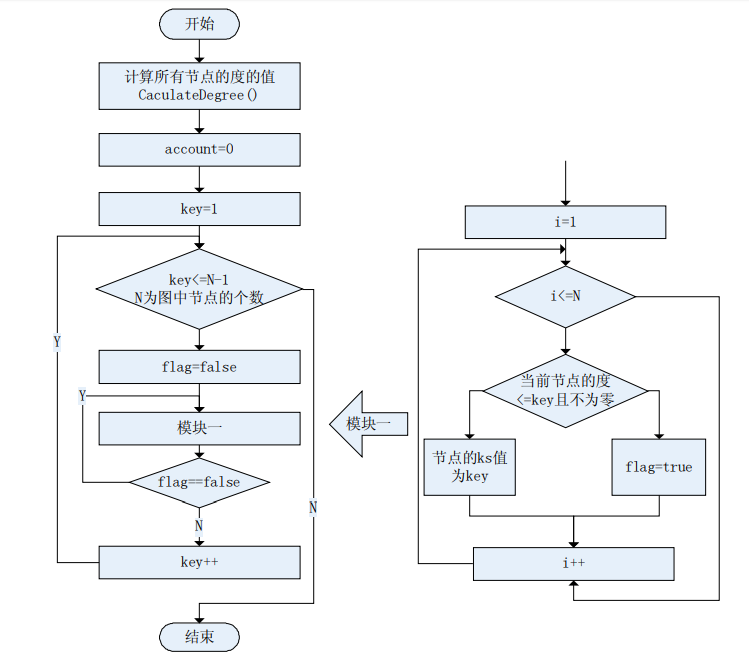


图2-2 k-shell 分解算法实现流程图

k-壳分解法计算复杂度低, 在分析大规模网络的层级结构等方面有很多应用。然而, 此方法也有一定局限性。第一，k-shell分解法有很多不能发挥作用的场景。比如在树形图, 规则网络和 BA 网络中, 所有(或大部分)节点都会被划分在同一层。更极端的例子是星形图, 显然中心节点有最强的传播能力, 但是 k-shell分解的时候, 星形网络的所有节点会被划分在同一层(ks=1)。第二, k-shell分解法的排序结果太过粗粒化, 使得节点的区分度不大。k-shell分解法划分的层级比度中心性方法划分的层级少很多, 很多节点处在同一层上, 它们之间的重要性难以比较。第三, k-shell分解法在网络分解时仅考虑剩余度的影响，这相当于认为同一层的节点在外层都有相同的邻居数目, 显然不合理。Zeng 等人提出了在每一步剥去一部分外围节点之后, 同时考虑节点剩余的邻居数和节点已经移除的邻居数的方法,这种k-shell分解法能够很好地区分树形图以及 BA 网络中不同节点的传播能力, 并且分层的层数大大增加(甚至可超过度中心性), 提高了节点传播能力的区分度。另外, Liu 等人指出壳数相同的节点传播能力差距可能很大, 并提出了一种可以进一步区分具有相同壳数的节点的传播能力的排序方法, 从而较 Kitsak 等人的方法有所进步; Hu 等人将 k-shell分解法与社区结构相结合, 提出一种改良指标, 在 SIR 模型上的实验表明该方法较 Kitsak 等人的方法略佳。

### 介数中心性

节点的介数中心性（Betweenness centrality，BC）指网络中所有最短路径中经过该节点的路径占所有节点对的最短路径总数的比例，是以经过某一节点的最短路径数目来刻画节点重要性的指标。它体现了网络中节点对网络信息流动的影响力。该项指标表明了信息传播到该节点的可能性。节点 v 的介数中心性为：

（4）

其中， 为从节点  到节点  最短路径的数目， 为从节点  到节点  最短路径  中有  个经过节点 。

整个算法函数最难的地方在于需要记录每一对节点的最短路径，而根据常见的最短路径算法，例如迪杰斯特拉算法，只能计算指定节点的一条最短路径。 因此，首先需要对迪杰斯特拉算法进行改进，使其将计算出的所有节点的最短路径都输出， 为了后续介数中心性算法的实现，可以用一个二维数组记录每对节点对的所有最短路径的相关信息。再对该二维数组的信息进行统计，得出每对节点对之间所有的最短路径数，以及每条最短路径经过指定节点的个数。之后，利用一个三维数组，记录所有节点对的最短路径的条数和其中经过指定节点的最短路径的条数。

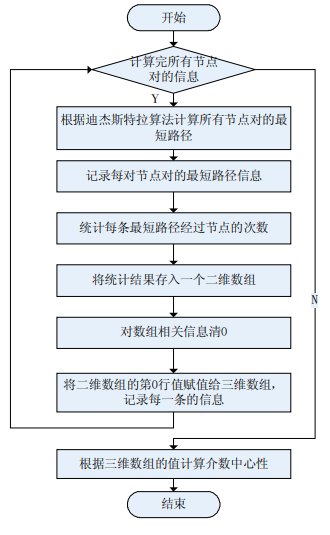


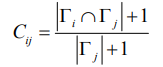
图2-3 介数中心性算法实现流程图

介数中心性可用于设计网络的通信协议、优化网 络部署、检测网络瓶颈等。王延庆将介数应用于负载网络, 提出用过载函数法研究网络的连接失效问题。此外, Goh 等人提出的负载中心性(traffic load centrality)采用类似网络中信息包传递的机制：每一对节点之间沿着最短路径传输一个单位的网络流, 如果最短路径不止一条, 则在几条最短路径的分叉处将网络流平均分配到这些最短路径上。忽略时延, 网络中所有节点对之间都互不干扰地传输一个单位的信息流时, 一个节点上传输过的网络流的数量称为该节点的负载。一个节点的负载越大, 该节点就越重要。介数中心性的计算时间复杂度较高, 使其在实际应用中受到限制, 相关讨论可参见文献。

### 个体对群体影响力算法

#### 算法分析

社交网络中节点之间的共同邻居越多说明节点间的关系越密切。其对彼此的影响力就不仅与自身所处的拓扑位置有关，还与共同邻居作用有关。现实生活中，节点  对  的影响程度和节点  对  的影响程度是存在区别的，因此引入贡献度这一概念，节点  对  的贡献度定义见公式（5）， 其中  和  分别为节点  和  邻居节点的集合。

（5）

如果说个体间的影响力刻画的是节点之间影响力传递能力的大小，那么个体对群体影响力就是刻画节点影响力的全局特性，即节点的影响力扩散对全网的影响程度。

定义：个体对群体影响力为节点  对群体 R 中所有其它节点的影响力之和，记为 。 在实际计算中， 从源节点依次计算至整个 群体，显然复杂度太高。在研究中发现并非要计算至整个网络，原因有二：其一，微信以及 Facebook 最新数据分析，复杂在线网络两节点间平均距离为 3 左右；其二，实际计算比对中发现，路径距离 >3 时，路径权重α 的大小对于计算结果排序并无明显影响。基于上述分析，顾亦然等人将  进一步定义如下：个体  对群体 R 的影响力  为  对距源节点路径距离  ≤ 3 所有节点的影响力之和。

公式为：

（6）

其中，为距节点  路径距离 ≤ 3的节点集合。

#### 算法步骤

输入：G=(V,E)

输出：节点  对群体 R 的影响力

1. 开始，通过广度优选搜索算法获取节点 到集合 中节点的路径与距离表 D\_path

2. 计算网络节点间贡献度矩阵

3. 计算网络每节点 BC值

4. 计算节点  到集合 中每个节点的影响力 

5. 由公式（6）将每条路径影响力加和，求得 

6. END

#### 仿真分析

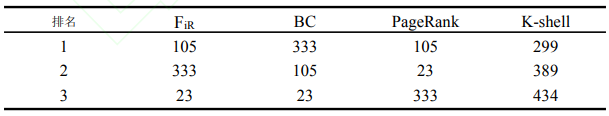
算法刻画的是节点影响力的全局特性，即节点的影响力扩散对全网的影响程度。显然，个体对群体的影响力是衡量节点重要性的指标之一。另外影响力大的节点往往是信息传播中的重要节点，信息传播能力强，传播至整个群体 速度较快。基于这一思想，顾亦然等人使用不同网络数据，通过经典的传染病模型进行仿真可以较为直观的分析节点信息传播情况，以验证算法的有效性。此实验使用的网络分别 email 和 Zachary 网络，两网络参数如表 2-1 所示。 此次实验首先使用 SI 传染病模型，对 email 网络进行传播验证，其传播概率为β=0.03。

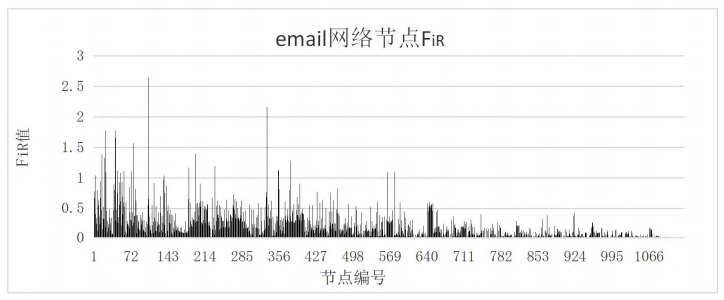
表2-1 各网络参数



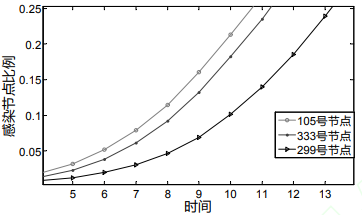
对email 网络进行 、BC、PageRank和K-shell 算法分析，结果如表 2-2 所示，其中  分析结果如图 2-4（a）所示。

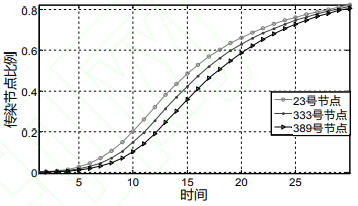
表2-2 email网络各算法排名前三节点





（a）email网络各节点

（b）email网络105,333,299 号节点传播结果



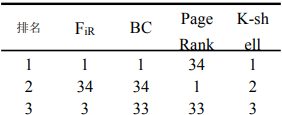
(c) email 网络 23,333,389 号节点传播结果

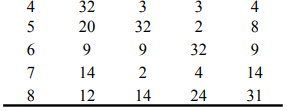
图 2-4 email 网络部分节点传播图

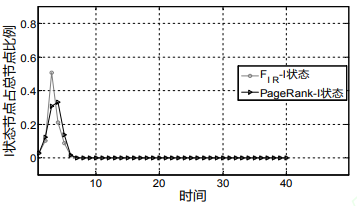
根据表 2-2 排序结果，首先将四种算法排在第一位的 105、333、299 号节点进行传播仿真，仿真结果如图 2-4(b)所示，从中可以看出 计算出来的 105 号节点在传播初期显优于其他影响力指标得出的节点；基于上述的思想将  计算出来的 23、333 节点以及 K-shell 排序第二的 389 号节点进行传播仿真实验，仿真结果如图 2-4(c)所示，从图中亦能看出在整个传播仿真中 23 号节点较优于 333 号节点 和 389 号节点；综上可知，此算法能够较为准确的衡量个体对群体的影响力。

在 Zachary 网络中分别计算节点的 ，BC， PageRank 和 k-shell 得到结果如表 2-3 所示。为更全面验证本算法，本小结传播仿真采用 SIR 传播。 将本算法排名前 20%节点分别和其他三个算法得出的排名前 20%节点传播比对，传播比对过程中去掉两算法中相同的节点，选取时间步为 40，传播 100 次取平均，结果如图 2-5 所示。图 2-5（a）为本算法与 BC 得出的节点传播对比图，由图可知本算法选取的节点不但感染峰值比例高于后者，而且到达峰值的时间也比 BC 的早；图 2-5（b）两者的差异更为明显；图 2-5（c）中两个最大感染比例差异虽没有图 2-5（a）、2-5（b）大，但到达最大值的时间却比较滞后。通过上述分析对比，本算法在 SIR 传播仿真中依然有较好表现。

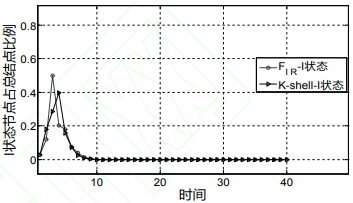
表2-3 Zachary 网络各算法排名前 20%节点







(b) 与 PageRank 排名前 20%对比



(c) 与 k-shell 排名前 20%对比

图 2-5 Zachary 网络传播仿真结果对比

## TPCF推荐算法

在互联网时代，线上资源信息的数据量急速膨胀。海量信息及资源为用户提供丰富服务的同时，也为用户便利地选择符合自身需求的服务带来了挑战。针对数据超载问题，做到安全、快速、准确地为用户进行服务推荐已经成为目前网络服务研究的热点之一。传统推荐系统多是基于协同过滤算法、内容数据过滤算法以及混合推荐算法等进行实现的。其中，协同过滤算法较为直观、准确，也是目前使用较为广泛的一种算法。其特点是只需处理用户对项目的评分数据，不需深入考虑用户及项目特性就可进行推荐。然而，协同过滤算法存在数据稀疏、可用性差以及冷启动等问题。

为提高服务推荐质量并解决传统 算法出现的问题，国内外研究人员做了大量的研究工作。Liu 等通过使用社交亲密度和个体用户影响力建立模型，设定多个端对端的性能属性约束，从而寻找复杂社交网络中的最 佳信任路径，得到最佳推荐。Gartrell等通过一个共识函数为每个组员评分，根据组内的属性及关系来获取组预测评分， 以建立推荐系统。Tang等提出了结合全局信任和局部信任关系从而给出信任值的信任度衡量标准。Zhao等提出通过加入社交网络连接的SBPR（社交贝叶斯个性化排名）模型来缓解冷启动等问题。Golbeck等提出了采用扩展广度优先搜索算法获得推荐者列表的TodalTrust算法，以最短路径距离寻找推荐者，并以活跃用户与推荐者之间的信任值为权重预测目标项 目评分。Hang通过准确惩罚或奖励服务 来提供服务质量，从而建立信任模型。jalami等提出了TrustWalker方法，其以随机漫游搜索信任网络，利用项目间的相似性来降低搜索算法的时间复杂度。Ziegler等在语义网络上使用本地组信任管理的方式，结合用户信任与不信任关系获取社交网络的信任关系。Tang 等分析了在线网络上的信任演化过程，并提出eTrust模型。Deng等提出了一种分为两个阶段的推荐流程，在初始化过程中利用深度学习来确定矩阵分解中的初始化社交信任度，将用户的兴趣及其可信赖的朋友的兴趣与社交效应的影响一起用于推荐。Yang等提出了集成用户提供的稀疏评级数据和社交信任 网络来提高协作过滤性能的方法。采用矩阵分解技术，根据用户的信任关系将用户映射到低维潜在特征空间，旨在更准确地反映用户对自己意见形成的相互影响，并更好地学习以为用户提供优质模式的高质量建议。Xiong等提出了一种新的基于信任的Top-k的推荐算法BTRank，其集成了 评分和信任信息，构建了评分排序模型，有效地提高了所有用户的Top-k项目列表的质量。Guo等提出了一种基于信任的在 线社交网络隐私保护推荐方案，其中在线社交网络用户通过多跳信任链，应用其属性来寻找匹配朋友，并与陌生人建立社交关系。

通过用户社交网络的信任传播能够有效解决数据稀疏和冷启动问题，Wang等人提出了一种基于信任和偏好的协同过滤的算法——TPCF（Collaborative Filtering Algorithm Based on Trust and Preference）算法。TPCF算法的基本思想是：在局部网络中使用传播激活模型建立信任网络，根据各用户间的信任关系获得用户间的信任值；提出偏好相似度、产品体验相似度及兴趣相似度的概念，设计偏好相似度计算指标并获得 用户偏好相似度；动态结合信任度和偏好相似度权重得到用户推荐权值，在已知用户评分及推荐权值的情况下，使用最近邻协同过滤算法计算出预测评分，并给出相应的推荐列表。

### 信任评分场景

推荐场景可以定义为社交网络下的服务推荐 SOE-SN（Service-Oriented Environments with Social Network）。在 SOE-SN中存在用户间社交网络G以及用户对服务的评分网络Rating两部分。定义G（V，E，W）表示社交网络，其中V 表示网络中用户组成的集合，E 表示用户间信任关系的集合，W 表示信任边上的权值。评价网络Rating(V’,E,W)中，V’表示所有参与评价服务的用户集合，且；S 表示在服务推荐环境下所有可能获得的不同服务集合；R 表示用户对服务的评价值集合。

信任模型如图2-6所示。用户间的有向边表示信任关系，边上的权值 w(a ,b)表示用户a与用户b 间信任程度的强弱，是一个介于0到1之间的实数。用户与服务间的有向边表示用户对服务的评价，边上的权值代表用户对服务的评价分数，本文定义用户的评价分数为 R(a ,b)。其中 R(a ,b) ∈{1，2，3，4，5 }。

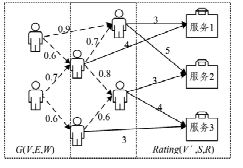


图2-6 信任评估场景

### 信任特性

在信任传播和信任聚合的过程中，需综合考虑信任的以下特性。

1. 信任关系的传递性。

如图2-7所示，用户a直接信任用户b，用户b直接信任用 户c，于是a间接信任用户ｃ，即。用户a对用户c的信任程度可以根据网络的连接情况和传播激活模型获得。



图2-7 信任传递示例

1. 信任关系的不对称性。

网络中有很大概率出现 ，整个用户信任网络是加权有向网络。

1. 信任关系的衰减性。

随着链路增长及信任程度的逐渐衰减，无直接联系的用户间间隔增多，两者间的信任程度降低。

1. 信任关系的多路性。在实际信任网络中，两个节点间的信任传递路径通常不止一条。关系链路越多，两个用户间的信任度就越大。

### 偏好相似度

此算法中偏好相似度被定义为用户间评分服务类型和评分值的综合相似程度。用户会对感兴趣的服务进行评价，而评分结果高也代表用户对此类项目的偏好程度大。因此，可 通过用户对服务的评价情况获得用户之间的偏好相似度。本文使用sim(a ,b) 表示用户a和用户b之间偏好的相似度，其值是介于0到1之间的实数。

### 算法分析

基于信任关系与用户偏好的服务推荐算法TPCF的框架如图2-8所示。给定评分矩阵和信任矩阵，算法的具体步骤如下：

1. 由评分矩阵得到任意两个用户之间偏好的相似度；
2. 由用户信任关系矩阵构建信任网络，建立信任传播模型，获得信任度矩阵；
3. 动态结合信任度和偏好相似度，得到推荐权值；
4. 采用协同过滤推荐，得到最终评分预测并获取推荐列表。

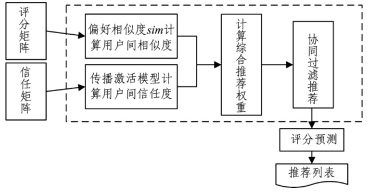


图2-8 TPCF框架

使用传播激活模型计算信任网络的传递过程和信任值。在传播激活模型中，如图2-9所示，源节点a注入的总信任能量为e，能量沿着有向边传播到后续节点。能量可以按照边上的权重进行分流，权重大的边所传播能量的比例大。整个信任传播过程中网络的能量保持守恒。

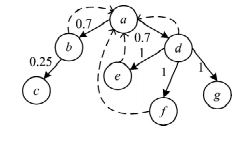


图2-9 能量的传播过程

每个终叶节点都有一个回溯到源点a的边且权值为１， 因此能够保证能量在整个网络中守恒且不进入死循环。

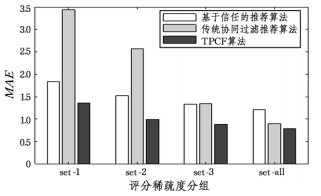
在推荐系统中，用户更倾向于选择与自身偏好相似的用户所选择的服务。此算定义两用户评分数据的相似程度为偏好相似度。不妨从两方面考虑用户的偏好相似度，一方面考虑用户共同评价过服务的数量，另一方面考虑两用户间服务评分的差异。

加入用户间的社交关系，并把社交关系量化成信任度，这使得在获取较少历史数据的情景下依靠用户间的信任关系也能获得较为准确的推荐结果。然而，如果只考虑用户间的社 交关系而忽略用户间的偏好相似度，则会导致在社交关系稀疏的情况下推荐结果偏差较大。于是将偏好相似度和信任度进行结合，有助于得到更加准确且合理的推荐结果。

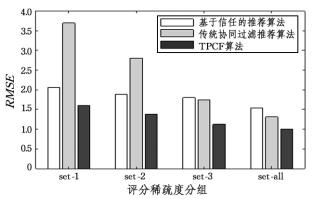
### 种推荐算法对比

实验数据集：王佳蕾等人采用包含22166个用 户、296277个产品、922267条用户评分数据的Epinions数据集1）进行实验，数据集中存在一半冷启动用户（评分次数少于５次），该数据集是一个非常稀疏的数据集。

TPCF算法是通过动态结合信任度和偏好相似度两部分指标得到推荐权重，并进行协同推荐的算法。图2-10对比了TPCF算法与单纯的信任度协同推荐和传统协同过滤算法推 荐效果。其中，set-1 表示采样评分数据数量小于3的源点集合，set-2是采样评分数量位 于3到5之间的源点集合，set-3是采样评分数据数量位于5到10之间的源点集合，而set-all 表示整体Epinions集合中随机采样源点集合。从图2-10中可以看到，相比于其他两种方法，TPCF方法 的 MAE（平均绝对误差 ）值及 RMSE（均方根误差）值较小，在预测的准确率上优于其他两种算法。



1. MAE 的对比



1. RMSE 的对比

图2-10 三种推荐算法性能的对比

在评分数据稀疏时，计算用户间的偏好相似度的误差较大，因而经典CF算法的推荐效果不佳。但基于信任权重的推荐的准确性取决于信任数据是否稀疏，与评分数据的数量 无关。TPCF算法由于动态地结合了经典CF算法以及基于信任权重的推荐，受到评分数据稀疏的影响较小，因此在评分数据稀疏时依然能够获取较高的推荐准确率。

TPCF综合了信任与偏好关系，使得用户间存在更多的可衡量关系，结合了两种推荐算法的优点，扩充了推荐算法的覆盖率。如图2-11所示，采样了编号0-5000的全体用户，并将其平均分成5组，每一组为连续的1000个用户及其评价值。计算其各自不同的平均覆盖率。由图2-11可知，单纯基于信任的推荐方法由于数据集中信任数据稀疏的问题而平均覆盖率较低；传统协同过滤算法在用户评分数据 稀疏时的平均覆盖率也较低；TPCF算法使用了社交网络数 据，对只根据用户评分计算整体信任权重的协同过滤方法进行了补充。尽管单一使用社交网络数据形成信任度并进行推 荐算法的平均覆盖率并不高，但结合两者优势的TPCF算法在整体覆盖率上有较大提升。

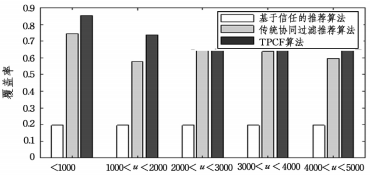


图2-11 3种推荐算法的覆盖率比较

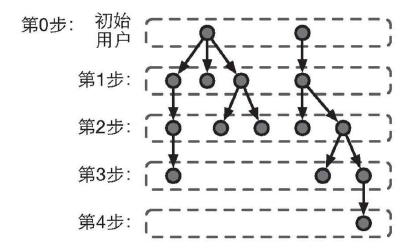
## 信息传播模型

社交网络中的信息传播近年来吸引了大量研究者的关注，主要分为传播话题挖掘、传播过程建模、传播最大化三个主流课题。本文的研究工作主要集中在传播过程建模，包括微观用户传播行为预测模型，与宏观传播规模预测模型。同时本文也涉及一些传播最大化问题中的关键用户识别算法。

### 微观个体行为模型

给定一条消息以及一个社交网络，网络中包含用户及用户间的关系，微观个体行为模型旨在预测哪些用户、在何时会帮助传播给定的消息，亦即描述该消息在社交网络中的传播过程。

该类方法包括两个基本模型：独立级联模型（Independent Cascade，IC）和线性阈值模型（Linear Threshold，LT）。两者均假设每位用户处于两种状态之一：还未接收到消息的用户处于非激活状态（inactive），而已经帮助传播消息的用户则处于激活状态（active）。IC模型和LT模型进一步假设处于激活状态的用户数随着时间的增长而单调递增，也就是已激活的用户无法回到非激活状态。同时，一位非激活用户被激活的概率与周围已激活的好友数成正比。亦即，周围尝试将消息推送给一位用户的好友越多，该用户被激活的概率越高。



IC信息传播模型

IC模型中，网络中的每条边上存在一个权重，用于表征一位用户将消息成功推送给另一位用户的概率。而在LT模型中，除了边上的权重外，每个结点还存在一个对应的阈值。

具体来说，在LT模型中，每个结点v的阈值往往从某一个分布中采样得到。而用户v与其每一位邻居u间的边权则被定义为往。在任意时间点，对于处于非激活状态的用户v，倘若其周围已被激活的好友的边权和达到或超过了，用户v就会在下一个时间点变成激活状态。

而在IC模型中，每当一位用户v在时间t被激活，他便有一次机会将消息推送给其未激活的好友u，其成功的概率为。而一旦用户v推送成功，则用户u会在时间t+1处于被激活状态。上图展示了IC模型中信息的传播过程。这两个不同的过程反应出，IC是以消息发送者为中心的模型，而LT则是以消息接收者为中心的模型。

### 宏观传播效应与模型

一条信息的传播规模一般被定义为在传播过程中，接收并帮助传播该消息的用户数，传播规模一般也被称为一条信息的受众度（popularity）。而信息传播规模预测模型的研宄目标，便是信息受众度随时间推移的动态演化过程。这类模型的特点是，并不要求输入一个网络结构，也不会描述用户的转发行为决策，而是从宏观层面直接对传播规模（受众度）进行建模。

对传播规模建模的方法分两种，无需提取特征的方法，和基于特征提取的方法。其中前者又可以分为传染病模型和基于Hawkes过程的模型。下面对这些建模方法进行详细描述。

#### 传染病模型

以SI、SIR、SIS为代表的传染病模型将用户分为不同的阶段或状态。例如，这里S代表“易感染”（Susceptible），亦即尚未接受消息的用户；I代表“感染”（Infected），亦即已经接收并帮助传播消息的用户；而R则代表“康复”（Recovered），可以理解为不再会接收消息的用户。在上述三种模型中，处于易感染期的用户以一个固定概率进入感染期。在SIR模型中，感染用户则以另一个概率进入康复期，亦即接收消息的用户不再传播同样的消息。而在SIS模型中，感染用户则会以一定概率回到易感染状态，亦即用户会重复传播消息。以最基础的SI（Susceptible-Infected）模型为例，该模型假设被感染的用户会尝试以的概率将信息（疾病）推送（传染）给其他用户。因此，若网络中共包含N位用户，且用I(t)表征时间t的感染者数量，S(t)用表示时间t的易感染者数量，那么I(t)和S(t)的动态演化过程可以用如下公式表述：

上述公式描述了在时间t感染者与其他用户之间会发生共次接触，而每次接触以概率令一位处于易感染期的用户被感染（接收消息）。对于SI的扩展模型，则有除上述公式外的更多动态方程用于其它状态之间的用户转移量。其它常见的传染病模型还包括MSIR、SEIR等。

#### Hawkes过程

而另一类传播规模预测模型则基于Hawkes过程。在这类模型中，传播起始于一个外部信息源E(t)或一些早期的传播者，这些用户以乘以一个常数项c为权重对信息进行传播，而则随着时间t的递增而逐渐衰减。如下式所示：

其中表示用户i接受消息的时间。由该类模型所生成的传播规模，其演化过程往往会伴有一个长尾下降模式，这种现象与很多实际数据所吻合。例如，Crane等人提出基于Self-Excited Hawkes Conditional Poisson Process的模型，对YouTube上的视频点击率进行建模。而Zhao等人和Yu等人则用类似的模型分别对Twitter上和微博上的帖子转发量进行预测。该类模型的另一个特点是假设网络中有无穷多的用户，因此其计算所得的信息传播规模可能会趋于无穷大，这是此类模型的一个缺陷。

#### 基于特征的方法

与上述模型不同，该类方法往往需要首先提取一系列详细、与传播规模潜在关联的特征，例如消息的内容特征，用户网络结构特征，时间特征，原帖的特征等。然后将这些特征应用于一些学习算法中，对信息传播规模进行预测。常用的学习算法包括简单的回归模型，基于概率的协同过滤模型，回归树，基于内容的模型，以及被动主动感知器。然而这些方法存在一些问题，例如在海量数据中提取特征所带来的时间成本和计算成本，定义特征所需的人力成本，并且模型的预测结果高度依赖于特征的质量。同时，在实时的社交数据中，每秒都有大量帖子被发布或转发，因此对每个帖子提取特征显得并不现实。为了解决计算复杂度问题，Choudhury等人研宄了如何利用采样算法对传播规模进行估计，他们所提出的算法既考虑了网络的拓扑结构，也考虑了用户的个人特征。此外，Sadikov等人提出了一种基于K-树的模型，通过部分可观察的传播过程，对传播规模进行估计。

#### 宏观长尾效应

信息的传播规模往往服从一个长尾分布，这种现象也被称为信息传播过程中的长尾效应。常见的长尾分布包括幂律分布和对数正态分布。

幂律分布的概率密度函数如下式所示：

其中，是一个幂次参数，而C则是一个归一化因子。本质上，幂律分布刻画的是两个变量间的函数关系，其中一个变量的变化量是另一个的指数级倍数。例如在万维网上，网站的超链接数，就随着网站数量的增长而呈指数级增长。因此，以网站与网站间的超链接关系为结点和边，建立一个网络后，该图的度数便服从一个幂律分布，类似的现象还能在社交网络中边的权重的统计中发现。

除幂律分布外，对数正态分布则经常在动态网络中被观察到。从定义上来讲，服从对数正态分布的变量，其取对数后的值服从正态分布。该分布最早由Mood所定义，其概率密度函数如下所示：

其中，和是随机变量x取对数后所服从的正态分布的均值和方差参数。

### PowerInfect动态演化模型

PowerInfect模型是结合传播规模与用户状态的动态演化模型，由于该模型可以在理论

上推导出服从幕律分布（Power-law distribution）的传播规模结果，因此将之命名为PowerInfect模型。该模型有两个主要组成部分：传播序列的生成与转移，及用户状态的生成与转移。

#### 受众度演化

传播序列的生成过程包括两步：1）选择被转发的帖子；2）选择转发帖子的用户。接下来对这两步进行详细的描述。

**生成传播信息。**为生成发生在单位时间t的传播事件，模型首先决定D个帖子中，哪一个涉及该次传播。定义受众度r的值为的帖子被选择的概率为：

其中，函数用于描述帖子d被选中的概率，与其当前的受众度之间的关联关系。例如，对任意r，如若定义，则所有帖子将会被等概率选择。而如果定义，其中常数C>>r，则表征受众度越大的帖子越容易被转发。当的值与r成正比时，不同帖子的受众度将服从一个幂律分布。

**生成传播者。**在信息传播过程中，一位用户是否进行转发的决策会被其周围的好友所影响。而不同的用户会受到不同程度的影响，从而决定转发与否的概率也不尽相同。PowerInfect模型定义了K种“用户状态”，并在每个时间点为每一位用户分配一种状态。模型进一步假设，处于相同状态的用户，其决定是否转发一条消息的概率相同。例如，在实际应用中，可以根据一位用户v周围尝试将消息推送给自己的好友数定义v的状态。在下图将用户划分为两种状态：周围有少于三位好友推送消息的用户被划分为状态1，而其他用户则被分为状态2。因此，在图中，用户由于只有一位已推送消息的好友，因此被划分为状态1，而用户和用户则被划分为状态2。具体每一位用户的状态，及其周围推送消息的好友数，均是未知的。可以理解为，PowerInfect模型通过训练数据尝试学习对用户状态的定义方式。

PowerInfect模型定义（有且）为处于状态k的用户所占比例，并且定义转发率参数（）用于表征状态为k的用户转发一个帖子的概率。基于上述定义，可以得到一位用户属于状态k并且转发一条帖子的概率为。

**生成过程整合。**定义为在时间t受众度为r的帖子数。如前文所述，PowerInfect假设在每个单位时间片内只会发生一起转发事件，因此至多增加或减少1。具体来说，对于r>0，若增加一个帖子，当且仅当有一个受众度为r-1的帖子被转发。其概率如下所示：

相似地，当且仅当受众度为r的帖子被转发时，会减少一个帖子，其概率

为：

接下来，为了描述的动态演化过程，需要定义其在单位时间内的变化量。

而综上所述，对于，可以得到在单位时间的变化量的期望值如下式所示：

对于，可以得到在单位时间的变化量的期望值如下式所示：

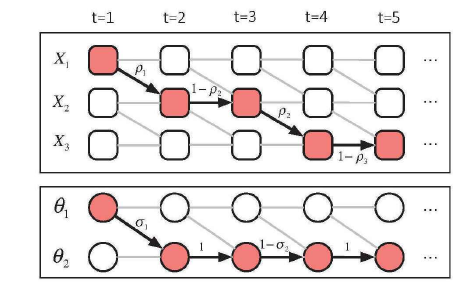
#### 用户状态演化

在上述小节中，模型将用户状态均作为恒定值看待。然而现实应用中，用户的状态往往是会发生改变的。例如，若模型学习得到，周围没有已传递消息的朋友的用户处于状态1，而其他用户则处于状态2。在时间t用户转发了消息，这会导致的周围，原先没有转发消息的好友的用户，从状态1跳转至状态2。用户状态由1至K进行编号，编号大的用户状态表征周围更多的好友推送了消息。由于在信息传播过程中，周围推送消息的好友随着时间的流逝在单调递增，因此用户仅可能从较小编号的状态跳转至较大编号的状态。定义当一位用户状态为K的用户转发一条帖子后，其状态变为K+1的概率为。因此，增加的概率由当前被选中的用户成功转发帖子的概率和状态发生改变的先验共同决定。具体地，对k>1时，增大的概率为：

类似地，减小的概率由被选中用户转发失败的概率、及用户状态未发生的概率决定，具体如下式所示：

简单来说，对于任意时间点，当一位用户转发一个帖子时，有一定概率变化，而没有转发时，—定不会变化。

下图展示了帖子受众度与用户状态共同演化的一个案例。在图中，当t=1时，模型选择了一个受众度为1的帖子及一位状态为1的用户。该用户以的概率成功转发了帖子，而这次转发促使一位用户以的概率由状态1转移到了状态2。此后，当t=2时，模型选择了一个受众度为2的帖子和处于状态2的用户。这次，转发以的概率失败了，因此根本没有发生任何改变。当t=3时，转发发生了，然而仍以的概率保持不变。



PowerInfect模型示意图

当K>１时，的演化过程可以由下述动态方程描述：

当K=1时，仅可能减小，因为在倍息传播过程中，周围推送消息的好友随着时间的流逝单调递增，因此没有处于更小编号状态的用户可以跳转至状态１。的的演化过程为：

完整的PowerInfect模型包含了对帖子受众度及用户状态的动态演化过程的描述。具体来说，对于r>0且K>1，PowerInfect模型包含了下述动态方程：

# 应用案例

随着互联网时代的到来，各种社交媒体盛行，社交网络分析应用也越来越广泛。本文主要从用户个体和社交群体两方面来阐述社交网络分析的应用案例。

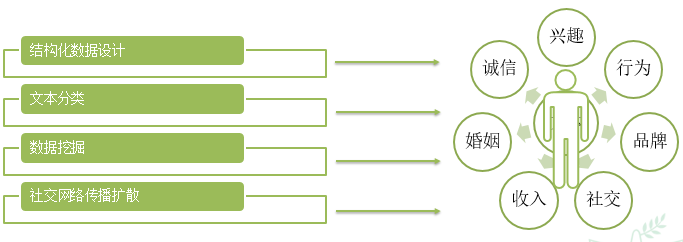
## 构建用户画像

### 用户画像

用户画像（persona），最早由交互设计之父Alan Cooper提出：“Personas are a concrete representation of target users.”，是指真实用户的虚拟代表，是建立在一系列属性数据之上的目标用户模型。随着互联网的发展，用户画像又包涵了新的内涵——根据用户人口学特征、网络浏览内容、网络社交活动和消费行为等信息而抽象出的一个标签化的用户模型。

用户画像的作用分为以下几方面：

1. 精准营销：精准直邮、短信、App消息推送、个性化广告等。
2. 用户研究：指导产品优化，甚至做到产品私人定制等。
3. 个性化服务：个性化推荐、个性化搜索等。
4. 业务决策：排名统计、地域分析、行业趋势、竞品分析等。



### 用户画像的生产

用户特征的提取即用户画像的生产过程，大致可以分为以下几步：

1. 用户建模，指确定提取的用户特征维度，和需要使用到的数据源。
2. 数据收集，通过数据收集工具，如Flume或自己写的脚本程序，把需要使用的数据统一存放到Hadoop集群。
3. 数据清理，数据清理的过程通常位于Hadoop集群，也有可能与数据收集同时进行，这一步的主要工作，是把收集到各种来源、杂乱无章的数据进行字段提取，得到关注的目标特征。
4. 模型训练，有些特征可能无法直接从数据清理得到，比如用户感兴趣的内容或用户的消费水平，那么可以通过收集到的已知特征进行学习和预测。
5. 属性预测，利用训练得到的模型和用户的已知特征，预测用户的未知特征。
6. 数据合并，把用户通过各种数据源提取的特征进行合并，并给出一定的可信度。
7. 数据分发，对于合并后的结果数据，分发到精准营销、个性化推荐、CRM等各个平台，提供数据支持。

## 社交媒体

### 推荐系统

一个好的推荐系统，对于电子商务网站来说：1.提高用户体验，好的个性化推荐，更好的向顾客展示商品以提高转换率。2.发现顾客的潜在需求，提高客单价。电子商务的先驱——亚马逊号称推荐系统提高了其30%的销售。

亚马逊推荐系统使用了至少十几种推荐算法。推荐形式主要包括三种：（1）基于用户的推荐；（2）基于内容的推荐；（3）基于图的推荐。

QQ、微信、人人网的好友推荐也是一种基于社交网络分析的推荐算法应用。在网络社交平台中有许多人都使用虚拟身份来进行交友，某些人因为虚拟身份的伪装，认为自己不用为在网络上的行为负责，在社交平台上肆无忌惮损害其他用户的权益，为社交网络的安全性能带来隐患。因此，好的好友推荐算法至关重要，基于社交网络分析的信任推荐算法很好了降低了该风险。

### 网络营销

社交网络具有**开放性（**如：任何人都可以加入其中并与外部世界产生联系）、**互动性**（如：通过回复和转发，用户可与朋友就一部电影或一个事件进行互动）和**时效性**（如：用户可随时更新状态信息）。利用社交网络信息传播，可以达到快速的病毒式营销（viral marketing）的目的。

如何基于网络结构评价用户的权威性和影响力？如何实现影响力最大化？根据个体影响力分析，选取圈子中影响力较大的人进行推广，可以快速扩展产品的广度和知名度，为针对用户群所熟知。如：微博流量明星的营销推广，一般有百万级甚至千万级的点赞、转发和分享，而这些行为必定有部分转化为产品销售额。同时，社交网络分析可以帮助商家针对流失群体进行分析，以此为基础调整策略等。

### 特定群体发现

社交网络分析还可以用于特定群体发现。

1. 疾病传播：

由一个中心点向外扩散，切断网络中的关键节点就可以有效阻止传染病的传播；

1. 抑郁症群体：

Facebook通过人工智能技术救助有自杀倾向的用户。它可以检测出含有自杀倾向的可疑帖子，而后Facebook的社区运营团队会对这些可以帖子进行审查、核实等。该技术主要依赖用户发现好友的可疑内容，对含有自杀倾向的可疑帖子、评论或视频直播内容等进行标识，用户的好友们可以很容易在页面找到这些被突出标识的内容。

## 公共安全领域

社交网络分析技术为公安执法部门的智慧警务建设提供了一种有利手段。在实际的警务工作中，公安干警对案件的研判会利用犯罪嫌疑人既有的复杂社会关系，对犯罪嫌疑人的可能动向、嫌疑性大小进行分析。与社交网络分析相同，整个研判过程可以使用网络结构进行建模。网络中的一个节点可以表示一个人、一个公司、一个组织，而边则代表着二者之间不同的信息交互类型，例如买卖双方的毒品交易、恐怖分子之间的通信联系、歹徒与受害者之间的联系等。由此构建的社交网络图，首先通过可视化技术，既可在视觉上直观地体现出犯罪嫌疑人之间、犯罪嫌疑人与受害人之间、犯罪嫌疑人与其亲人朋友之间的联系，又将抽象的案件分析具体化、数据化；然后，利用社交网络分析技术的定量分析，能够刻画出案件中的关键人物、团伙间的协作流程，为警员的案件侦破或者复盘提供丰富多角度的信息。

### 犯罪团伙挖掘

随着社会的发展，犯罪手段和技术也在不断地演进，呈现出团伙作案的趋势。在一个犯罪团伙内部，成员们分工明确，通过相互协作实现隐蔽作案的目的。在办案过程中仅仅打掉其中一个环节或某几个环节，效果都不尽如人意，团伙仍有可能重新组织人手继续犯罪。因此，明确团伙成员、分析清楚团伙内部成员的地位等级，将给执法办案提供极大的帮助，同时也是需要解决的一大难题。

社交网络分析为上述这一难题的解决带来了希望。社区发现是社交网络分析的一项基本内容。而基于以往的经验了解到，犯罪团伙间都存在或多或少的联系，团伙内部的关系与团伙外部的关系有着显著的差异，比如，团伙的结构相对稀疏，但外围的人员与外界关系频繁等。根据上述经验，设置合理的模块优化目标，可以利用社区发现的思想得到一个粗略的群体，然后再利用精细的分类方法，比如基于图核函数的聚类算法，从群体中分离出小团体。

通过社交网络分析技术，能有效地从海量数据中筛选出可疑度较高的团伙，从而降低了警员挖掘潜在团伙的人力成本。

### 犯罪团伙中的核心成员识别

在分析出犯罪团伙以后，需要进一步识别团伙中各个成员的角色和地位。比如处于聚簇边缘、与外界联系相对较紧密（与其他聚簇相连）的节点，可能是联络人，也可能是新入伙的人；在聚簇中心的可能是组织中的核心人物。所谓“擒贼先擒王”，首先识别出网络中的核心人员，警方再根据分析结果，选择特定的突破口，对该组织实施打击粉碎行动。这里利用节点重要指标可以迅速定位挖掘犯罪团伙中的核心成员。

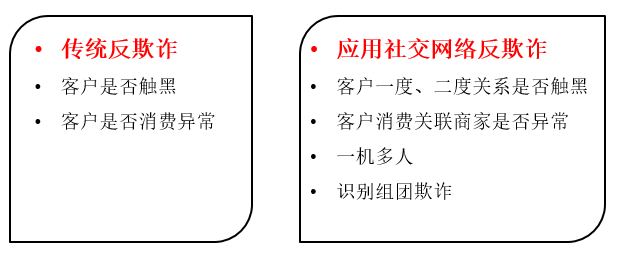
根据实际网络情况，选择合理的指标对网络中的重要节点进行识别，如：介中心性、紧密中心性等等。比如，在一个涉恐网络中，领导人物的命令要以最快的速度传播到整个网络中，那么紧密中心性将很好地识别出这类信息发布角色的重要节点。以美国“9·11”恐怖袭击的犯罪网络]为例，使用新闻报道资料为基础构建了“9·11”袭击中的19个恐怖分子构成的关系网络。然后利用紧密中心性和中介中心性等中心性指标识别出了网络中的关键节点——Mohamed Atta，而事后证实确认该关键节点即该团伙的领导者。此例子给社交网络分析技术在公共安全中的应用提供了强有力的证据。

### 串并案挖掘

串并案是公安机关提高办案效率的一个有效途径，它是指把多个可能相关的案件一起进行研判分析。这里的相关包括同一伙犯罪成员或者是相似的作案手段等。通过对多个案件的串并分析，不仅能起到信息互补的作用，而且能解决不少冤假错案。社交网络分析有利于将同一犯罪团伙所为的案件进行串并，使警员有效洞察其背后千丝万缕的联系，为案件侦破提供线索。

以美国某州的警局针对该地区一系列便利店的抢劫案为例，警方利用社交网络分析进行并案，从而抓捕了多个之前未被发现的犯罪嫌疑人。他们首先把单独的案件中涉及的人、事、物映射到一张网络上，然后在调查中发现某个正在被调查的人A与其他辖区的有案底的人B存在联系。于是，他们以这两人（A和B）的姓名为种子，利用类似标签传播的算法，构造了一张未知网络，即网络中出现了多个未被关注的人员。对这个未知网络进一步分析，发现该有案底人员A又与多个青少年有频繁的联系，于是就把之前未被人视为同一伙人作案的抢劫案给串联起来了。最后，通过可视化展现的方式，对多张网络进行浏览和管理，为干警破案提供了快速有效的线索。

### 金融反欺诈



传统反欺诈和应用社交网络反欺诈进行对比，我们可以发现，应用社交网络进行反欺诈具有以下几个优势：

1）可以更大概率检测出客户是否触黑；

2）可以检测出组团欺诈；

3）可以检测出客户付款方是否存在欺诈行为，降低客户被欺诈率。

参考文献

[1] Newman M E J. Networks: An Introduction. Oxford: Oxford University Press, 2010

[2] Albert R, Jeong H, Barabási A L. Error and attack tolerance of complex networks. Nature, 2000, 406: 378–382

[3] Cohen R, Erez K, Ben-Avraham D, et al. Breakdown of the Internet under intentional attack. Phys Rev Lett, 2001, 86: 3682

[4] Weng J, Lim E P, Jiang J, et al. Twitterrank: Finding topic-sensitive influential twitterers. Proceedings of the Third ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2010. 261–270

[5] Ding F, Liu Y, Cheng H, et al. Read and Reply Behaviors in a BBS Social Network[C]// the 2nd ieee international conference on advanced computer control. IEEE. Shenyang China 2010:571-576.

[6] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine[C]// International Conference on World Wide Web. Elsevier Science Publishers B. V. Brisbane Australia 1998:107-117.

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

[8] Freeman L C. A set of measures of centrality based on betweenness.[J]. Sociometry, 1977,40(1):35-41.

[9] Burt R S, Minor M J, Alba R D. Applied network analysis: A methodological introduction. Sage Publications Beverly Hills, 1983

[10] Kitsak M, Gallos L K, Havlin S, et al. Identification of influential spreaders in complex networks. Nat Phys, 2010, 6: 888–893

[11] Kim S J, Lee S H. An improved computation of the pagerank algorithm. Adv Infor Retr, 2002, 2291: 73–85

[12] Brin S, Page L. Reprint of: The anatomy of a large-scale hypertextual web search engine. Comput Netw, 2012, 56: 3825–3833

[13] Bollen J, Rodriquez M A, Van de Sompel H. Journal status. Scientometrics, 2006, 69: 669–687

[14] Bhat H S, Sims B. InvestorRank and an inverse problem for PageRank. Doctor Dissertation. Merced: University of California, 2012

[15] Chen P, Xie H, Maslov S, et al. Finding scientific gems with Google’s PageRank algorithm. J Informetr, 2007, 1: 8–15

[16] Walker D, Xie H, Yan K K, et al. Ranking scientific publications using a model of network traffic. J Stat Mech Theor Exp, 2007, 6: P06010

[17] Ma N, Guan J, Zhao Y. Bringing PageRank to the citation analysis. Inform Process Manag, 2008, 44: 800–810

[18] Jomsri P, Sanguansintukul S, Choochaiwattana W. CiteRank: Combination similarity and static ranking with research paper searching. Int J Int Technol Secur Tran, 2011, 3: 161–177

[19] Zeng A, Zhang C J. Ranking spreaders by decomposing complex networks. Phys Lett A, 2013, 377: 1031–1035

[20] Hu Q, Gao Y, Ma P, et al. A new approach to identify influential spreaders in complex networks. In: Web-Age Information Management 2013. New York: Springer, 2013. 99–104

[21] 汪小帆,李翔,陈关荣.网络科学导论[M].高等教育出版社,北京,2012,159-162

[22] ZHANGS, ZHOU Y. Template matching using grey wolf optimizer with lateral inhibition [J].Optik-International Journal for Light and Rlectron Optics,2017,130:1229-1243

[23] NIZAMIIF, MAJIDM, AFZALH, etal.Impactof Feture Se-lection Algorithmson Blind Image Quality Assessment [J]. Ara-bian Journal for Science & Engineering, 2017,43(4):1-14

[24] 邵浩,陈东方,刘欣.复杂网络算法中K-shell与介数中心性算法的实现[J].现代计算机(专业版),2014(17):7-11.

[25] 王佳蕾,郭耀,刘志宏.基于社交网络信任关系的服务推荐方法[J].计算机科学,2018,45(S2):402-408.

[26] 顾亦然,马德营,孟繁荣.社交网络中个体对群体影响力分析[J/OL].计算机技术与发展,2018(12):1-6[2018-12-26]

[27] 任晓龙,吕琳媛.网络重要节点排序方法综述[J].科学通报,2014,59(13):1175-1197.

[28] 杨洋，社交网络中的信息传播模型研究，2016，6

[29] 方滨兴, 许进, 李建华. 在线社交网络分析[M]. 电子工业出版社, 2014.

[30] Severyn, Aliaksei, Alessandro Moschitti. "Twitter sentiment analysis with deep convolutional neural networks." Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2015

[31] Wang, Xin, et al. "Social Recommendation with Strong and Weak Ties." Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. ACM, 2016

[32] Althoff, Tim, P. Jindal, and J. Leskovec. "Online Actions with Offline Impact: How Online Social Networks Influence Online and Offline User Behavior." Tenth ACM International Conference on Web Search and Data Mining ACM, 2016:537-546

[33] Zhang, Jiawei, et al. "Intertwined viral marketing in social networks." Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on. IEEE, 2016

[34] Forsati R, Barjasteh I, Masrour F, et al. Pushtrust: An efficient recommendation algorithm by leveraging trust and distrust relations[C]//Proceedings of the 9th ACM Conference on Recommender Systems. ACM, 2015: 51-58.

[35] Wang Y, Yin G, Cai Z, et al. A trust-based probabilistic recommendation model for social networks[J]. Journal of Network and Computer Applications, 2015, 55: 59-67.

[36] Zhang Z, Liu H. Social recommendation model combining trust propagation and sequential behaviors[J]. Applied Intelligence, 2015, 43(3): 695-706.