浙江大学

本科生毕业设计
开题报告

学生姓名：夏立伟

学生学号：3120101964

指导教师：陈为、巫英才

年级与专业：大四 计算机科学与技术

所在学院：计算机学院

一、题目：对 MMORPG 中多种行为的可视化分析系统

二、指导教师对开题报告、外文翻译和中期报告的具体要求：

开题报告能够对所做的项目有一个明确的介绍，包括背景介绍，技术要求，设计目标等。同时对项目有一个明确的时间进度安排。对所做的目标有一个合理的可行性的分析，并且能够分析完成项目的一些关键技术要点和技术方案

外文翻译翻译一篇与社交网络相关的英文论文。

中期能够有一个初步的系统框架以及能够完成数据的整理以及数学建模的工作。

指导教师（签名）_____

年　　月　　日

毕业设计开题报告、外文翻译和中期报告考核

导师对开题报告、外文翻译和中期报告评语及成绩评定：

成绩比例	开题报告 占（20%）	外文翻译 占（10%）	中期报告 占（10%）
分 值			

导师签名_____

年　　月　　日

答辩小组对开题报告、外文翻译和中期报告评语及成绩评定：

成绩比例	开题报告 占（20%）	外文翻译 占（10%）	中期报告 占（10%）
分 值			

开题报告答辩小组负责人（签名）_____

年　　月　　日

-

目录

本科毕业设计开题报告

1. 游戏数据可视化背景与简介

大型多人在线角色扮演游戏(MMORPG)建立了一个虚拟社会，许多的玩家可以通过扮演不同的角色并且加入游戏中的不同的门派和工会进行玩家与玩家之间的行为交互。由于 MMORPG 中交互性和用户自参与的特点，成千上万的玩家们在不同的 MMORPG 中投入了金钱精力以及社会资源，而这些投入都成为了当今社会中电子商务中的重要的组成部分。据估计，2015 年全世界范围内 MMORPG 的收入达到了 1100 万美元，并且在 2017 年有望达到 1300 万美元。

所以如果能够理解玩家在游戏中的行为模式和行为动态将会带来巨大的经济价值。除此之外，玩家在 MMORPG 中的个人行为的时间戳节点信息能够很好的帮助社会科学家来估计玩家的多种行为是如何被其他玩家影响的，并且探究这些影响是如何限制或者促进 MMORPG 游戏所创造的虚拟世界的。

之前对于 MMORPG 中玩家的行为模式的研究收到了一些大的限制。首先，大多数的研究知识对 MMORPG 游戏中个体的成长行为的一些片断进行了研究，这些研究在很大程度上限制了我们对玩家行为的连续动态行为的研究。其次，之前的研究将玩家脱离于 MMORPG 这个虚拟社会的上下文的情景中，并且将玩家的行为当作是个人决定的结果。这样的假设严重的限制了我们对个体，核心玩家以及其他玩家动态行为的社会结构所扮演的角色的理解。除此之外，大多数的研究知识将关注点放在 MMORPG 中的单种行为并且忽视了玩家的行为的多样性，这样的做法让社会学家很难找到多种行为之间的动态联系。

为了研究 MMORPG 玩家多种行为之间的动态关联并且调查不同社会结构所扮演的角色，我们必须面对两个主要的挑战。一个挑战就是如何如何测量玩家不同行为之间的相互作用并且对基于这些相互作用的不同的社会结构进行建模。在 MMORPG 游戏中每一个玩家都是从属于这个虚拟社会的多种社会结构中的，从最简单最基础的二元关系结构到三角关系结构最后到全世界级的结构。除此之外，这些社会结构都是互相渗透的。如何选取理论上重要的社会结构和从经验上获取他们所扮演的社会作用是我们所要面对的挑战。

第二个挑战是如何去可视化多种行为之间的动态关联并且可视化社会结构对动态行为的作用。动态相互作用包括了不同的商品以及不同组的核心玩家。尤其是这种相

互影响会随着时间动态的变化类型，包括不同群体的玩家之间的影响以及不同组的玩家对商品消费的影响。因此，很难去建立一个简洁的，内容充足的并且具有高辨识度的可视化系统对这种影响力进行一个时间上的总结。除了对这些关系提供可视化之外，对不同组玩家之间对不同商品购买的影响力的异同点的解释也是一个挑战。同时由于玩家在游戏中的行为活动和他们自身的多种动态属性是在不同层面进行可视化的，所以如何可视化来进行这些探索和分析仍然是个难题。到现在为止，没有什么可视化方面的研究可以帮助分析动态行为和他们随时间变化的多种关系联系。

为了解决第一个挑战，我们系统的探索了那些会影响 MMORPG 游戏里玩家行为的社会结构并且找出了两个比较基础并且重要的结构化的机制:社会影响力和三元闭包。我们在这基础上提出了一种对影响力敏感的模型用来调查研究上述社会结构对于玩家行为互相之间作用的影响。为了解决第二个挑战，需要开发一个可视化分析系统 BeXplorer，这个系统是基于一个比较流行的 MMORPG 游戏的开发商的需求进行开发的。

2. 主要工作和目标

我们主要的工作分为两个部分，一个部分是对所获得的网易游戏的数据进行整理，分析，然后进行数学建模来构建一个对影响力比较敏感的模型来测量和分辨不同社会结构对玩家不同行为的影响。另外一个部分是构建一个可视化系统，整个系统由两个主要的可视部分组成的，一个是基于流的可视化，用来提供一个不同核心玩家之间以及玩家对不同商品消费影响的总体视图，另一个是基于像素的可视化，用来展示和可视化玩家的细节数据，比如玩家在游戏中的活动行为和玩家的属性值。这样就可以有利于我们去对基于流的可视化的一些特征进行更加深入的解释。我们所需要解决的问题如下：

- 在时间维度上，玩家之间的交流是如何影响他们的消费行为的？不同社会结构之间的影响力有什么不同？
- 哪一种玩家对其他人的影响力更大?那一类玩家更会受其他玩家的影响?
- 不同商品的消费受到的影响有什么不同？
- 在时间维度上，消费行为和交流行为是如何相互影响的？
- 当玩家影响的特定模式出现的时候玩家的交流网络结构是怎么样的？我们如何在那个时间点观察每一个玩家的社会行为?
- 当某种特定模式出现的时候，可视化系统是如何对假说在信息层面以及可靠性层面提供支持的？

3. 可行性分析

A. 数据检查及用户需求

我们从一个拥有百万活跃用户的大型 MMORPG 中获取了数据。整个时间的区间是 49 周。整个游戏在许多不同的服务器上同时自发的运行。在不同时间戳上的不同玩家的不同属性以及活动行为都被存储在了 CSV 文件中。

游戏中的每一个服务器在研究中都被当作一个虚拟的世界。每个玩家在游戏中创建的角色被标记为 *PlayerID*，每一个角色的属性都被分类成为静态属性（随着时间并不会变动）和动态属性（随着时间改变）。前者包括角色的性别（*avatar_gender*）和职业（*job*），后者包括：*online_time, grade, chat_frequency, count_pvp, consumption_records. VIP_level, Practice* 和 *Mastery*。这些属性的细节如下:

静态属性：

- *avadar_gender*: 角色的性别是由玩家在创建时候决定的。
- *class*: 角色的门派是由玩家在创建角色时选择的。在各个游戏中一共有 8 个职业，并且每个职业都隶属于一个门派，现在一共有四个门派，分布是神机营，昆仑山，逍遥宫和天工阁。为了简便，我们将他们命名为门派 1，2，3，4。由于同一门派在功能和行为特征上有相似之处，我们将在门派的层面对玩家的行为进行研究。

动态属性：

- *online_time*: 玩家在游戏上所花的累计时间。
- *grade*: 玩家的等级，从 1 级到 150 级。
- *chat_frequency*: 玩家和其他玩家之间的交流频次。玩家通过是以一对一的形式和他人交流的。每一次交流都是有时间戳以及交流上方的玩家 id 的记录的。对于他们所谈论的内容并不知道。
- *count_pvp*: 游戏允许玩家在 pvp 的模式下与其他玩家进行战斗。而 *count_pvp* 记录了玩家发生这种行为的频次。
- *consumption_records*: 玩家可以使用他们的虚拟财富来购买商品。虚拟财富包括元宝和银两。元宝的获得是通过充值现金或者用银两和其他玩家交换得到的。而银两则是玩家通过完成日常任务，击败其他玩家或者兑换元宝来积累的。为了让流通的货币统一，我们将银两通过每日的银两-元宝转换率转换成了元宝。所有的商品被分成了五类。商品的 ID，购买时间戳，购买数量以及花费都被记录了下来

- *VIP_level*: VIP 的等级是从 1 到 10 级，游戏中充值更多，那么玩家的 VIP 等级也就越高。
- *Practice*: 学习和提高技能或者完成不同的任务能够提升修为。
- *Mastery*: 一旦玩家加入了一个帮会，他会被赋予修炼属性，修炼会随着玩家获得的技能以及对帮会做的贡献而提升。

这些数据基本上满足了数据分析的需要。

B. 社交网络的行为分析

在 MMORPG 游戏中的众多可能的行为里，有两种行为被认为是尤其重要的：消费和交流。玩家消费行为的重要性非常好理解，因为它直接关系到开发者的收益和他们对这个游戏持续的投资。然而，玩家的消费是十分容易收到其他玩家的影响的，尤其是那些有影响力的玩家。而这种影响力可以通过三种机制来进行计算，分别是：直接交流，社会影响力以及三元闭包。

在 MMORPG 游戏中，玩家可以通过直接的交流对其他玩家施加影响。玩家之间的交流纽带将会成为 MMORPG 游戏的虚拟社会系统中商品以及意见的首要传输途径。关于商品的重要的功能质量信息都能通过直接交流进行传播。

社会影响力机制表明社会个体会由于同伴的行为而改变自己的行为已使得行为接近一致。通过观察 MMORPG 游戏里面其他玩家的行为和技能，一个玩家可能会去消费某些商品来提高自己。三元闭包机制提出一个玩家和另一个玩家在一个网络里拥有更多的共同交流伙伴，那么这个玩家将会更倾向于去消费更多商品来提升自己从而维持与其他玩家的交流。

直接交流，社会影响机制以及三元闭包机制并不相互独立。相对的，它们将会一起影响用户的消费行为。

C. 社交网络可视化

社交网络可视化在近几年已经受到了相当大的关注。而这些研究让我们的系统提供了很好的理论支持。

早期的社交网络可视化的工作主要集中在网络的拓扑结构，比如 Vizster 和 NodeTrix。Vizster 主要使用节点连接的可视化方法来可视化个人为中心的图，并且能够探索图的连接和网络结构。NodeTrix 是一个混合的展示方式，通过用矩阵的方式展示密集的连接节点。有些技术强调展示节点和边的属性，比如 Semantic substrates 在平面上赋予了节点不同的属性值。SaNDVis 提供多种视图和交互让用户来探索个体与实体之间的关系。

也有很多的方法被用来支持动态网络的可视化分析。Steffen 使用了聚类的超图来展示动态的聚类和聚合属性，使得用户能够找到并分析在时间上比较相类似的属性和子结构。Van den Elzen 将高维空间的网络分布图投影到了二维平面来帮助可视化。两

种方法都在宏观层面描述了网络，但是如果将这个直接应用到社交网络也许会丢失一些信息。

玩家与玩家，玩家与产品之间的复杂联系以及两种联系之间的动态关联是研究的中心。在某种程度上，我们的研究也被视作为社交网络的分析，用来辨认网络中节点和边的多种联系和结构。然而，当我们设计系统时候，我们需要展现更多的信息，比如交流，消费行为，参与的活动以及多种属性。因此，在这样的工作中需要一个更加统一的可视化设计来编码多种信息。

D. 暂存数据可视化

现在已经有许多可视化的技术用来寻找暂存数据的模式。主题流和故事线是非常经典的用来展现暂存数据的可视化技术。许许多多的研究都是基于这些可视化技术的。Tanahashi 优化了故事线的布局并且提出了将故事线的可视化应用于流型数据的框架。StoryFlow 将一种故事线的布局作为一种高效混合的优化方法，这样让实时的交互成为了可能。

最近几年比较流行的用来展现不断变化的文本和事件的方法是基于流的设计。LifeFlow 将连续的时间聚合到一棵树上，并且将他们通过一系列不同的彩色的方块进行呈现。同这个研究相同，Outflow 将不同的事件聚合到了一个图中，用边来代表同一时间内的转移。

我们所使用的游戏数据包括了玩家的动态属性，游戏的活动以及不同形式的行为。这些特征却让我们不能直接应用上面所说的方法。所以我们需要拓展现有的方法使得系统能够展现所种类型的随时间变化的的关系。

4. 初步技术方案和关键技术考虑

整个项目将会分成两大块来进行，一个是对现有的数据进行整理分析并进行建模计算结果，另一个是构建可视化的系统，能够将整理分析以及计算出的数据进行一个合理的交互展示，并从中发现有用的特征模式。

数据建模方面我们首先需要计算出 MMORPG 中玩家的核心度，然后通过一个影响与受影响的模型来计算不同类型之间玩家的影响力，不同玩家对不同商品消费的影响力。这些影响力都是通过特定的模型进行计算的。在这一个过程中需要我们运用数据库的知识，也需要对 python 和 R 的编程有一定了解，能够运用 python 和 R 里面有关于社交网络等方面的库。

构建系统方面，我们打算整个系统是一套 B/S 的架构，也就是基于浏览器的系统，在框架方面我们选用了 mean.js 的架构，mean.js 是由 AngularJS, Express, bootstrap,

node.js 和 mongodb 组成的框架，而我们所用的可视化的图形库是 javascript 上面对 svg 应用比较成熟也比较有名的 d3.js。

5. 预期工作结果

我们的系统希望能达到以下一些效果：

A. 对玩家动态的行为和影响力进行一个总结。

我们的系统希望能够提供一个有效的，简洁的对动态数据的总视图。需要用不同颜色的彩带来代表不同玩家在不同时间的不断变换的关于交流和消费的影响力。不同的彩带是平行放置的，这样就可以比较容易的比较不同类型的玩家或者不同类型消费之间的异同点。同样彩带需要自带明暗，用来编码关于消费的信息。这样就可将玩家的交流行为，消费行为以及消费和交流的影响展现出来了。

B. 能够对一些有趣的特征进行深度探索。

像素块以及平行坐标会被放在中间视图中来展现每一个玩家个体的交流行为以及在游戏中的一些属性。不同视图之间的关联允许我们找到一些反常的交流行为的关键特征。总视图和细节之间组合能够对假说进行更加深入的论证。

C. 提供视觉隐喻。

为了让用户有一个更加简洁和直觉的体验，我们的设计中有很多不同的隐喻。商品的消费信息被编码成流，交流信息的影响被编码成了流中的彩带，这些彩带根据流进行起起落落。这些设计构成了总视图。而交流的影响通过不同的方面展现出来：直接交流，社会影响机制以及三元闭包机制。

D. 独立的比较不同类别的玩家。

在不同缩放下的比较是必须的。在对总视图进行探究之后，我们可以将不感兴趣的消费流通过点击节点去除，在留下的流之间进行对比。

6. 进度计划

1、 完成数据建模： 2-3 周
2、 进行可视化设计： 2-3 天
3、 搭建系统框架： 3-4 天
4、 完成主视图： 2 周
5、 完成中间视图： 2 周
6、 完善细节视图： 1 周
7、 测试完善： 1 周

本科毕业设计外文翻译

文献题目：Different Aspects of Social Network Analysis

作者：Mohsen Jamali and Hassan Abolhassani, Web Intelligence Research Laboratory, Computer Engineering Department, Sharif University of Technology, Tehran, Iran

摘要：

社交网络是一群人（或者组织或者其他社会实体）的集合，通过朋友关系，工作关系或是信息交流所连接的。社交网络分析着重于对人际关系、组织和组织模式。社交网络分析同时提供了对人际关系的视觉和数学的分析。网络同样可以被当做社交网络。社交网络在网络上由超链接和网页构建而成。在这篇文章中，我们将对社交网络分析中的一些常用方法进行讨论

正文：

1. 介绍

社交网络是一种在个体和组织之间的社会结构。它表明了实体连接的紧密性，包括从初识到紧密联系。电子邮件，交通，疾病传染和重要活动都能通过社交网络所建模。

社交网络分析则是对人、组织、动物、电脑以及其他信息体之间的联系进行测量和映射的一种方法。网络中的节点是人和组，同时节点之间的连接则是表明了节点之间的关系。社交网络框架中比较有趣的一点是它的子结构是由群组和圈组成的。而这些子结构的数量，大小和连接情况可以告诉我们整个社交网络的行为状况。

从定义可以知道，社交网络数据可以被看做是由实体和联系组成的社会关系系统。社会关系系统也包含了实体属性和多重关系等额外的信息。

2. 社交网络模型

1) 用形式化的方法来展现社交网络在社交网络分析中用数学或者图形化的方法的原因之一是可以简洁系统地呈现对网络的描述。还有一个原因就是形式化的方法，尤其是数学，可以允许我们使用计算机去对社交网络的数据进行计算分析。第三个原因是这种方法对数据的呈现方式正式我们比较所希望看到的。

在整个社交网络分析中，大概可以分成以下三种方法：

- 描述方法，也可以通过图形的方法。

- 分析过程，经常是基于对邻接矩阵的分解。
- 基于概率分布的数据模型

2) 使用图来展示社会关系

网络分析使用由点和线组成的图来展现实体和联系。当社会学家从数学家中得到这种图后，他们将这种图称为社会关系网图。

社会关系网图有很多的变种，但是它们都有相同的特点：所有的实体都是用一个带标注的圈来表示的，而连线则表示两者之间有一定的联系。用社会关系网图这种视觉化的方法能够提供对社交网络数据的第一手描述。不过虽然对于小的关系网是可以满足需求的，对于那些比较复杂的关系网络就比较困难了。

3) 用矩阵的方法来表现社会关系

在社交网络分析中最普遍的矩阵是由实体组成的行和列以及关系所组成的元素构成的。最简单并且最常用的是二值矩阵。也就是说，如果两者之间有联系，那么就在对应的矩阵栏里填上 1，反之则填 0。这类的矩阵可以说是所有社交网络分析的初始矩阵，并且由于它能够表现在我们的社会空间谁与谁比较接近，这种矩阵被称为是邻接矩阵。一般来说，在一个有向图中，关系纽带的发出自是列，纽带的目标是行。

4) 社交网络分析的数据模型

对社交网络的数据化分析已经了持续了 60 多年。在 1970 年后，一个主要的方向是对实体之间进行可能性分析，虽然刚刚开始只是对小的群体进行分析。社交网络模型中的统计文献假设有 n 个实体和他们之间的联系信息。二元关系将会表现成为 n ×n 的矩阵 Y，其中如果 i 对 j 有某一种联系，那么 Y_{ij} 就为 1，反之则为 0。比如说 Y_{ij} 为 1 如果 i 认为 j 是他的朋友。实体通常表现为节点而关系则是节点之间的箭头。如果矩阵是 Y 自反的，那么关系箭头就可以当作是无向的。

更一般的 Y_{ij} 可以是有值的而不仅仅是二值的，代表了 i 和 j 之间的关系强度。除此之外，每一个实体都可以有特征集作为他们的统计信息。那么 n 维的向量 $X = x_1,....,x_n$ 就可以作为模型中被观察的协变量数据了。

但是已知的模型当中仍然有许多的问题，比如退化，拓展等问题。所以还是需要更多的去研究和发展。

3. 社交网络特性

社交网络有许多重要的特性比如大小，密度，度，可达性，距离，直径，侧地距离等。我们在这里会描述一些在社交网络中更加复杂的特性。

1) 最大流

当问到两个实体间总共有多少连接，就是问总共有多少不同的实体在到通往目标的路径上。如果我要向你传递一条信息，但我只能通过一个人来向你传递，那么我的连接是比较弱的，即使我可以通过的那个人有非常多的途径传递信息给你。另一方面，如果我可以通过四个人来传递信息，每一个人都有一条以上的途径将信息传达给你，那么我的连接是强壮的。这种流的方法表名了我对你联系的紧密度不会比连接中最弱的连接要强。

2) Hubbell 和 Katz 聚合

最大流问题关注的是两个实体之间联系的易碎性和冗余性，类似于弱连接的强壮性的命题。作为一种可选择的方法，我们也许回去考虑所有连接的强度。如果我们对两个实体之间互相影响的强度或者对同一处境的感觉的分享感兴趣的话，他们之间所有的连接都必须要考虑在内。

即使我们要考虑两个实体之间所有的连接，把长度为 10 的路径和长度为 1 的路径视为同样重要是没有什么特别大的意义的。Hubbell 和 Katz 的方法计算了实体之间连接的综述。但是对每一个连接都根据它的长度给了一个权重。长度越长，那么连接越弱。

3) 中心和权力

所有的社会学家都赞成权力是社会结构的基础特征。但是对于什么是权力却没有一个很好的定论，同样的对它的产生和结果也是众说纷纭。下面是一些社交网络分析所采用的分析权力的主要的方法。

度：指的是一个实体所拥有的连接数，度越大代表机会越多可变性越高。

亲密度：到其他实体的路径的长度，亲密度越高那么就更能与其他实体进行交互。

中间人：位于在两个实体之间，处理两者之间的关系来决定是维持关系还是分离他们。

4. 社交网络中的群组和子结构

社交网络分析中最令人感兴趣的是网络中的子结构。对网络结构的很多研究都着重于查看联系混合的紧密度以及这些连接是如何发展成为派系或者小群体的。

在社会结构中将实体分成派系或者小群体是非常重要的。这个有助于我们取认知

网络作为一个整体是怎么去运作的。比如说，假设一群在同一个网络中的实体组成了两个不重合的派系，在另一个网络中，实体组成了有重复成员的两个派系。我们可以观察到有重叠的两个派系之间的冲突要远远少于那些没有重叠的派系之间的冲突。当群组重叠的时候，动员和扩散行为将会快速散播到整个网络，当群组不重叠的时候一个群组的特征不太可能会传播到另外一个群组。

依照派系或者群组，一个图的特征能够很，明显的被观测出来：

- 子图的分离度有多少？
- 一个连接的子图有多大？是有一些大的群组还是非常多小的群组。
- 是不是有一些实体扮演了网络中的特定角色？

1) 派系

派系的概念是相对来说比较简单的。从最大众的角度来说派系就是指网络里一群联系非常紧密的人。大多数的人按照年龄，性别，种族，宗教等其他方式建立一个派系。

对派系更加强力的定义是指一个组里每一个人和其他剩余的所有人都有联系。最大完整子图就是这样的一个群组。

更加严格的派系定义是每一个人要和其他所有人都有直接的联系，我们一般不考虑这样的情况。我们可以将条件适当的放宽，允许有一些成员只有一些艺术的和其他成员的联系，这样对研究也能带来方便。

2) K-Plexes

一个最大完全子图的可行的方案是如果一个个体和除了 k 个成员之外的人都有连接那么这个个体也是这个团体的成员。比方说：如果 A 和 B、C 有连接，但是和 D 没有连接；同时 B 和 C 都与 D 相连，那么在 K-Plexes 的方法中这四个个体都属于同一个团体。在这种方法中，一个节点是一个大小为 n 的团体的成员需要这个节点与这个团体中的 n-k 个成员直接相连。这个方法旨在注重重叠和中心而不是独立和影响范围。

5. 社交网络中的群组和子结构

互联网是社交网络的一个例子。互联网网络是由网页与网页之间的超链接所构成的。为了利用超链接，需要建模互联网成为一个图，在这个图中顶点是网页而超链接是边。网页也许会被认为同文本或者多媒体内容比较近似，一个超链接往往是这个网页的作者认为的另一个网页与之有很大相关的明确指示。

推出和收集个人信息的可能性是互联网能从开始就成功的重要因素。很明显，直

到 2003 年互联网才成为绝大部分用户作为社交的活跃空间。那一年爆发式的出现了一种新的类型的网站，被统称为社交网络服务网站。最初的社交网站的创建者仅仅在几个月内就吸引了超过五百万的注册用户。

这些网站提供了对个人信息的共享和在线社交的访问途径和结构的形成。伴随着注册，这些站点允许用户去公布基础的信息，去邀请其他的用户来注册并且去连接其他用户的描述信息。这些网站同时也能够探索和可视化网络来发现一些有共同之处的朋友，或者是潜在的有共同兴趣的新朋友或者是之前失去联系的朋友。有些主题性的站点迎合了更加细分的目的，比如建立一些商业联系或者发现一些浪漫的关系之类的。

1、对互联网应用一些社交网络的分析。

从 1996 年开始，科学家们就对互联网图进行了一系列的社交网络的分析及应用，以此来达到对用户的查询能够反馈出最为权威的网页的查询回答。

1）谷歌的页面排名：如果一个用户在互联网上漫游无限的时间，对当前网页进行随机的跳转，跳出的可能性为 1-p，跳转到另一个随机网页的可能性为 p，那么不同的页面被访问的几率是不同的：一些比较受欢迎的网站由于有很多的进入的连接从而导致可能会被更多的访问。这种对受欢迎程度的测量方法被称为网页排名，被递归的定义为：

$$PageRank(V) = p / N + (1 - P) \sum(PageRank(u) / OutDegree(u))$$

这里是对网页 u 连接到网页 v 的集合进行求和，其中 N 代表的是在互联网这个图中所有的节点的总数。因此谷歌公司的搜索引擎大概的模拟了这样一个随机的在互联网图中的满目来达到估算网页排名的目的，而这个网页排名则可以被用来作为评估网页的受欢迎程度的重要依据。需要我们注意的是，上面所获得的网页的受欢迎程度的分数是被事先计算好的，是独立于用户所要查询的查询集的。这也是为什么谷歌公司的搜索引擎在用户搜索的时候能比其他相似的搜索引擎要快很多的原因。

2）Hyperlink induced topic search (HITS)

超链接诱导话题搜索可以说跟上面的相比是由些许不同的，如果是用这种方法的话，它不会去网络中抓取或者事先去处理互联网，而是取决于搜索的引擎。如果一个用 HITS 的查询语句被用于一个搜索引擎的话，它会先取回互联网图的一个子图，在

这个子图中的节点能够匹配查询语句。这些网站所引用的站点和那些引用了这些网站的站点都会被包含进来。每一个在这个扩展图中的节点 u 有两个相关系数，分别是 hu 以及 au,，这两个相关系数在初始化的时候都为 1。这样的话，HITS 就被迭代的定义为：

$a_v = \sum h_u$ 与此同时 $h_u = \sum a_v$ 。

在上面的公式中，在每一次的迭代之后，求和式都要做标准化的处理。a 这个相关系数代表了对网页权威性的一个测量，同样的，h 这个相关系数则是对网页中心性的一个测量。因为这样查询的结果的图的构建是基于查询语句的，使用 HITS 方法做搜索是会比谷歌公司慢的。这种技术的多种方法经常被用于基于连接的互联网的分析。这些方法通过预先抓取大量的成百上千的互联网页面来优化它们的速度。

6. 推理研究互联网上的群体

群体信息是互联网上的一个重要的活动。整个互联网上拥有着非常大数量的群体。一个群体是一群互相之间具有连接关系的内容创造者。很多研究人员的目标是在拥有很大一堆页面的情况下能够找出互联网上潜在的群体。互联网的连接结构呈现出了相当可观数量的潜在的人类的信息，因此可以对互联网结构的研究有着很有意义的帮助。现在有越来越多的研究工作被用于整合互联网上的文本和连接的信息，来用于组织，可视化并且搜索互联网上的各种超文本。在我看过的文章中有种方法可以从连接的拓扑结构中识别群体：

二分核心

一个完全的二部图是一个节点可以被分成两个集合 L 和 R 的有向图，其中 L 和 R 的并集是 V 并且 L 和 R 的交集是空集，每一个 L 中的节点与 R 中的每一个节点相连。用符号 K_{lr} 来表示一个完全的二部图，其中 $l = |L|$ 并且 $r = |R|$。首先，一个二部核心，K_{lr}，有一种特性那就是所有的在 L 中的节点有一个不低于 r 的书目耦合值，并且所有的 R 中的节点都有一个不低于 l 的协同引证耦合值。因此，二部子图是由相似的最小度的节点组成的。而且这些子图展现出来的结构是经验主义上的互联网社会中的核心。

7. 作为社交网络的博客空间

博客已经成为了互联网中非常显著的一个社交媒体，它能够让用户快速并且简便的将一些非常有自己想法的内容给发布出来。一个博客是一个典型的站点，它有不同的书写的时间顺序并且被用户用一些比较特殊的工具来维持。既然一个博客的入口可

以被超链接到网页或者别的博客的入口，博客和连接的结构信息可以被看做多个社会的网络。

一个博客本质上是一个在互联网上可行的日记。写博客是更新博客的一种方式，那些坚持记录博客的人是博主。博客是可以每天用软件更新的互联网的站点，博客甚至允许那些对电脑技术并不是非常在行的人进行日志的记录和书写。博客是按照时间顺序进行发布的，并且最近的日志将会在最显著的位置。

一些博主会列出一些别的博客通过博客链接，也会阅读以及评论其他博客的文章。博主们经常会阅读其他人的日志，并且这种阅读和评论他人日志的现象在一个用户在线浏览的活动中是非常的常见的。这些文章和评论能够传播最近的感兴趣的事情并且让人发现一些深刻的博主偶然遇到的信息。这些信息都是作者发表出来希望大家能够看到并且有所体会的。

网络上的博客是互联网的一个子集，所以互联网上的博客也可以被认为是一个社交网络。但是如果我们考虑到评论和博客的入口连接。网络博客的受欢迎的程度可以由以下两个决定：连接到这个博客的超链接数以及这个博客的被访问数。很多研究着付出了很多的努力去分析博客的评论以及博客受欢迎程度和评论模式之间的关系。通过网络博客中的链接结构来建立一个互联网的推荐系统。

Cameron Marlow 采用了一种社交网络的分析方法来描述博客的社会结构。他探索了两种权威性的测量方法：通过测量博主们的公众喜爱程度来展现受欢迎的程度以及测量通过引用他人文章说带来的影响。Ko Fujimura 提出了一种新的算法，这种算法是通过给每个入口一个分数，这个分数是通过评估博客的接口以及通过基于特征向量的计算来获得的博主的权威值。这个算法能够使一个高的分数被用于一个好的博主所推送的博客，但是不会将这个分数联系到其他的之前被引用过的相关博客。通过这个算法，博客中心以及权威评分是通过博主进行计算的，并且通过其他博主的评分给这个博客进行权重。博客的受欢迎程度并不能单纯通过接入的连接来进行预测。

8. 总结

这篇文章中，我们大概浏览了社交网络，形式化的表现方式以及社交网络特征。社交网络分析方法为社交网络分析结构的各个不同的方面提供了很多有用的工具。

网络自身可以被当作是社交网络。在互联网的社交网络中，文本是社会图的节点，文本之间的超链接是社会图的边。网络博客可以作为互联网社交网络的一个特殊的子

集。我们已经讨论过了互联网博客的一些特殊的连接结构。