**班 级 1705012**

**学 号 17050120011**

****

本科毕业设计论文



**题 目** 复杂网络大数据分析中的链路预测：

局部路径指标的精度分析

**学 院** 物理与光电工程学院

**专 业**  电子科学与技术

**学 生 姓 名**  农 昂

**导 师 姓 名** 张 睿 挺

**院内导师姓名**

西安电子科技大学

毕业设计（论文）诚信声明书

本人声明：本人所提交的毕业论文《复杂网络大数据分析中的链路预测：局部路径指标的精度分析》是本人在指导教师指导下独立研究、写作的成果，论文中所引用他人的无论以何种方式发布的文字、研究成果，均在论文中加以说明；有关教师、同学和其他人员对本文的写作、修订提出过并为我在论文中加以采纳的意见、建议，均已在我的致谢辞中加以说明并深致谢意。

本论文和资料若有不实之处，本人承担一切相关责任。

论文作者： （签字） 时间： 年 月 日

指导教师已阅： （签字） 时间： 年 月 日

# 摘要

复杂网络一般是由网络节点与节点之间形成的连接组成。若将我们身边的很多领域进行抽象看待，会发现它们中的大部分都可以用复杂网络来进行表示。而链路预测，是对复杂网络进行研究的一个重要手段，具有巨大的科研与应用价值。具体来讲，链路预测就是对网络中的链路进行预测，是一种利用已有的信息去判断网络中的某一条连边有多大的概率存在的研究。近年来，该方向的研究获得了很多的成果，一些优秀的链路预测算法也相继被提了出来。在本文中，我们主要对局部路径指标的链路预测算法的精度进行分析。通过在两个模型网络以及六个真实网络中进行大量的数值模拟表明，与两个众所周知且广泛使用的指标:CN指标和Katz指标相比，局部路径指标具有较高的准确率和效率。事实上，局部路径指标提供了与Katz指标一样具有竞争力的预测精度，同时比Katz索引需要更少的CPU时间和内存空间，因此局部路径指标是大型网络数据挖掘中潜在实际应用的有力候选。

**关键词：复杂网络 链路预测 局部路径**

# ABSTRACT

Complex network is generally composed of network nodes and the connections between this nodes. If we look at many fields around us abstractly, we will find that most of them can be represented by complex networks. Link prediction is an important means of complex network research, which has great value in scientific research and application. To be specific, link prediction is to predict the links in the network. It is a kind of research that uses the existing information to judge the probability of the existence of a certain link edge in the network. In recent years, a lot of achievements have been made in this field, and some excellent link prediction algorithms have been proposed. In this paper, we mainly analyze the accuracy of link prediction algorithm based on local path index. A large number of numerical simulations in two model networks and six real networks show that the local path index is more accurate and efficient than two well-known and widely used indexes :CN index and Katz index. In fact, the local path index provides the same competitive prediction accuracy as the Katz index, while requiring less CPU time and memory space than the Katz index. Therefore, the local path index is a strong candidate for potential practical applications in large-scale network data mining.

**Key words: complex network link prediction local path**

目录

[摘要 Ⅰ](#_Toc8479)

[Abstract Ⅱ](#_Toc30079)

[第一章 绪论 1](#_Toc13771)

[1.1研究背景及意义 1](#_Toc24467)

[1.2链路预测问题研究现状 2](#_Toc31077)

[1.3本文主要研究内容 6](#_Toc8413)

[第二章 相关理论介绍 9](#_Toc23648)

[2.1 图及其类型 9](#_Toc8709)

[2.2 网络的拓扑性质 11](#_Toc32440)

[2.2.1 度、平均度、网络密度与度分布 11](#_Toc23852)

[2.2.2 网络的连通性与路径 12](#_Toc11223)

[2.2.3 直径、平均路径长度与效率 13](#_Toc27317)

[2.2.4 聚类系数 14](#_Toc12973)

[2.2.5 社团结构 14](#_Toc32129)

[2.3链路预测理论概述 15](#_Toc25472)

[2.3.1 问题描述 15](#_Toc4608)

[2.3.2本文主要算法介绍 15](#_Toc2366)

[2.3.3 性能评价指标 17](#_Toc28345)

[第三章 实验结果及分析 19](#_Toc8790)

[3.1实验设置及实验环境说明 19](#_Toc19010)

[3.2实验数据集 22](#_Toc32100)

[3.3结果与分析 24](#_Toc18875)

[第四章 结论与展望 27](#_Toc16436)

[致谢 29](#_Toc25664)

[参考文献 31](#_Toc21480)

# **绪论**

## 1.1研究背景及意义

联系，是一种特征，体现了不同个体之间的相互关联，在各种系统中我们都能看到它的存在。如果我们将各种真实的系统简单的看待，只考虑系统中的对象以及对象之间的联系，就会发现它们其实就是一个一个的复杂网络。对象对应于节点，联系就是节点之间存在的连边。如今各种技术都在飞速发展，在我们的日常生活中复杂网络已经成为非常普遍的存在，比如因为人与人之间不断交流而形成的社交网络、经过通信人员多年的建设而形成的通信网络、以及等待着研究学者们去探索的人脑神经网络等。可以说，无形之中网络这个东西已经将我们包围，我们就是生活在各种各样的网络中。如果我们能揭开复杂网络的神秘面纱，了解其演化机制，那么对于我们的生活以及日常的生产活动都将有着非常大的帮助。因此，复杂网络的研究现在已经成为许多学科共同关注的研究问题。

链路预测作为复杂网络研究的重要分析手段之一，其再在推动网络科学和信息科学的发展方面有着非常大的作用。在一个真实的网络中通常存在着丰富多样的隐含信息。链路预测分为两步，第一步是对网络中现有的信息进行收集并整理，第二步就是利用上一步获得的信息对网络的机构演化进行预测。预测的目标一般来说可以分为两种，一种是当前已经存在但还没有被发现的连接；一种是当前还没有出现，未来才会形成的连接[1][2]。简单来说，链路预测就是预测网络的演化机制，判断网络中的某一条连接是否存在。

毫无疑问，链路预测具有重大的实际应用价值，因此越来越多的研究学者将目光投向对它的研究。例如，在生物研究领域，蛋白质是一个及其重要的研究热点。将蛋白质的相互作用抽象成网络，在一般的情况下，如果我们想要知道某条边是否存在，那就需要进行大量的实验。但截至到目前为之，人们对蛋白质的研究依然处于探索阶段，对蛋白质间存在的相互作用了解的还很少，不到0.5%[3][4][5]。在这样的情况下，单纯的仅仅依靠实验去进行网络中相互作用的验

证，工作量巨大，将是一项非常消耗人力、物资和时间的工作。因此我们需要做的是尽量减少对一个网络进行预测时的工作量和花费的成本，这时，链路预测算法将体现巨大的作用。先根据现有的信息，设计出一种合理且对目标网络运行高效的算法，然后将这个算法运用该网络，根据运行出来的结果，我们就能知道那些条边存在的概率比较大，最后着重对存在概率大的边进行实验验证，这样我们就能提高实验的成功率并且减少一些不必要的实验，更快的揭示该网络的本质。另外，在收集一些社会网络的信息时，可能会出现由于技术不成熟导致收集到的网络信息结构具有一定缺失的情况，这对于对该网络的研究将产生不良的影响，这时候通过链路预测将网络的拓扑结构补全，然后再对该网络进行分析[6][7]就能消除这些影响。更加重要的是，凡是那些可以抽象成网络形式的系统链路预测都能够发挥其作用，对系统中可能存在的相互作用进行预测。由于这一特点，把链路预测应用于在线社交网络[8]或者电商网络中，将会产生巨大的商业价值。比如在社交网络中，通过已有的人际关系网络，去预测新的网络链接，从而推荐现实中可能认识的人互加好友，从而提高用户对产品的忠诚度。此种方法也同样适用于电商网络。链路预测还有修正的作用，比如通过链路预测可发现原先在网络中存在的链接可能是错误的[9][10]，就可以通过人为的去对错误的链接进行修正，将网络结构进行优化。这一特点在构建蛋白质网络时就曾有过运用[9]。

综上所述，我们知道无论是在探索未知领域，还是在优化现有系统方面，链路预测都在发挥着巨大的作用。链路预测这一项工作以较少的研究成本，带来了可观的经济效益，是具有非常大的研究价值的。

## 1.2链路预测问题研究现状

复杂网络的研究距今已有20年的历史。在这20年里，科学家们几乎在每个领域里都进行了链路预测的研究。不同的领域所形成的网络特点总是有所差别的，而面对这些特点不一的网络结构，学者们也对此提出了适用不同网络结构的算法，方法分类如图1.1。

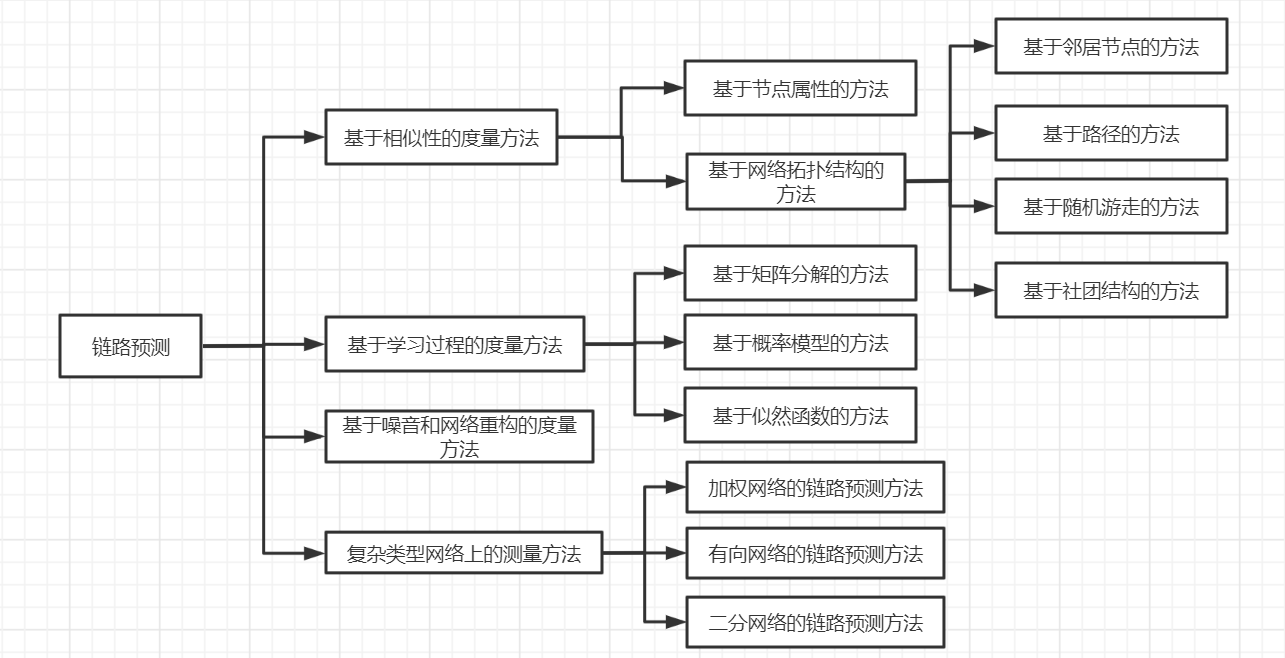


图1.1 链路预测方法分类图

最早的链路预测采用的方法比较简单，主要是通过利用网络中节点的属性信息进行预测。这种方法取得了较好的预测效果，但存在一个弊端。因为在大多数情况下，节点的属性信息是很难获取到的。正是由于这样的原因，很多学者不得不去寻找一种更容易实现的方法进行链路预测，而通过不断的思考，学者们想到了通过网络的结构去进行预测。网络结构是一个比较宏观的东西，相比于节点信息不那么具体，但这也意味着更容易获得到，并且一个网络的结构一般不轻易改变，具有更好的可靠性，同时这一类方法能够非常好的适用于那些具有相似结构的网络。在进行链路预测时我们通常会尝试去获取到网络节点和网络的拓扑结构信息，如果将这两种信息都利用上的预测算法一般我们称之为基于相似性的算法。按研究范围大小来划分，这类算法大致可以分为三类[11]，基于局部、全局以及准局部三类。

基于网络局部结构的预测算法是比较简单的，但也是最直观并且相对有效的。这类算法考虑的信息非常少，仅仅是利用了网络节点的邻居信息，不考虑额外的因素，计算复杂度很低，因此非常适合在大规模网络中应用。但是受到信息量的限制，它的预测精度往往不太理想，比如主要考虑共同邻居数量的共同邻居指标[2][12]、主要考虑节点度乘积信息的偏好连接指标[2][13]、主要考虑共同邻居度值信息的Adamic-Adar 指标[14]、主要考虑共同邻居相似性的Jaccard指标[2][15]、被称为余弦相似性指标的Salton[2][16]和从资源角度出发的资源分配指标[17]等相似

性指标。相比之下，基于全局结构的相似性算法考虑了整个的网络结构信息，因此在进行预测时能有更多的参考信息，使得预测精度有了较大的提升，但由于需要计算全局的节点，因此计算时间也大大增加了，时间复杂度较高。这类算法的代表有考虑网络所有路径的Katz指标[2][19]，与其相似的LHN-II指标[2][18]，通过引入随机游走粒子的重启的随机游走指标[20]、SimRank指标[2][22]，以及考虑时间因素的平均通勤时间指标[2][21]等。而基于准局部结构的预测算法，则是可以看作是以上两种算法的中和，它考虑了比局部指标更多的因素，但又不需要考虑与全局指标一样多的信息量，具备了较好的算法精度同时时间复杂度也较低。这类算法的代表有只考虑局部的局部随机游走指标[23]和只考虑局部路径的局部路径相似性指标[17]以及存在叠加效应的局部游走指标[22]等。

一直以来对基于相似性算法的研究从未间断过，经过学者们的努力，一些优秀成果也相继出现。在引用网络方面，贾等人通过对大量的网络实验进行总结，提出了一个全新的指数：H指数[24]，并将网络中度的概念清除了，改为用H指数表示。与先前的研究方式不同，他们对每条引文都设置了权重，通过权重来区分引文的重要性，然后再在 Sorenson指标[2][25]、Salton指标[2][26]和 AA指标[14]三种链路预测方法的基础上进行了改进。大量的实验结果表明他们的改进方法很有效，在引文网络中拥有优秀的表现。经典相似性指标一般都与共同邻居节点有着联系，Cannistraci 等人[27]将这种联系进行了增强，成功设计出了一种新的链路预测算法：资源分配相似性预测算法CAR。周等人[28]提出了在复杂网络中知识是依靠路径进行传播的假设，定义了知识在复杂网络中的传播机制。在这个假设的基础上，吕等人[29]又提出了KDLP全路径预测指标。高等人[30]将网络分成一个一个的局部结构，并以网络中最小的局部结构三元闭包为单位，单独考虑节点在相邻的三元闭包中的权重，并把这个权重作为节点相似性指标的一个影响因素，由于这样做使获得的节点信息更精确，极大的提高了算法的预测精度，但同时也令算法的时间复杂度大幅上升。刘等人[31]提出了扩展资源分配指标，如果一个端点与另一个端点的共同邻居与非共同邻居有着很大的资源交互，那么这两个端点就具有很高的相似性，反之则很小。孙等人[32]通过节点亲密关系这一角度出发，提出了局部亲和结构指标。武等人[33]设计了一个考虑网络节点聚类系数的预测指标CCLP，并在该指标上进行扩展，在计算相似性时不仅将节点聚类信息考虑进

来，同时也把边聚类信息也加以考虑，提出了NLC算法[34]，最大似然估计类算法是通过模型与现实结构进行对比来进行预测，不断更改模型的参数，让其与已知的结构能进行最大概率的结合，以此来寻找网络中缺失的链接。目前，这类算法主要存在两种，一种是在网络的层次结构信息基础上进行研究的，由 Clauset 等人[35]在 2008 年提出。这类方法对两节点之间是否存在连接进行判断时，不是考虑节点本身，而是通过节点的共同祖先的存在概率值来判断。两节点存在共同祖先的概率越高，它们之间越有可能存在连接，类似于一种家庭关系。另外一种是随机块模型方法，是由Guimerà[9]提出的。如图1.2所示，在这个模型中，如果集群具有很高的密切度，那么两个节点之间将有很大的概率进行连接。另一种最大似然方法是闭路模型方法，是由吕等人[36]提出的。寻找出网络中的闭合回路，并将其看成是一种局部性，网络中的似然值通过闭合环路的多少来进行定义，如果我们想要知道某一条缺失边存在的概率，那么向该网络添加这条边后计算出网络的似然值，这个值就是我们想知道的缺失边存在的概率。总而言之，这种基于最大化网络似然函数的预测算法通过模拟网络的最大结构，可以很清楚的将缺失的连接表示出来，并且还能让我们看到网络更深层次的特征，但存在的问题是这类算法一般都具有较高的时间复杂度，因为它对网络顶点的个数非常敏感。还有一种是基于矩阵分解的预测算法[37][38]，一般网络数据中会存在着许多潜在的特征，这种方法是先将这些潜在特征寻找出来并且对其进行学习，利用潜在特征的信息去对网络连接进行预测。每个节点都会有一定数量的潜在特征，通过节点的潜在特征相似程度来对节点的相似性进行衡量[39]。

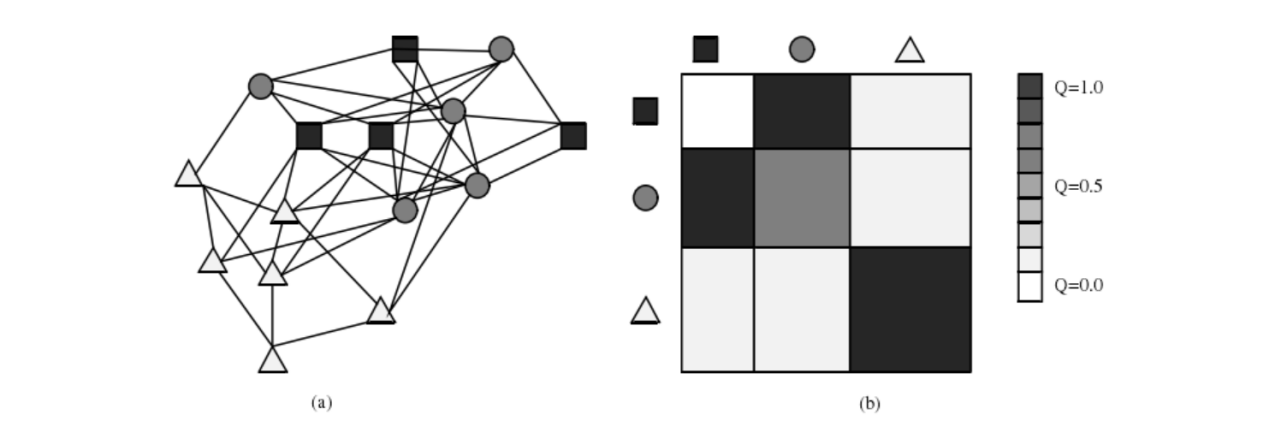


图 1.2 网络（a）与随机块模型（b）

除了以上提到的算法，在对网络的噪音过滤以及对网络进行重构方面还存在

一些特定的链路预测方法。这类方法考虑到了网络可能存在噪音等情况，在算法逻辑中进行了特殊的处理，使得算法具备良好的健壮性。如Zhang 等人[40]为了弥补局部路径上路径提供的参考区分度不大的缺点提出了加权的局部路径链路预测指标。为了解决在进行网络重构后网络的结构特性很难保持与原网络一致的难题，Zeng等人[41]提出了一种预测模型，该模型由共同邻居和网络中边的中心性混合生成。实验表明，这个模型很好的解决了上述问题。吕等人[42]又从邻接矩阵的扰动因素出发，提出了结构一致性预测指标。类似的工作还有对于存在噪音的网络环境下，然后对算法的健壮性进行评价的评价方法[43]，还有研究能将网络中出现的噪音过滤的链路预测方法[44]等。

在目前的链路预测算法研究中，由于设备与技术的限制，研究对象更多是规模较小的小型网络，对于大规模网络的研究还是比较稀缺的。同时，现在网络的表现形式大都是以离散邻接矩阵来进行表示的，学术界目前主流的研究方向也是对该种表现形式网络的研究。然而，随着科学技术的飞速发展，大规模信息的网络时代即将来临，这种网络表示方式也出现了越来越多的问题。一方面来说，离散邻接矩阵表示的网络能获取到的信息较少，只能够获取到顶点之间的相邻关系，在这种情况下，当需要构建当下复杂网络中更复杂的高阶网络结构关系的时候，比如路径、频繁的子结构等，则会因信息不足而导致构建出来的子结构不是很完整，同时在这种网络表现形式下，如果我们想要获取到节点除属性外所携带的外部的信息也会存在较大的困难。另一方面，当系统需要处理的网络规模非常巨大时，网络的离散邻接矩阵将也会变得非常巨大，因此想要完成链路预测就需要对处理器的性能和存储有较高的要求。由于以上两个原因，部分学者们意识到想要克服今后网络链路预测所可能遇到的困难，就需要从网络表示的方式寻找突破口，让其对于未来更多多样化的链路预测需求能提供更好的帮助。遗憾的是，该方向的研究进展并不是很顺利，到目前为止，网络表示算法的数量仍然是很少，而其中大部分又都是和任务无关的实现。

## 1.3本文主要研究内容

链路预测的研究存在许多困难。一个是目标网络的稀疏性，这就导致了网络中的链路具有很小的先验概率，这是一个非常严重的问题，使得在建立统计模型

时需要面对很大的困难。另外一个问题是真实系统的网络规模往往是巨大且复杂的，因此在对系统进行计算时就需要非常高效的算法。但是，作为在实际应用中的一个非常关键的因素，算法的计算时间以及算法所需要内存的复杂性都还没有得到系统性的研究。一般来说，一个算法的精度往往与它的计算复杂度是成正相关关系的，即越高的精度也通常意味着越高的计算复杂度。需要注意的是，如果消耗的时间或内存是不可接受的，那么任何高度精确的算法都将变得毫无意义。因此，设计一个准确又快速的算法是一个很大的挑战，尤其是对于稀疏和庞大的网络。

在本文中，主要是对局部路径（LP）指标链路预测算法进行分析研究。对模拟网络和真实网络的大量数值模拟实验表明，该相似性指标同时兼顾精度与效率，其预测精度远高于普通邻居（CN）指标，与Katz指标具有竞争力，并且高效，计算所需要的时间和空间远小于Katz指标。特别是，当网络很大时，局部路径指标与Katz指标相比将显示出很大的优势，因为计算后者要求CPU时间缩放为网络大小的立方，而计算前者要求线性CPU时间作为网络大小。最后，我们先是在组成真实世界网络的两种常见模型，小世界网络模型和无标度网络模型中进行实验，初步验证三种预测指标的精度分析，然后再通过六个真实网络数据实验结果的对比，完成对LP指标算法的进一步分析。

# 第二章 相关理论介绍

## 2.1 图及其类型

网络描述的是对象与对象之间的关系，本质上是一种集合。在进行网络的研究时，一般是将其抽象为图的结构。图的节点就是网络中的对象，图中的边则是各个对象之间的联系。可以说，图是进行复杂网络研究的一种重要工具，能帮助我们对网络的结构进行更好的理解。

通常，网络都可以抽象成图。图在数学上的定义为 G=(V,E)，其中V是由网络中一个一个的顶点组成的集合，E 则是网络中联系的边的集合。假设定义一个网络，在这个网络中存在的节点总数为Y，边的数量总共为X，而节点对组成的全集是P，那么存在有Y=|V|，X=|E|,P=N(N-1)/2。在后面的描述中，本篇文章对以下三组用词不做任何区分，每个组的两个词等价。（1）图和网络；（2）节点和顶点；（3）链接和边。

图具有很多不同的种类，它们根据连边的特性进行区分，目前主要有四类。图2.1是四种不同类型图的示意图，为了方便理解，我们还将它们之间的关系画了出来，如图2.2。节点与节点之间的连接也会存在强弱之分，在这里，我们用权来表示这种强弱程度。计算机实际分析中，常用于表示图结构信息的方式主要有两种：邻接表和邻接矩阵，邻接矩阵是这篇文章的主要研究对象。对于网络信息与邻接矩阵的关系我们是这样规定的：如果一个网络含有M个节点，那么它的邻接矩阵就是一个M阶的方阵，用数学表示即为： A={bij}M\*M。在矩阵中我们用下标来确定元素的位置，比如bij就是表示第i行第j列的元素，这个元素的定义与网络类型有着密切关系，不同的网络类型也会导致该定义有所区别。

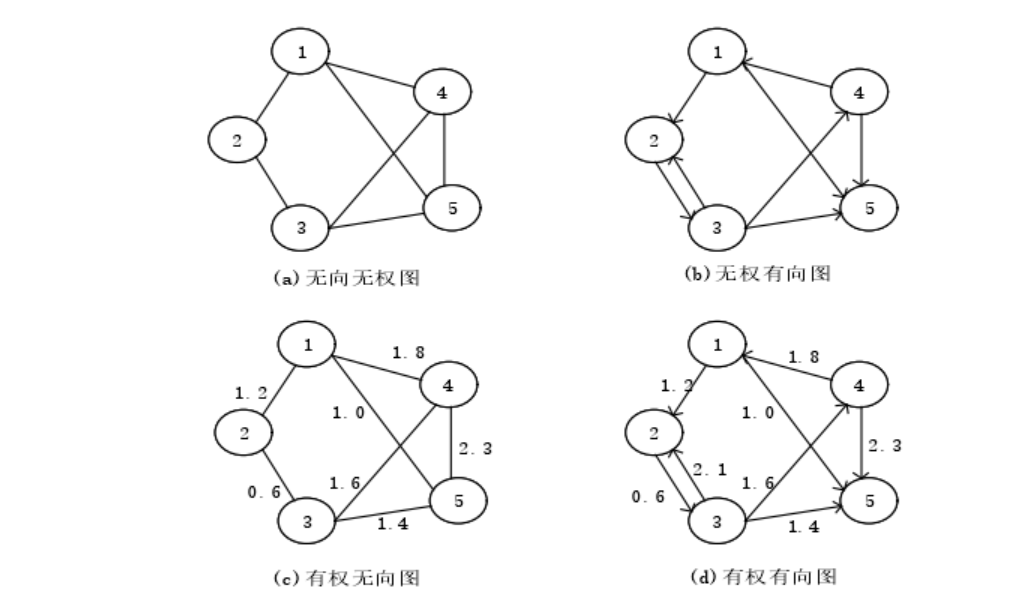


图2.1 四种类型的图的示意图

当一个图中只是单纯着存在连边关系，节点之间的连边没有重要程度的区别，并且也没有方向之分的时候，这样的图就是无向无权图。它的示意图如图2.1（a），比如同班同学之间的好友关系网络。这一类网络的邻接矩阵A中的元素定义为：

（2-1）

有权无向图表示的是图中节点之间的连边存在方向的概念，但是依然不存在强弱程度的这一类图。示意图如图2.1（b）。这类网络中比较常见的比如微博用户与它们关注的博主之间的关系，明星与粉丝的关系等。这个时候bij的定义如下：

（2-2）

在有权无向图这一类图中，节点之间的连边存在强弱程度的区别，但没有方向之分。示意图如图2.1（c），比如演员与导演合作网络，两者之间的连边不区分方向，二者一起合作作品的数量就是它们连边的权重。网络对应的bij定义

为：

（2-3）

有权有向图是四种图中最复杂的一种，它节点之间的连边既存在强弱程度的区分，同时还有连边方向的不同。示意图如图2.1（d）。这类网络在现实生活中也比较常见，交通运输网络就是这类图。交通网络中的节点就是站点，站点之间的连边方向代表了运输的方向，连边的强弱程度代表了两个站点之间的运输频率。这类网络中bij的定义如下:

（2-4）

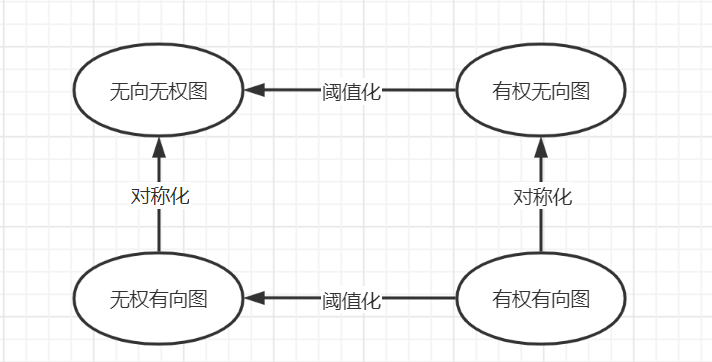


图2.2 四种类型网络图的关系

在本文中，我们主要是对无向无权图这种网络进行研究。

### 2.2 网络的拓扑性质

在科学研究过程中，常常会出现一些对研究内容结论影响不大但研究对象又具备的因素的情况，为了使我们能对研究对象能更直观的观察，常常会画出对象的拓扑图。对于网络来说，拓扑性质[45][46]是仅需要对网络中的节点和连边进行考虑的一类性质。在本章节中，我们将对部分基本的拓扑性质进行一个简要介绍。

2.2.1 度、平均度、网络密度与度分布

度是节点的一个非常重要的属性，用k来表示。节点的度表示了该节点存在

连边的数量。网络中会存在很多节点，为了能更好的衡量网络的连边密度，学者们又提出了一个平均度<k>,代表了整个网络中节点度的平均值。当一个网络G的邻接矩阵为：A={aij}M\*M，那么它的<k>和ki的计算表达式分别为：

(2-5)

(2-6)

当一个网络的边存在方向性时，节点的度又可以被分为两种：入度以及出度。从上面度的介绍中我们知道，度是节点连边的个数，入度，是指连边的方向是从其他节点开始，而指向当前节点的连边的数量。出度则与入度相反，始于自己，终点是其他的节点。如果一个网络的节点总数为N，存在M条边，那么这个网络的密度ρ定义如下：

(2-7)

在一个网络中，并不是每个节点的度都是一样的，因此出现了度分布p(k)这个概念，表示了某个度在网络中的分布情况，定义如公式(2-8)。经过多年的探索，学者们也将一些常见的度分布进行了归类，比如高斯分布，幂律分布等。度具有幂律分布的网络在我们生活中是比较常见的，比如在线社交网络、电商网络等。幂律分布的特点就是节点中的度呈幂次的特征，也就是说网络中节点度的大小一般会相差很大。因此，学者们针对度分布特性又提出了一个新的指标：度异质性H，通过它我们就能知道网络中节点度的分布均匀程度。它的定义如公式（2-9）：

(2-8)

(2-9)

除了度分布以外，网络之中还存在一个联合度分布，它的定义如下：

(2-10)

2.2.2 网络的连通性与路径

存在一个网络G=(V,E),如果网络中不存在孤立的节点，换句话说，网络中的

每个节点都至少与其他一个节点产生路径连接，则这个网络G是联通的，否则，这个网络是不连通的。既然产生了链接，那么就会存在路径。在网络中，我们采用节点序列来代表一个路径。比如p={v0,v1,v2,…vl}，规定（vi,vi+1）∈E（0≤i<l），就是代表了一条路径p，这条路径的长度就是集合p里节点的数量。每个节点对都有可能产生路径，并且有可能产生多条，如果每一条路径都用上述形式表达，那将是非常麻烦的工作，并且也不直观，因此对于不同节点之间存在的路径数我们使用邻接矩阵来进行表示。比如若有这样一个节点对（vi,vj）,它们之间仅存在一条连边并且路径的长度为1，则用bij=1表示。又比如若节点对（vi,vj）存在多条路径，数量位置，但可以确定路径的长度都为2时，那么这个节点对之间的路径总为：

(2-11)

根据以上的推论，我们可以知道节点对（vi,vj）之间存在长度等于 l 的不同路径的数量为：

(2-12)

2.2.3 直径、平均路径长度与效率

在一个复杂网络中，两个节点常常存在多条路径，有的路径包含3个节点，有的包含5个，在这些路径中，我们将包含节点数目最少的成为最短路径，记作dij。网络的直径D就是这些最短路径中的最大值，即：

(2-13)

为了刻画网络中整体路径的长度，学者们提出了平均路径长度<d>[47]该值越大，证明网络中的路径普遍都很长，反之则很短，它的定义如下：

(2-14)

由上面的式子可以知道，如果网络中的存在孤立节点的情况，那么将导致<d>变成无穷大。使我们得到错误的网络参数。为了解决这个问题，学者们又提出了网络效率 e[47]来表示节点间发送消息的效率，定义如（2-15）所示。效率高意味着路径会很短，反之，效率底的话证明网络中的节点之间存在的路径普遍都很长。

(2-15)

2.2.4 聚类系数

复杂网络中节点的紧密程度我们使用聚类系数来进行衡量。对于度为ki的节点vi，它的聚类系数Ci[46,47]定义为：

(2-16)

对于上面的定义，Watts[50]等人从一个新的角度出发，在另几合图形中做实验，发现节点 vi的聚类系数 Ci等价于包含vi的三角形数目在以节点vi为中心的连通三元组中的比例，即：

（2-17）

与平均路径类似，网络中也同样存在平均聚类系数[46][47]，用<C>表示，定义如下：

(2-18)

2.2.5 社团结构

如同我们生活中所见到的社团一般，社团结构就像是一个群体，社团中的个体会因某种联系而聚集在一起。在网络中，同样会由于节点的聚集而形成一个一个的社团结构。相对于不同社团之间的节点来说，同一个社团内部的节点间会有更紧密的连边关系。

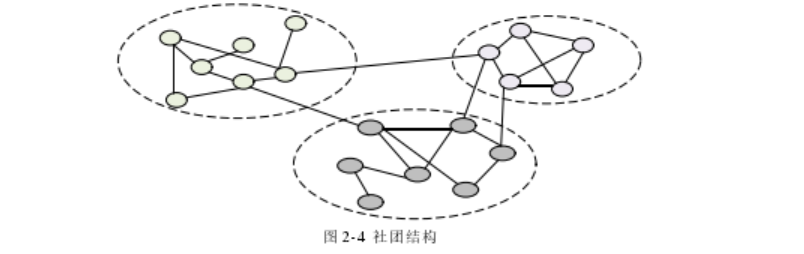


图2.3 社团结构

## 2.3链路预测理论概述

### 2.3.1 问题描述

复杂网络中链路预测简单来说就是对网络中连边的生成进行预测，去发现尚未被发现或者尚未存在但将来有很大概率存在的连边。这项工作主要分为以下几步：第一步就是先将网络的信息抽象成图；第二步将得到的图进行处理，将其分成两个部分：训练集和测试集。在这里，我们将为每一个节点对分配一个分数值，该分数值通过算法计算而得。分数值的大小衡量了两个节点之间产生连边的概率。即分数值越高，它们之间存在连边的概率越大，反之越小。第三步，在训练集中根据预测算法将不存在连边的节点对的相似性分数计算出来，并按这样相似性将节点集合进行降序排序，这个集合就是预测所存在的边的集合，最后通过测

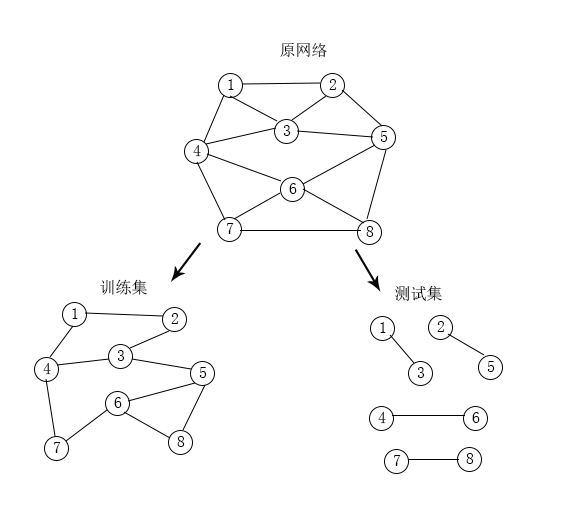
试集来对这个算法的精度进行判断。图2.4是训练集与测试集的划分示意图。

图2.4 训练集和测试集的划分示意图

### 2.3.2本文主要算法介绍

在本文中，主要目标是对LP指标的精度进行分析，为了能更直观的了解该指标的特性，我们选取了另外两种指标来进行比对：CN指标以及Katz指标。以下是它们的定义及介绍：

1. 公共邻居（CN）指标：

该指标将共同邻居的数量作为链路预测的主要参考因数，两个节点的共同邻居越多，将来则越容易形成链路。对于一个节点x，让Γ（x）表示x的邻居集合。CN指标的定义如下：

(2-19)

可以知道Sxy=A2xy，其中A网络的邻接矩阵，和前边介绍的一样，在邻接矩阵中，如果节点x和节点y是相连的，那么就用Axy =1进行表示，否则Axy=0。而A2xy则代表了x与y相连的路径中长度为2的路径数。纽曼[48]在进行协作网络的研究时，对这个量进行了使用。一些更复杂的指标，比如Sørensen指标、 Jaccard指标以及Adamic-Adar指标等，也可以归类为基于CN指标的指标。然而大量的实证分析表明,CN指数比那些复杂的变体表现得更好。因此，在本文种选择CN作为所有基于CN的度量的代表。虽然CN耗时少，在众多局部指标中表现相对较好，但是由于信息不足，它的精度赶不上基于全局信息的测度。

1. Katz指标

Katz指标是将节点对的所有路径的作用都作为了链路预测的参考因素，当然，它会设置一个参数，控制了不同路径在预测中所能引起的影响程度，较短的路径具有的影响权重较大，接下来影响权重按路径长度指数递减。数学表达式如下：

(2-20)

其中pathsxy<l>是长度为l的所有连接x和y的路径的集合，βl是控制不同的路径权重的一个自由参数。很容易看出，βl非常小的路径产生了接近于CN的度量，而当路径很长时，βl会变得越来越小，路径产生的贡献也会变小。S矩阵也可以写成（I-βA）-1-I。注意β必须低于矩阵A的特征值最大值的倒数，因为这样才能保证方程的收敛性。

1. 局部路径（LP）指标

该指标是本文的重点研究指标，是一种准局部性相似指标，可以认为它是CN的扩展版。同样是将节点邻居作为链路预测的重要参考因素，但与CN指标只考虑二阶邻居不同，它考虑到了三阶邻居的作用，它的定义如下：

(2-21)

其中，S表示相似矩阵，是一个可以进行设置的参数。从它的定义式我们可以很清楚的看到，如果将设置为0，那么它将和CN指标完全相同。A3xy的含义与上面的介绍相同，表示节点x与节点y长度为3的相连路径数目。

选择这三个指标进行比较，是因为它们都可以归类为路径依赖的相似性，都可以统一形式为。其中对于CN，l=2；而对于LP,l=2,3;对于Katz，l=1,2,3，...，∞。因为本文只对间接连接的节点对感兴趣，所以Katz指数可以被视为考虑l=2，3，...，∞。

### 2.3.3 性能评价指标

在链路预测的算法研究中，一般采用AUC来作为算法的评价指标，本文也不例外。AUC是一种从整体上衡量算法精确度的指标，可以这样简单理解：先从测试集中选一条边，这条边的选取方式是随机的，将这条边的相似性分数值记录下来（这个链接可能出现的概率）为a，然后再随机选取一条不存在的边，记这条边的分数值为b，那么AUC就是a>b的概率。具体的计算规则是这样的：通过前面的数据处理，我们会获得两个集合：测试集和不存在的边的集合。进行计算的时候，每次都会从上述两个集合中各随机的选取一条边，并且记录下它们的分数值，为了方便叙述，在这里假设从测试集获取到的边的分数值为x，从不存在的边中获取到的边的分数值为y。此时主要记录出现的两种情况：1.如果x>y，那么加1分；2. x=y就加0.5分。跟具这样的计算规则独立的进行n次比较，如果第一种情况出现了n′次，第二种情况出现了n′′次，则AUC的定义就为：

(2-22)

从上面是式子我们可以知道，如果可以进行的比较次数足够多，则n足够大，并且节点之间的相似性分数都是随机产生的，那么n'趋向于0，n''趋向于n，则AUC=0.5。因此AUC值大于0.5的程度衡量了算法的预测性能，超出越多，也

意味着算法的精确度也越好。

# 第三章 实验结果及分析

## 3.1实验设置及实验环境说明

在网络上查找到的复杂网络数据一般具有节点数量多，节点之间连接复杂的特点，为了更好的理解CN、LP、Kazt指标之间的关系，对它们的精确的关系有个简单的了解，我利用python的networkx进行了网络构建，并通过另一个工具包matplotlib将构建的网络进行了可视化。

networkx是python中的一个工具包，在复杂网络的分析上有着非常优秀的支持。networkx本身具备了生成网络的功能，通过调用其中的方法，并设置我们想要的参数，就能很方便的生成我们想要的网络模型，这对于网络算法的分析研究工作提供了非常大的帮助。随着研究的进展，学者们发现现实中完全规则的网络或者是完全随机的网络是不存在的，真实系统中形成的网络大都是无标度网络与小世界网络的混合体，具有与这两个模型相似的统计特性。因此我利用代码构建出无标度网络模型[49]和小世界网络模型[50]，小世界网络中的随机重置链接概率固定设置为0.1，并在两个模型中生成两个节点数量固定为500的网络，通过改变两个网络中每个节点邻居的个数来对CN、LP以及Katz指标进行对比研究。以下是构建网络的部分代码：

import networkx as nx

import matplotlib.pyplot as plt

import random

G=nx.random\_graphs.barabasi\_albert\_graph(N,K)

with open() as f:

    for e in G.edges():

        f.write(str(e[0]) + " " + str(e[1]) + "\n")

nx.draw(G)

plt.draw()

plt.show()

将构建出来的网络保存为文本形式，然后再通过三种指标的算法进行计算，

得出各自的AUC值，图3.1，是两个网络模型各个阶段的部分网络拓扑图，其中n是节点个数，k是网络中每个节点的邻居个数：

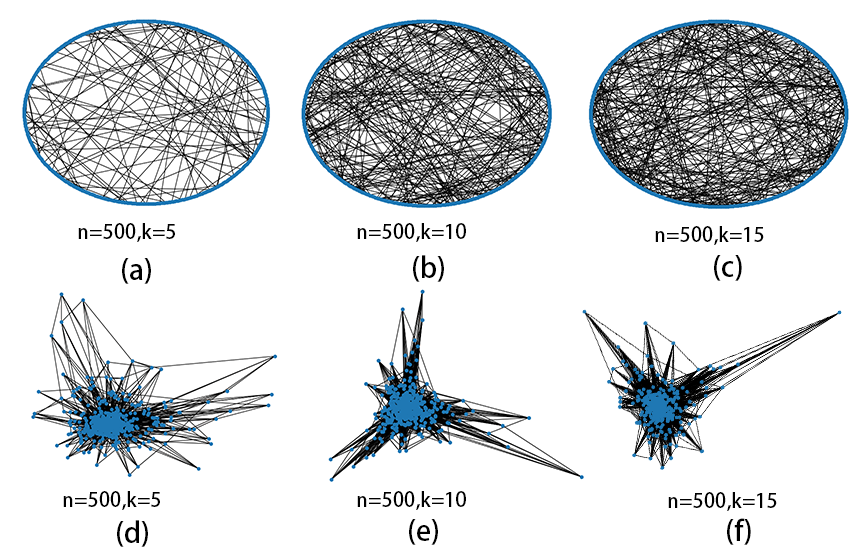


图3.1 网络不同阶段拓

其中(a)、(b)、(c)是小世界网络，(d)、（e）、（f）是无标度网络。将每个网络中节点的邻居个数从5个依次增加到40个，并将网络的数据保存下来，通过三种指标的算法算出各自的AUC。

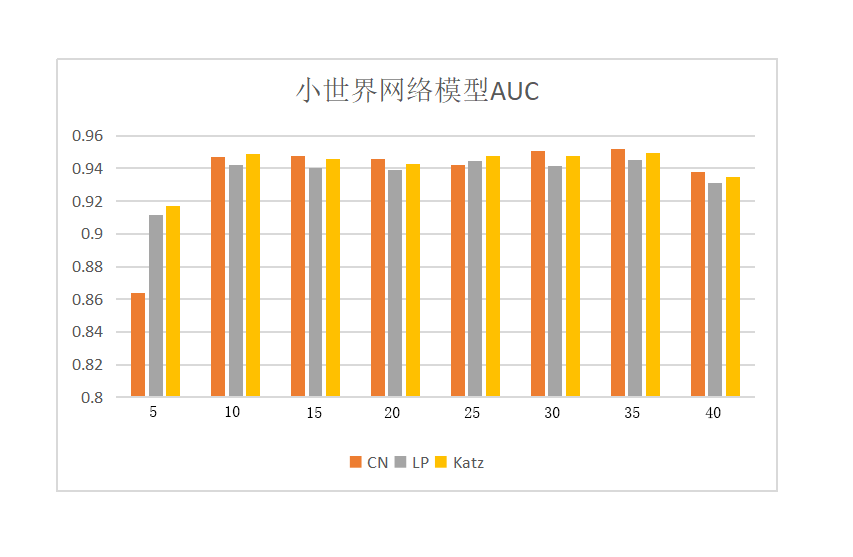


表3.1小世界网络模型AUC

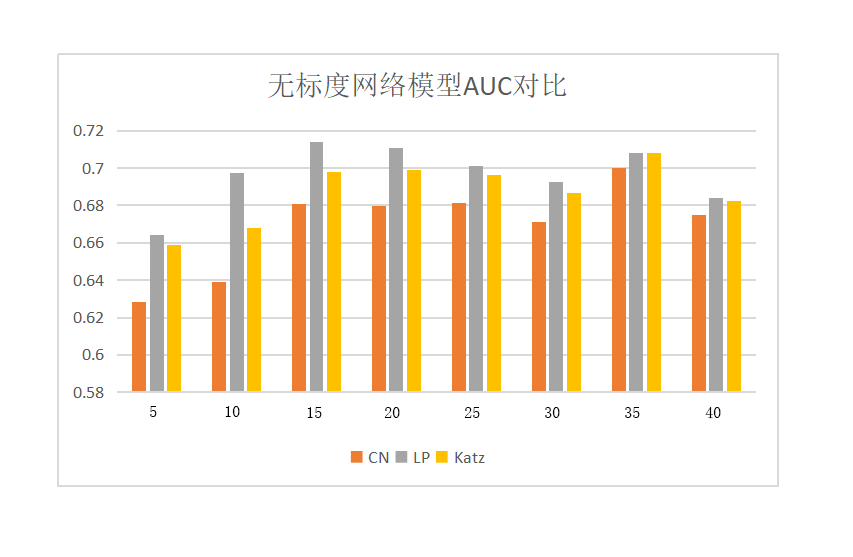


表3.2无标度网络模型AUC

在小世界网络中，节点的度比较均匀，网络中的平均路径也比较短，网络结构也较为简单。可以看出，三种预测指标都取得了较好的预测效果，几乎都能达到0.9以上，如表3.1所示。而在无标度网络里，虽然三种指标的预测精度都有所下降，但可以明显的看到LP指标的表现则最为突出，如图表3.2所示。无标度的网络中节点的度呈幂律分布，也就是说网络中有一小部分节点拥有着很多的连接，而大部分节点却很少。也就意味着有一小部分的节点是很多节点的共同邻居，而实际这些节点并不存在连接关系。CN指标用于只考虑了二阶的共同邻居，所获得的信息有限，因此在这种模型的网络中预测效果不是很好，而LP指标通过引入了下一阶邻居的作用，极大的提高了算法的预测精度。

接下来，我们讨论三个相似性指数的计算时间复杂性。在分析某种算法的复杂度时，一般我们使用大O表示法来描述算法复杂的上升趋势。假设在一个网络中，节点的总数量是N，节点平均度是<k>。当我们需要遍历网络中的所有节点时，时间复杂度为O(N),而计算一个节点的邻居的时间复杂度为O(k)。CN指标只需要考虑到二阶邻居，因此时间复杂度为O(NK2)；LP指标是在CN指标的基础上将三阶邻居对链路预测的作用也考虑了进来，因此总的时间复杂度为O(NK3)；而Katz指标考虑的则是网络中所有的路径，一般来说，网络中的节点个数要远远大于网络中节点的度，因此此时总的时间复杂度就会变成了O(N3)。综上，我们可以知道在算法预测准确性上，LP指标相比于CN指标有较为明显的优势，与Kazt指标也同样显现出很强的竞争力，并且在时间复杂度上比Katz指标要低得多。

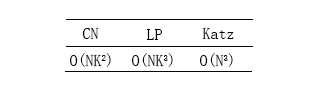


表3.3 各算法时间复杂度

## 3.2实验数据集

在上一个小节中，我们通过自己构建两种网络模型，对三种指标的精度进行了分析，也得到了一些初步的结论。但现实中存在的复杂网络一般非常庞大，具

有节点数量多，连边复杂等特点。为了能使我们得到的结论更加具有支撑性，在对周涛等人[51]所完成工作进行参考的基础上，我们选取了来自不同领域的六个代表性网络进行更深一步的实验，六个真实的网络数据集基本都可以从http://snap.stanford.edu/data/、 http://www-personal.umich.edu/~mejn/netdata/两个网站中下载得到:

1. Yeast[52]:蛋白质相互作用网络。含有2363个蛋白质和19021条边。
2. Karate[53]:描述一个美国大学里的俱乐部成员间关系的网络。是一个规模较小的网络，只有34个节点。
3. Grid：三级电网网络。这个网络包含了4941个节点，运行良好。
4. Metabolic[54]：描述的是线虫的新陈代谢网络，代谢物是网络中的节点。如果某一种代谢物可以在通过生物化学等反应生成另外一种代谢物，那么二者间存在连边。
5. INT:互联网的路由器级拓扑，由Rocketfuel Project收集。INT有4975个节点，连接良好，而且它是一个极度稀疏的网络，平均度只有2.49。
6. Football[55]:美式足球比赛的网络，每支球队对应于网络中的节点，两支球队之间比过赛则产生连边。表3.4总结了这些网络的巨型组件的基本拓扑特征。

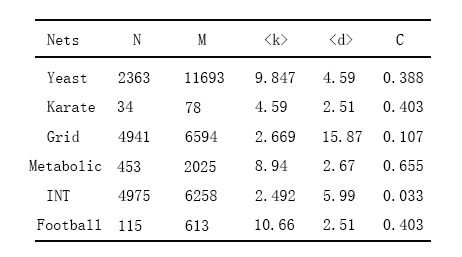


表3.4 六个网络基本拓扑特征

## 3.3结果与分析

我们将链路预测算法应用于六个真实网络，其精度如表3.5所示。显然，LP指标的表现总是优于CN指标，尤其是对于INT网络，AUC从0.663大幅提高到0.966。除了电力网格网络（Grid），LP指数给出的预测和Katz指数一样具有竞争力。

表3.5 三种相似性指标在六个网络中的AUC值

电力网格网络（Grid）是一个强局部化的网络，大多数链路的地理长度较短，因此，网格的拓扑距离<d>=15.87比其他五个示例网络大得多。虽然电力网格网络在地理上是局部化的，但是聚类系数相对较小，并且缺少短循环，因为从工程角度来看，这种循环是冗余的并且效率很低。实际上，在电力网格网络（Grid）中，当链接被移除时，通常很难找到非常短的路径，例如连接两个端点的长度为2或3的路径。因此，如果链路被移除，仅考虑非常短的路径，CN和LP指数不能细化两个直接连接的节点之间的相关性。因此预测精度相比于Katz指标来说较为不理想。

结合前面小节中分析到的三种指标的时间复杂度，显然，CN指标的计算时间成本是最低的。但需要注意的是，计算LP指标的计算复杂度对节点的平均度

非常敏感，而计算Katz指标的计算复杂度对网络大小非常敏感。因此，与采用Kazt指标的算法相比，采用LP的算法在大规模稀疏网络中具有很大的优势。由于对计算复杂性的真正挑战总是与大规模的真实网络有关，这些网络大多非常稀疏，所以LP指标比Kazt指标更加实用。

# 第四章 结论与展望

在本文中，我们通过将LP指标与常见的CN指标，Kazt指标进行比较分析，发现LP指标比CN指标提供的预测略准确，尤其是在高噪声情况下。我们进一步使用六个有代表性的真实网络来测试三个相似性指数，证明了LP指标可以提供与Katz指标相当的预测精度。并且与Katz指标相比，LP指标需要更少的CPU时间和内存空间，因此更实用。忽略节点度相关性，计算LP指标和Katz指标的时间复杂度分别为O(NK3)和O(N3)。因此，对于巨大的并且稀疏（即非常小的平均度k）的网络，LP指标的优势是惊人的。

高度准确的预测在实践中非常重要。比如许多生物网络，如蛋白质-蛋白质的相互作用网络、代谢网络以及食物网，在实验室或野外发现链接的成本很高。如果预测足够准确，而不是盲目地检查所有可能的相互作用，以预先预测已知的相互作用，并专注于最有可能存在的联系，可以大幅降低实验成本。对于其他一些，如网络社会中的友谊网络，很有可能向相关用户推荐尚未存在的链接，作为有希望的友谊的推荐。腾讯QQ中可能认识的人这一功能就是一个典型的应用。除了实际意义之外，值得强调的是，链路预测的研究还可以提供一些关于结构组织的理论见解。例如，在本文中，关于电力网格网络的意外结果给出了一些不明显的特定结构特性的证据。

本文仅仅只考虑静态网络中的链路预测问题。但是很多真实的网络都是一直在进化的，不同时间创建的链接原则上应该赋予不同的权重。许多时序链路预测的方法也相继出现，第一类是基于张量分解和矩阵实现的方法，Dunlavy [56]等人提出了两种分别基于张量以及基于矩阵的二部图的时序链路预测方法。他们通过将Kazt方法进行扩展，获取到二部图，并利用截断的SVD来设计一个可伸缩的方法来计算“截断了的”Kazt得分。他们还对数据的张量分解进行分析，证明了时间链接数据的自然三维结构的有用性。第二类是基于时间序列模型的方法，Chen[57]等人同时考虑了网络动态演化和链接之间的时间相关性，提出了一种时间序列模型和静态图表示下的混合算法，并通过实验证明，相比于六种常用的静态

图链路预测算法，该算法具有更好的性能。第三类是基于机器学习的方法，在文章[58]中提出了一种没有监督的预测方法。他们把时间事件定义为从前一帧到后一帧的两个节点间的特定活动。这个方法把主要事件相关的奖励与和次要事件相关的奖励进行了关联。真实的网络都是动态演变的，如果想要更好的掌握网络的演化机制，未来关于动态网络中的链路预测将是一个非常值得重视的方向。

# 致谢

时光匆匆而过，转眼之间大学四年已经临近结束了。回望当初初步入大学之时，是我第一次离开从小生长到大的小县城。一个人提着行李箱，跨越一千二百六十多公里的距离，独自面对这个完全陌生的环境，心中充满着紧张与期待。四年的时间说长不长，说短也不短，但毫无疑问，这四年是我人生中一个非常重要的部分，在这期间发生的点点滴滴都对我产生了很大的影响，它让我成长，让我活得智慧，变得更加的坚强。

首先感谢我的指导老师-张睿挺老师，在最开始进行论文课题内容研究时，面对其中难懂的理论公式时，我一时不知从何下手，是张老师给予我帮助才使我的工作逐渐取得进展，并且在后面撰写论文的过程中，张老师也给我提出了很多宝贵的建议，还帮我查找最新的相关资料，与我一同研究、分析。正因如此，我的研究课题才如此顺利的完成。在这里真诚的对张老师说一声：谢谢！

此外，我还想感谢我亲爱的母校以及我的室友们，感谢学校给我提供了这么好的平台，让我在这四年里拥有了幸福而又充实的时光。也非常感谢室友们在这四年里对我的照顾和帮助。让我在离家千里之外的城市里能感受到另一个“家”的温暖。

最后，我还要感谢我亲爱的家人们，谢谢你们一直以来都无条件的支持我，包容我。感谢父母，含辛茹苦的抚养和不辞辛苦的培养，你们的照顾和爱是我不断前进的动力。

再一次向各位表示衷心的感谢，也非常感谢各位专家和教授能在百忙之中抽空来审阅我的论文，非常感谢!

# 参考文献

1. Lü L Y, Zhou T. Link prediction in complex networks: A survey [J]. Physica A Statistical Mechanics & Its Applications, 2010, 390(6):1150-1170.
2. 张斌, 马费成.科学知识网络中的链路预测研究述评[J]. 中国图书馆学报, 2015, 41(3): 99-113
3. Lei C, Ruan J. A novel link prediction algorithm for reconstructing protein–protein interaction networks by topological similarity [J]. Bioinformatics, 2012, 29(3): 355-364.
4. Boucher B, Jenna S. Genetic interaction networks: better understand to better predict [J]. Frontiers in genetics, 2013, 4: 290.
5. Röttger R, Rückert U, Taubert J, et al. How little do we actually know? On the size of gene regulatory networks [J]. IEEE/ACM transactions on computational biology and bioinformatics, 2012, 9(5): 1293-1300
6. Pan L, Zhou T, Lü L Y, et al. Predicting missing links and identifying spurious links via likelihood analysis [J]. Scientific reports, 2016, 6: 22955
7. Al-Halah Z, Tapaswi M, Stiefelhagen R. Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning[C].Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 5975-5984.
8. Kleinberg J. Analysis of large-scale social and information networks [J].Philosophical transactions.Series A, Mathematical,physical,and engineering sciences.2013, 371(1987):20120378.
9. Liao H, Zeng A. Reconstructing propagation networks with temporal similarity [J]. Scientific reports, 2015, 5: 11404.
10. Das D. Positive and Negative Link Prediction Algorithm Based on Sentiment Analysis in Large Social Networks [J]. Wireless Personal Communications, 2018(2):1-16.
11. Chen B, Chen L, Li B. A fast algorithm for predicting links to nodes of interest [J]. Information Sciences, 2016, 2016(329):552-567.
12. Newman, M.E. Clustering and preferential attachment in growing networks

[J].Physical Review E. 2001, 64(2):025102.

1. Xie Y B, Zhou T, Wang B H. Scale-free networks without growth [J]. Physica A: Statistical Mechanics and its Applications. 2008, 387(7):1683-1688.
2. Adamic L A, Adar E. Friends and neighbors on the web [J]. Social Networks. 2003, 25(3):211-230.
3. Jaccard P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura [J]. Bulletin De La Societe Vaudoise Des Sciences Naturelles. 1901, 37:547-579.
4. Salton G,McGill M J.Introduction to modern information retrieval [J]. McGraw-Hill. 1986, 41(4):305-306.
5. Zhou T, Lü L Y, Zhang Y.C. Predicting missing links via local information [J]. The European Physical Journal B. 2009, 71(4):623-630.
6. Everett M G, Borgatti S P. Regular equivalence: General theory [J]. Journal of mathematical sociology. 1994, 19(1):29-52
7. Katz L. A new status index derived from sociometric analysis [J]. Psychometrika. 1953, 18(1):39-43.
8. Brin S, Page L. Reprint of: The anatomy of a large-scale hypertextual web search engine [J].
9. Klein D J, Randić M. Resistance distance [J]. Journal of mathematical chemistry. 1993, 12(1):81-95.
10. Jeh G, Widom J. SimRank: a measure of structural-context similarity[C].Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002: 538-543.
11. Liu W, Lü L Y. Link prediction based on local random walk [J].Europhysics Letters. 2010, 89(5):58007.
12. Jia Y, Qu L.Improve the Performance of link prediction methods in citation network by using H-Index[C].2016 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC). IEEE, 2016: 220-223.
13. Sørensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons [J].

Biol. Skr. 1948, 5:1-34.

1. Salton G, McGill M J. Introduction to modern information retrieval [J]. McGraw-Hill. 1986, 41(4):305-306.
2. Vittorio C C, Gregorio A L,Timothy R.From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks [J]. Sci Rep, 2013, 3(4):1613.
3. Zhou W, Jia Y. Predicting links based on knowledge dissemination in complex network [J]. Physica A: Statistical Mechanics and its Applications. 2017, 471:561-568.
4. Lü L Y, Zhou T, Zhang Q M, et al. The H-index of a network node and its relation to degree and coreness [J]. Nature communications. 2016, 7:10168.
5. 高杨, 张燕平, 钱付兰,等.基于三元闭包的节点相似性链路预测算法[J].计算机科学与探索, 2017, 11(5):822-832.
6. Liu S X, Ji X S, Liu C X, et al. Extended resource allocation index for link prediction of complex network[J].Physica A Statistical Mechanics & Its Ap-plications,2017, 2017(479):174–183.
7. Sun Q, Hu R, Yang Z, et al. An improved link prediction algorithm based on degrees and similarities of nodes[C]. 16th International Conference on Computer and Information Science. IEEE, 2017: 13-18.
8. Wu Z H, Lin Y F, Wang J, et al. Link prediction with node clustering coefficient[J]. Physica A Statistical Mechanics & Its Applications, 2016, 2016(452):1-8．
9. Wu Z H, Lin Y F, Wan H Y, et al. Predicting top-L missing links with node and link clustering infor-mation in large-scale networks[J]. Journal of Statistical Mechanics Theory & Experiment, 2016, 2016(8):083202.
10. 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.
11. 吕琳媛,周涛. 网络链路预测：概念与前沿[J].中国计算机学会通讯, 2016, 12(4) : 12 – 21.
12. Cai D, He X F, Han J W, et al. Graph Regularized Nonnegative Matrix Factorization for Data Representation [J]. IEEE Trans Pattern Anal Mach Intell, 2011, 33(8):1548-1560.
13. Jiao P F, Cai F, Feng Y D, et al.Link predication based on matrix factorization by fusion of

multi class organizations of the network [J]. Scientific Reports, 2017, 7(1):8937.

1. Zhao Y, Li S, Zhao C, et al.Link prediction via a neighborhood-based nonnegative matrix factorization model[C].The Proceedings of the Third International Conference on Communications, Signal Processing, and Systems. Springer, Cham, 2015: 603-611.
2. Zhang C.J, Zeng A. Prediction of missing links and reconstruction of complex networks [J]. International Journal of Modern Physics C. 2016, 27(10):1650120.
3. Zeng A, Cimini G. Removing spurious interactions in complex networks [J]. Physical Review E. 2012, 85(3):036101.
4. Lü L Y, Pan L M, Zhou T, et al. Toward link predictability of complex networks [J]. Proceedings of the National Academy of Sciences of the United States of America,2015, 112(8):2325-2330.
5. Zhang P, Wang X, Wang F, et al. Measuring the robustness of link prediction algorithms under noisy environment [J]. Scientific Reports. 2016, 6:18881.
6. Ouyang B, Jiang L, Teng Z. A Noise-Filtering Method for Link Prediction in Complex Networks [J]. PloS one. 2016, 11(1):e0146925.
7. 姚亚兵. 基于复杂网络拓扑结构的链路预测方法研究[D]. 兰州大学, 2017.
8. Barabási A L. Network science [M]. Cambridge university press, 2016.
9. Battiston F, Nicosia V, Latora V. Structural measures for multiplex networks [J]. Physical Review E, 2014, 89(3): 032804.
10. M. E. J. Newman, Phys. Rev. E 64, (2001),025102(R).
11. Barabási A L, Albert R. Emergence of scaling in random networks [J]. Science,1999,286(5439): 509-512.
12. Watts D J, Strogatz S H.Collective dynamics of “small-world” networks [J].Nature.1998, 393(6684):440-442.
13. Lü L, Jin CH, Zhou T. Similarity index based on local paths for link prediction of complex networks. Phys Rev E Stat Nonlin Soft Matter Phys. 2009 Oct;80(4 Pt 2):046122
14. Von M C, Krause R, Snel B, et al. Comparative assessment of large-scale data sets of protein-protein interactions[J]. Nature, 2002, 417(6887):399.
15. Porter M A, Onnela J P, Mucha P J. Communities in networks [J]. Notices of the AMS, 2009,

56(9): 1082-1097.

1. Duch J, Arenas A.Community detection in complex networks using extremal optimization [J]. Physical review E. 2005, 72(2):027104.
2. Girvan M, Newman M E. Community structure in social and biological networks [J]. Proceedings of the National Academy of Sciences of the United States of America, 2002, 99(12):7821.
3. Dunlavy D M, Kolda T G, Acar E. Temporal link prediction using matrix and tensor factorizations[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2011, 5(2): 1-27.
4. Chen H, Li X, Huang Z. Link prediction approach to collaborative filtering[C]Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'05). IEEE, 2005: 141-142.
5. Soares P R S,PrudêNcio R B C.Proximity measures for link prediction based on temporal events[J]. Expert Systems with Applications, 2013, 40(16): 6652-6660.