兰州大学

毕 业 论 文

（本科生）

中文标题 基于链路预测的国际原油贸易潜在关系分析与预测

英文标题 Analysis and Prediction of Potential Trade Relation in International Crude Oil Trade based on Link Prediction

学生姓名 王杰

指导教师 李龙杰、工程师

学 院 兰州大学信息科学与工程学院

专 业 计算机科学与技术（基础理论班）

年 级 2015 级

兰州大学教务处

诚信责任书

本人郑重声明：本人所呈交的毕业论文（设计），是在导师的指导下独立进行研究所取得的成果。毕业论文（设计）中凡引用他人已经发表或未发表的成果、数据、观点等，均已明确注明出处。除文中已经注明引用的内容外，不包含任何其他个人或集体已经发表或在网上发表的论文。

特此声明。

论文作者签名：_____ 日　　期：_____

基于链路预测的国际原油贸易潜在关系分析与预测

摘 要

原油是重要的工业原料,原油和原油产品对各国的经济发展起着重要的作用，并且在国际贸易中被大量交易。由于各国的原油矿藏和消费水平的差异，大部分国家需要通过国际原油贸易来丰富自己的原油储量以保证工业的顺利发展。因此，评估潜在的贸易联系、研究原油贸易演变机制并定量分析国际原油的贸易格局对于挖掘国际原油贸易数据隐含信息，合理规避国际原油贸易风险有着非常重要的作用。本文研究中，以年为单位，将大量的国际原油贸易数据抽象为以国家为节点、原油贸易活动为连边的复杂网络图，并使用基于节点相似度的链路预测方法来进行复杂网络分析。基于在复杂网络中评估潜在连边并分析其产生的原因和推测网络演变机制，本文使用了基于局部网络结构的多个节点相似度度量方法来评估节点对之间潜在的贸易联系，然后使用 Ranking Score 量化评估了不同相似度度量方法的准确度。再基于 Ranking Score 值区别不同因素对于国际原油贸易网络演变的影响力大小以推断国际原油贸易网络的演化机制。另外，为了进一步提高相似度度量方法的准确度，本文结合了时间近因改进原有的节点相似度度量方法来对国际原油贸易网络的潜在网络进行分析预测。实验结果表明：(1) 共同贸易伙伴是国际原油贸易联系的有力保证。(2) 大部分国家拥有的贸易伙伴数量相近。(3) 国际原油贸易市场稳定，国家倾向于保持已有贸易联系。

关键字：原油，演变，链路预测，相似度，时间。

ANALYSIS AND PREDICTION OF POTENTIAL TRADE RELATIONS IN INTERNATIONAL CRUDE OIL TRADE BASED ON LINK PREDICTION

Abstract

Crude oil is an important industrial raw material. Crude oil and crude oil products play an important role in the economic development of various countries and are traded in large quantities in international trade. Due to the differences in crude oil deposits and consumption levels in various countries, most countries need to enrich their crude oil reserves through international crude oil trade to ensure the development of industry. Therefore, assessing potential trade links, studying the evolution mechanism of crude oil trade and quantitatively analyzing the trade pattern of international crude oil plays an important role in mining the implicit information of international crude oil trade data and avoiding the risk of international crude oil trade. In this study, the large-scale international crude oil trade data is abstracted into a complex network map with the state as the node and the crude oil trade activity as the edge, and the link prediction method based on the node similarity is used for the complex network analysis. Based on the evaluation of potential edges in complex networks and analysis of their causes and speculative network evolution mechanisms, this paper uses multiple node similarity measures based on local network structure to evaluate potential trade links between node pairs, and then use Ranking Score quantitatively evaluates the accuracy of different similarity measures. Based on the Ranking Score, the influence of different factors on the evolution of the international crude oil trading network is inferred to infer the evolution mechanism of the international crude oil trading network. In addition, in order to further improve the accuracy of the similarity measure method, this paper combines the time factor to improve the original node similarity measure method to analyze and predict the potential network of the international crude oil trade network. The experimental results show that: (1) Common trading partners are a strong guarantee for international crude oil trade links. (2) The number of trading partners owned by most countries is similar. (3) The international crude oil trade market is stable and the country tends to maintain existing trade links.

Keywords: crude oil, evolution, link prediction, similarity, time.

目 录

第一章 绪论

1.1 研究背景及意义

原油是一种重要的不可再生的矿石资源。第二次工业革命后，人类进入电气时代，对于石油的深度利用也使得石油成为现代工业的重要基础之一。原油自地壳开采出来后，经过冶炼可以制成多种衍生品，有汽油等一次性能源，也有诸多化学原料（如乙烃、苯、甲苯、二甲苯），这些原油产品可以作为三大合成材料（合成橡胶、合成纤维、合成塑料）的基本原料，是医药、农药、炸药的重要原料。原油在没有新的替代品出现之前，作为现代工业血液的原油，其意义和价值重大，是各个国家的工业命脉之一。并且，在人们的日常生活中，对于原油的利用也深入现代人生活的方方面面。无论衣、食、住、行都离不开对于原油的利用。合成纤维、液化石油气、化肥、农药、塑料、汽油、润滑油、沥青等等都是石油工业的产品。

不同的国家拥有的原油资源并不均衡，有些国家原油资源丰富，而有些国家原油资源贫乏且消耗大。因此，对于原油资源的竞争，常常成为国际矛盾纠纷的重要因素之一，甚至是战争的导火索。就我国而言，我国是世界人口第一大国，也是世界能源消费第一大国。人民生活水平的提高，工业的发展和能源转型，都促使我国的原油消费水平也在不断提高。当今时代，和平与发展是主题，因此国际原油贸易作为当今最重要的原油资源的再分配的方式，一直备受多方关注。

原油是国际上交易范围最大，交易量最多的货物之一。各国之间通过国际原油贸易，保证自己的原油储量以保障自己的资源安全和能源安全。国际原油贸易数据显示原油资源丰富的国家常与原油资源贫乏的国家之间产生贸易联系，而原油资源贫乏的国家常同时与原油资源丰富的国家和原油资源贫乏的国家之间产生贸易联系以扩大其原油贸易关系。然而，国家之间还存在着现有的数据并不能直接说明的潜在贸易联系[1]，评估这些潜在的贸易联系对探索国家间原油贸易关系非常重要[2]。国际原油贸易关系中，部分国家联系松散没有明显的贸易联系，但是国家间的贸易模式和其他国家的贸易联系可能会促使这些国家间在未来产生贸易联系。并且对于各国政府来说，评估潜在贸易联系也有重要的意义。政府部门需要提高自己对国际原油贸易的了解，以评估国际原油贸易趋势，合理规避原油贸易风险[3]。

1.2 相关工作

国际原油贸易中，各个国家的原油贸易数据可以形成一个以国家为节点，以贸易活动为边，以贸易量为权重的动态复杂网络。在这种全球性的国际原油贸易网络中，复杂网络分析是一种合适的国际贸易分析方法[4]。例如，Figiolo 等人通过复杂网络分析研究了世界贸易网络的演变[5]；Duan 等人用复杂网络分析研究了世界贸易网络的演化模型[6]。

以往的研究中，对于国际原油贸易网络的研究大多集中在国际原油贸易的格局和演变[7]。Zhong 等人利用加权和非加权的网络结构研究了国际原油贸易网络中社区的演变[8]。Yang 等人发现国际原油贸易网络存在“小世界现象”[9]。Zhang 等人[10]和 An 等人[1]研究了国际原油贸易网络中的国际竞争关系。Helpman 等人[2]和 Foschi 等人[11]使用国家属性（国内生产总值 GDP）来推断贸易国家之间的已知联系的贸易流量。这些研究的重点在于已知的贸易联系或者是明显的贸易联系，忽视了国家之间潜在的贸易联系。然而贸易网络的波动变化影响着国际贸易关系，使得贸易网络中形成新的联系或者旧的联系消失。

为了分析预测国际原油贸易网络中的潜在联系，Guan 等人基于链路预测方法，利用公共邻居 （Common Neighbors, CN）指标预测潜在的贸易联系，为全球贸易关系的研究提供了一个新的视角[12]。但是，他们只从公共邻居的角度来预测分析国际原油贸易网路中的潜在联系，没有考虑其他方面的因素。

国际原油贸易网络是一个动态的复杂网络，对于动态复杂网络的分析仅局限于网络结构的分析中是不够的。国际原油贸易网络的贸易联系不仅与网络的拓扑结构相关，贸易联系还会随着时间的变化而产生波动。

基于此，本文在对国际原油贸易演化机制的研究中，结合了多个节点结构属性和时间属性对于国际原油贸易的影响，进一步的深入研究了时间因素在国际原油贸易网络的演化机制中的作用，并提供了新的角度来研究贸易网络的演变机制。

1.3 论文组织结构

本文首先概述了研究的目的和意义以及相关工作。接下来的部分组织如下：第二章介绍了链路预测方法和一些常用的相似度方法，并且论述了时间近因和基于近因的相似度改进；第三章介绍了衡量相似度方法准确度的三个指标（AUC、Precision、Ranking Score）；第四章介绍了数据源和实验过程，并对实验结果进行分析；最后第五章对实验结果进行总结。

第二章 链路预测的方法

2.1 链路预测

自然界和现实生活中存在着大量的复杂系统可以抽象为复杂网络，如交通系统、电力系统、社交系统等。这样的网络由大量的节点以及节点与节点之间的连边组成，不同于规则网络和随机网络，有着其独特的结构和特征。

Liben-Nowell 和 Kleinberg 对链路预测的定义为：链路预测是为了准确的预测在网络 G 中，从指定的 t_1 时刻到目标 t_2 时刻内被添加到网络 G 中的边[13]。

链路预测是复杂网络研究中的一个重要方法，其目标是预测网络中节点对之间可能存在的连边，可用于推断分析复杂网络的结构和演化机制，其研究有重要的理论意义和实际应用价值[14]。链路预测方法量化计算预测的准确度，在网络演化机制的研究中，相较于直接建立演化模型，可以更为明确地分辨不同因素在网络演化上重要性，从而对比分析出驱动网络演化模型的重要因素[15]。链路预测的发展推动了网络演化模型研究的发展。

链路预测方法通常使用网络的结构特征来预测节点间存在的边[16]。基于节点相似度的链路预测方法是链路预测研究中的主流方法[15]。基于节点相似度的链路预测方法认为节点对的相似度越高，节点对之间有潜在联系的可能性就越大[14]，其核心是节点对相似度的定义[17]。

基于节点相似度的链路预测方法包含以下几步。

（1） 抽象数据集，将大量真实的复杂系统进行拓扑抽象，以节点来描述真实系统中的不同个体，以节点属性来代表真实系统中的个体属性，以边代表复杂系统中个体与个体之间的关系，以边的属性权重来描述个体之间联系的强弱。

（2） 定义节点相似度。用网络中节点的特征或网络中的结构特征来定义复杂网络中两个不同节点相似度，表示两个暂时没有联系的节点存在潜在联系或在未来出现联系的可能性。

（3） 对于已经抽象好的网络 G，将网络 G 中所有已经存在的边作为一个数据集 E，将网络中所有的节点所组成的完全图的所有连边作为一个数据集 U，将网络 G 中所有节点对之间不存在的连边作为一个数据集 E^N。为检测算法或者相似度在网络 G 中应用的准确性，在 E 中以一定比例随机选取一定数量的边作为测试集数据 E^P，剩下的边作为训练集数据 E^T。以节点集 V 和边集 E^T 构建一个新的网络 G^N，在 G^N 中计算 E^N 和 E^P 所表示的节点对之间的相似度，并将相似度

和节点对一起以相似度为度量由大到小排序，得到一个相似度序列 X。理论分析中认为，在相似度排名序列中，排名越靠前的节点对存在潜在联系的可能性越高。

（4） 对于 3 中所取得的测试集 E^P 和相似度序列 X，计算算法的精确度。

2.2 常用相似度度量方法

节点相似度的定义是链路预测中的核心问题。不同的相似度定义对应着复杂网络不同的结构特征。常见的有基于节点属性的定义[18]和基于网络拓扑结构的相似度定义[13]。节点属性定义中认为两个节点的共同特征越多，节点相似性就越高。但是在很多现实网络中，真实的节点信息获取困难，例如社交网络中的个人信息。与此相比，基于网络拓扑结构的相似度定义就更为可靠。

2.2.1 CN 方法

CN（Common Neighbors）是链路预测中对于节点对相似度最直接的定义。CN 以共同邻居的数量作为度量指标来计算节点间的相似度[19]，在 CN 的定义中两个节点的共同邻居数量越大，则节点对之间产生连边的可能性越大。

对于节点对(x, y)，以S_{xy}来表示节点对(x, y)的相似度，$C(X)$表示 x 的邻居节点集合，那么S_{xy}^{CN}可以表示为：

$$S_{xy}^{CN} = |C(X)| \cap |C(Y)| \qquad (2.1)$$

2.2.2 JC 方法

JC（Jaccard's coefficient）系数是信息检索中常用的相似度度量指标[13]。

JC 定义节点对(x, y)的相似度为 x，y 的共同邻居的数量在 x，y 的所有邻居中所占的比例[13]，即：

$$S_{xy}^{JC} = |C(x) \cap C(y)|/C(x) \cup C(y) \qquad (2.2)$$

2.2.3 AA 方法

AA（Adamic/Adar）基于 CN 的定义考虑了共同邻居的度对于节点对相似度的影响。AA 中认为共同邻居中度较小的节点对于节点对之间产生连边有更大的贡献。Adamic 和 Adar 通过对共同特性进行加权求和来计算个体之间的相关性，AA 定义节点对(x, y)的相似度为共同邻居的度的对数的倒数求和[20]，即：

$$S_{xy}^{AA} = \sum_{z \in C(x) \cap C(y)} {}^{1}/_{log_2|K(z)|} \quad (2.3)$$

这里，K(z)表示 z 的度。

2.2.4 PA 方法

PA（Preferential Attachment）认为在网络拓扑图中新添加一条边连接到节点 x 的概率正比于节点 x 的度的大小。PA 指标在无标度网络预测准确性较好。PA 指标定义为节点对的度的大小的乘积[21]，即：

$$S_{xy}^{PA} = K(x) \times K(y) \quad (2.4)$$

2.2.5 RA 方法

RA（Resource allocation）的定义是从资源分配的角度提出的节点对的相似度定义[22]。RA 中假定两个不相连的节点对(x, y)之间有共同邻居集合 T，x 通过集合 T 中的节点分配传递资源到 y，y 接受到的资源就是节点对(x, y)的相似度。对于 T 中的任何一个节点 t，t 拥有从 x 传递过来的一份资源，假定 t 将它的资源平均的分配给它的每一个邻居，那么 y 从 t 可以获得的资源就是节点 t 的度的倒数，则 RA 的定义为节点对(x, y)的共同邻居的度的倒数求和，即：

$$S_{xy}^{RA} = \sum_{z \in C(x) \cap C(y)} {}^{1}/_{|K(z)|} \quad (2.5)$$

2.3 时间近因

常用相似度方法基于网络拓扑结构和节点属性，但是对动态网络，单从网络拓扑结构或者节点属性方面研究有所不足。结合时间属性和网络拓扑结构来计算相似性可进一步提高预测的准确度[23]。

在动态网络中，经常产生变化的节点和两个节点对之间经常产生的连边称为活跃节点和活跃边。

网络中活跃节点和活跃连边由时间属性定义[23]。基于网络拓扑结构，活跃节点和活跃边的定义可分为两类。第一类，一段时间内度频繁变化的节点和长时间保持不变的连边为活跃节点和活跃边；第二类，活跃节点定义为距离当前时间最近出现度的变化的节点，活跃连边定义为距离当前时间最近产生的连边[24]。本文在对于国际原油贸易网络的研究中使用第二类定义。

近因是 Potgieter 等人首先提出的表示链路预测的时间性的度量标准[24]。他们使用一个非常简单的概念来定义近因——1 加上节点距离上一次通信所经过的时间步长。

本文使用了 Saravanan Mohan 等人[25]在近因概念的基础上建立的两个近因方程：

$$Recency_{sender}(u) = \sqrt{\frac{a}{i+1}} \qquad (2.6)$$

$$Recency_{reciever}(v) = \sqrt{\frac{a}{j+1}} \qquad (2.7)$$

上述两个方程应用在通信网络上，u 和 v 分别是发送方和接收方。i 和 j 分别为 u 和 v 距离上一次通信所经历的时间步长。a 是常数，对近因加权。

在网络中，如果不考虑发送方和接收方在网络的角色，即不考虑考虑节点度的出度和入度的区别，则节点 u 的近因定义如下：

$$Recency(u) = \sqrt{\frac{a}{i+1}} \qquad (2.8)$$

i 为节点 u 的度距离最近一次变化的时间步长。从公式（2.8）可得，节点 u 的度产生变化的时间点与目标时间点越近，节点 u 的近因越强，节点 u 越活跃。

同理，边(u,v)的近因定义如下：

$$Recency(u,v) = \sqrt{\frac{a}{\mathrm{i}+1}} \qquad (2.9)$$

i 为边(u,v)距离最近一次产生所经过的时间步长。从公式（2.9）可得，节点 u，v 之间的连边保持时间越长，与目标时间点越接近，边(u,v)的近因就越强。所以节点 u，v 之间活动越频繁，节点 u，v 越可能在未来产生联系。

综合节点近因的定义和边近因的定义，当节点 u 的近因越弱，节点 u 与邻居之间连边保持的时间越长，其邻边就越活跃，邻边的近因就越强。对于一个没有新加节点的动态网络，网络中的节点频繁建立新的连接，节点近因越强，而当网络趋于稳定的时候，边近因就越强。例如贸易网络图中，市场扩大增长期的时候，节点近因就强；当市场稳定时，节点近因减弱，边近因增强。

在节点近因公式和在边近因公式中，时间步长的单位可以是秒、分、时、天或者年。在本文中，所采集的国际原油贸易数据以年为单位，所以在近因公式中所使用的步长差也以年为单位。

2.4 结合近因的节点相似度

时间是动态复杂网络的一个重要指标。使用相似度度量方法结合时间近因来进行链路预测可以获得更好的预测精度。Farshad 等人在社交网络中以公式（2.6）和公式（2.7）作为节点近因结合相似度方法[23]，但本文研究中不考虑国际原油贸易的贸易流向，将其替换成公式（2.8），得到相似度度量 TDCN、TDJC、TDAA、TDPA、TDRA，定义如下：

$$S_{xy}^{TDCN} = \sum_{z \in C(x) \cap C(y)} 1 + Recency(x) + Recency(y) \quad (2.10)$$

$$S_{xy}^{TDJC} = \sum_{z \in C(x) \cap C(y)} 1 + Recency(x) + Recency(y) \Big/ |C(x) \cup C(y)| \quad (2.11)$$

$$S_{xy}^{TDAA} = \sum_{z \in C(x) \cap C(y)} 1 + Recency(x) + Recency(y) \Big/ log_2|K(z)| \quad (2.12)$$

$$S_{xy}^{TDPA} = (K(x) + Recency(x) + 1) \times (K(y) + Recency(y) + 1) \quad (2.13)$$

$$S_{xy}^{TDRA} = \sum_{z \in C(x) \cap C(y)} 1 + Recency(x) + Recency(y) \Big/ K(z) \quad (2.14)$$

同理，结合边近因的节点相似度 TCN、TJC、TAA、TPA、TRA，定义如下：

$$S_{xy}^{TCN} = \sum_{z \in C(x) \cap C(y)} 1 + Recency(x, y) \quad (2.15)$$

$$S_{xy}^{TJC} = \sum_{z \in C(x) \cap C(y)} 1 + Recency(x, y) \Big/ |C(x) \cup C(y)| \quad (2.16)$$

$$S_{xy}^{TAA} = \sum_{z \in C(x) \cap C(y)} 1 + \mathrm{Recency(x, y)} \Big/ \log_2|\mathrm{K(z)}| \quad (2.17)$$

$$\mathrm{S_{xy}^{TPA}} = \mathrm{K(x)} \times \mathrm{K(y)} \times (\mathrm{Recency(x, y)} + 1) \quad (2.18)$$

$$\mathrm{S_{xy}^{TRA}} = \sum_{\mathrm{z \in C(x) \cap C(y)}} 1 + \mathrm{Recency(x, y)} \Big/ \mathrm{K(z)} \quad (2.19)$$

上述近因公式中增强时间近因对节点相似度的贡献，结合节点近因的相似度度量为 MTDCN、MTDJC、MTDAA、MTDPA、MTDRA，定义如下：

$$\mathrm{S_{xy}^{MTDCN}} = \sum_{\mathrm{z \in C(x) \cap C(y)}} \mathrm{Recency(x)} + \mathrm{Recency(y)} \quad (2.20)$$

$$S_{xy}^{MTDJC} = \sum_{z \in C(x) \cap C(y)} \frac{Recency(x) + Recency(y)}{|C(x) \cup C(y)|} \quad (2.21)$$

$$S_{xy}^{MTDAA} = \sum_{z \in C(x) \cap C(y)} \frac{Recency(x) + Recency(y)}{\log_2 |K(z)|} \quad (2.22)$$

$$S_{xy}^{MTDPA} = \big(K(x) + Recency(x)\big) \times \big(K(y) + Recency(y)\big) \quad (2.23)$$

$$S_{xy}^{MTDRA} = \sum_{z \in C(x) \cap C(y)} \frac{Recency(x) + Recency(y)}{K(z)} \quad (2.24)$$

同理，结合边近因的节点相似度 MTCN、MTJC、MTAA、MTPA、MTRA，定义如下：

$$S_{xy}^{MTCN} = \sum_{z \in C(x) \cap C(y)} Recency(x, y) \quad (2.25)$$

$$S_{xy}^{MTJC} = \sum_{z \in C(x) \cap C(y)} \frac{Recency(x, y)}{|C(x) \cup C(y)|} \quad (2.26)$$

$$S_{xy}^{MTAA} = \sum_{z \in C(x) \cap C(y)} \frac{Recency(x, y)}{\log_2 |K(z)|} \quad (2.27)$$

$$S_{xy}^{MTPA} = K(x) \times K(y) \times Recency(x, y) \quad (2.28)$$

$$S_{xy}^{MTRA} = \sum_{z \in C(x) \cap C(y)} \frac{Recency(x, y)}{K(z)} \quad (2.29)$$

第三章 链路预测方法评估指标

评估指标是为了计算预测方法的准确性。不同的评估指标，侧重点有所不同。常见的相似度评估指标有 AUC、Precision 和 Ranking Score。AUC 从测试集整体的相似度值来计算相似度度量方法的准确度，Ranking Score 从测试集相似度排序的顺序来计算预测准确度，而 Precision 用相似度序列的部分来评估相似度度量方法的精确度。本文研究中采用 Ranking Score 作为预测准确性的评估指标。

3.1 AUC

AUC 评估指标计算随机从测试集 E^P 抽取一条边比在不存在边集 E^N 中随机抽的边的相似度值大的概率[26]。

假设随机的分别从 E^P 和 E^N 中抽取边 e^p 和 e^n n 次。在 n 次抽取中，如果抽取到的边 e^p 的相似度分数值大于 e^n，就取值 1 分；若 e^p 分数值等于 e^n，取值 0.5 分；若 e^p 分数值小于 e^n，则取值 0 分。

若 n 次随机取样中，$S_e^p > S_e^n$ 的次数为 n_1 次，$S_e^p = S_e^n$ 的次数为 n_2 次，$S_e^p < S_e^n$ 的次数为 $n-n_1-n_2$ 次。则总分值为$S_{count} = n_1 \times 1 + n_2 \times 0.5 + (n - n_1 - n_2) \times 0$。AUC 定义为$S_{count}/n$，即：

$$AUC = {(n_1 + n_2 \times 0.5)}/{n} \qquad (3.1)$$

AUC 定义中，AUC 的值越大越接近 1 表示相似度算法越准确。

3.2 Precision

Precision 从局部来衡量相似度算法的准确度，在计算的过程只考虑了在相似度排名序列 X 所取的前 M 位的节点对连边预测的准确性[27]。Precision 算法将所取的前 M 个节点对连边视作平等的，不区分其相似度分数值大小和排名先后。

假定在序列 X 所取的前 M 个节点对连边中有 m 条节点对连边在测试集 E^P 中，所以 Precision 的值为：

$$Precision = {m}/{M} \qquad (3.2)$$

Precision 的值越大越接近 1 表示预测越准确。

3.3 Ranking Score

Ranking Score 从相似度排名序列的顺序来计算相似度算法预测的准确度[28]。Ranking Score 的定义为测试集 E^P 中的边在序列 X 中的位置。

假定 E^P 中的边 $e^p{}_i$ 在序列 X 中的排名为 r_i，则 $e^p{}_i$ 在序列 X 中的 Ranks 值为 $r_i/|X|$。遍历 E^P 中的所有边的 Ranks 值取平均，得集合 E^P 的 Ranking Score 值为：

$$\text{Ranking Score} = \frac{1}{|E^p|}\sum_{i\in E^p}\text{Ranks} = \frac{1}{|E^p|}\sum_{i\in E^p} r_i/|X| \qquad (3.3)$$

Ranking Score 的值越小越接近 0 表示测试集 E^P 中的边在序列 X 中的位置越靠前，相似度算法预测的越精确。

第四章 实验

4.1 数据预处理

本文实验中，从 UN comtrade（https://comtrade.un.org/）上（HS Code 270900）获取了 1990 年至 2017 年的国际原油贸易数据，其中 1990 年到 1999 年的数据只用于对后面的数据进行近因分析。预数据处理中，清除掉原始数据中的重复数据和自回路数据。

不考虑其贸易流量的大小和方向，以年为单位将一年的所有国家的国际原油贸易数据转化为无权无向的国际原油贸易网络。实验中我们可以得到 28 个国际原油贸易网络来进行链路预测。如图 4.1 为 2010 年的国际原油贸易网络，图中大部分区域的边分布稀疏，少部分区域的边分布密集。

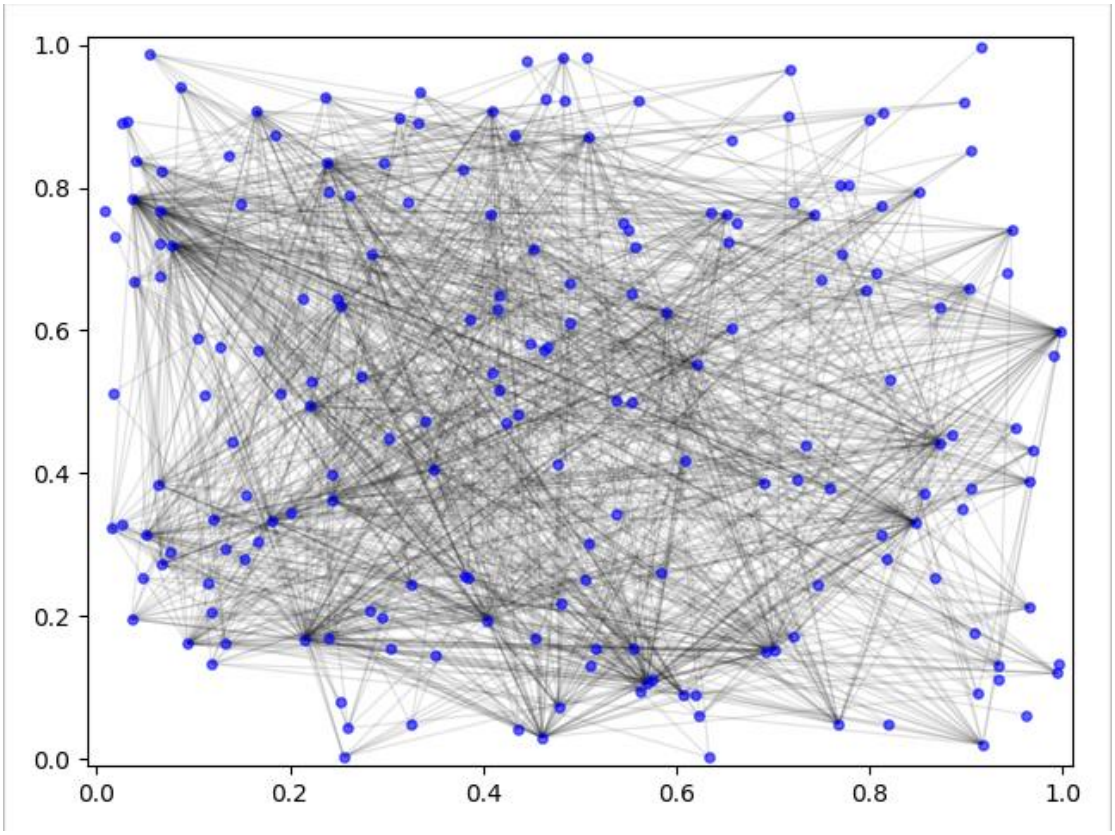

图 4.1 2010 年国际原油贸易关系图

进一步分析 2010 年的国际原油贸易网络的度分布图（图 4.2）发现，度为 10 以下的节点占据了网络中的大部分，少部分节点的度在 10 到 100 之间。网络中大部分节点只有少量的邻居，而少部分节点与大量的节点相连（图 4.3），意味着国际原油贸易网络的少部分国家拥有大部分的原油资源[12]。

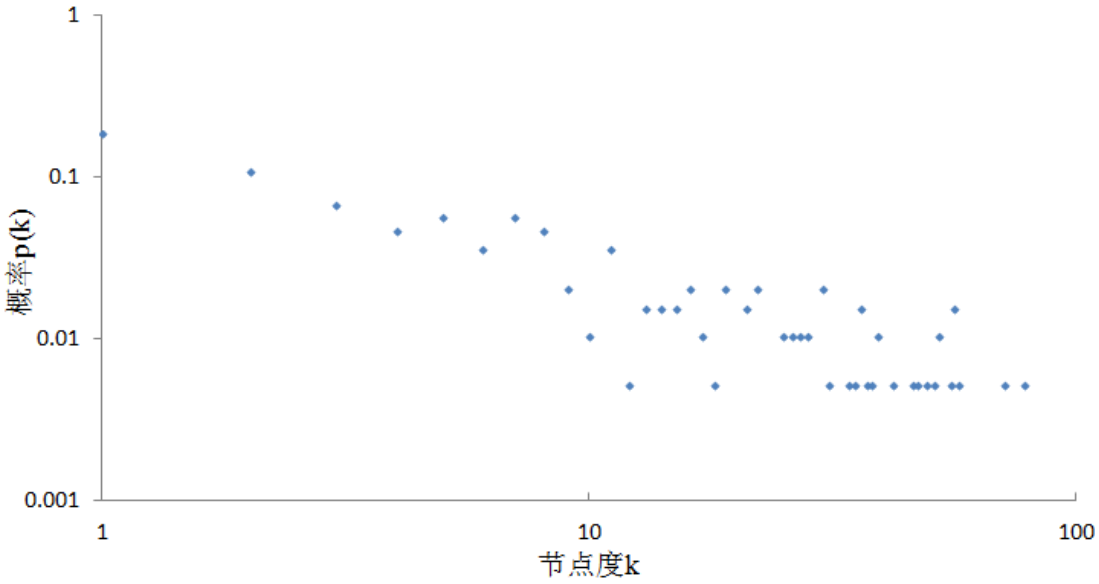

图 4.2 节点度概率分布图

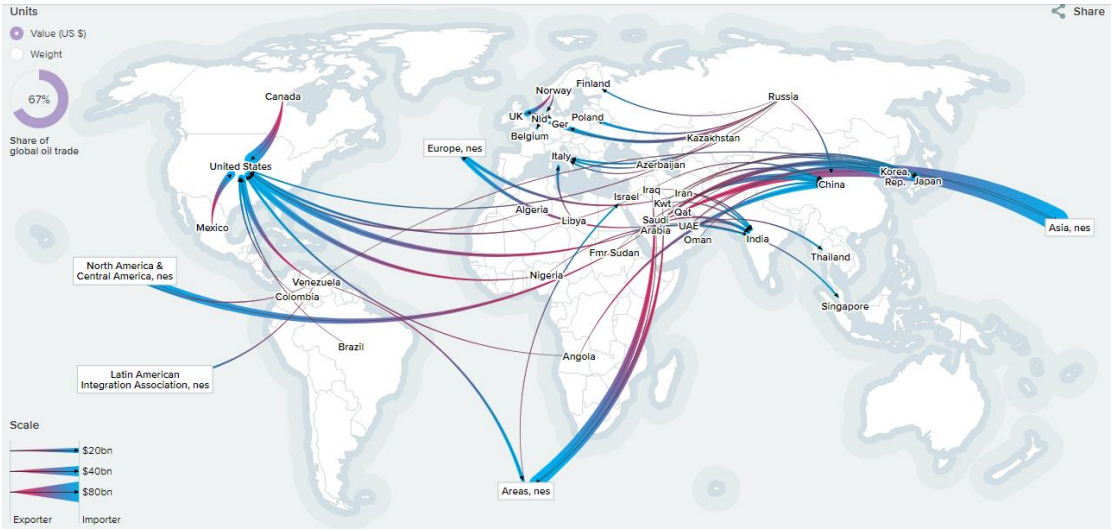

图 4.3 2010 年国际原油贸易流向图

4.2 实验过程

本文中以 10%为比例在边集 E 选取测试集 E^P 进行基于节点相似度的链路预测，所使用的相似度方法有 CN、JC、AA、PA、RA 以及分别结合节点近因和结

合边近因的相似度方法。每个国际原油贸易网络在各个方法下各进行了 50 次的独立重复实验。最后计算 Ranking Score 值来评估链路预测的准确度，进一步分析国际石油贸易网络的演化因素。

4.3 结果及分析

4.3.1 2000 年测试集分布

图 4.4 中展示了 2000 年的相似度排序中测试集的分布，直观的展示了几种相似度方法（CN、JC、AA、PA、RA）的准确度。图 4.4 蓝色的部分表示了测试集中边的排序位置。

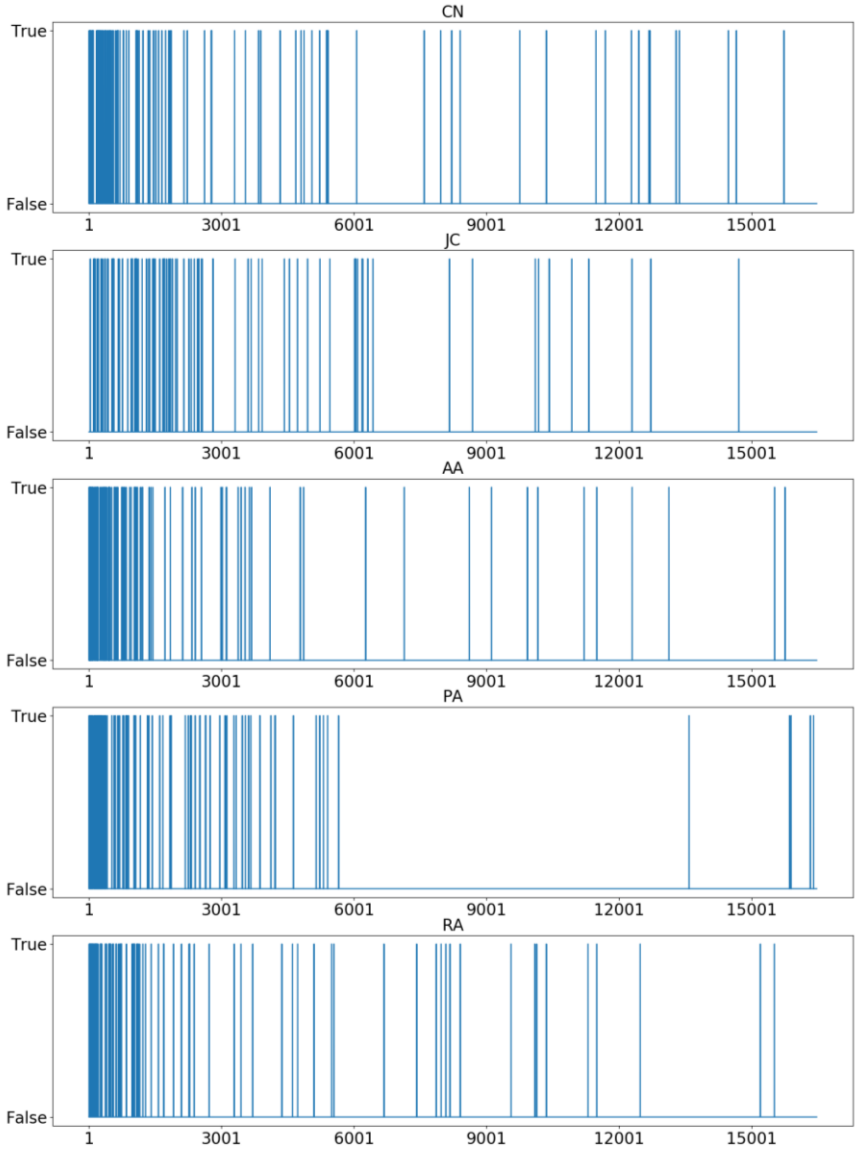

图 4.4 2000 年测试集分布

在图 4.4 的测试集分布中，所有相似度方法的测试集主要分布在相似度序列的头部位置，只有少量的测试集边位于相似度序列的尾部但不影响测试集的整体分布。结果表明几种常用的相似度方法对于国际原油贸易网络的链路预测是准确有效的。贸易伙伴和共同贸易伙伴的数量等都是影响国际原油贸易的潜在贸易或未来贸易的重要因素。

并且分布图中 CN、AA、RA 的测试集在头部的分布较为密集，而 JC、PA 的测试集在头部的分布较为稀疏。表明在国际原油贸易中，共同贸易伙伴对贸易关系的贡献大于贸易伙伴对贸易关系的贡献。

4.3.2 常用相似度方法的 Ranking Score 值对比

图 4.5 中以年份为横坐标，Ranking Score 值为纵坐标，量化地展示了相似度方法 CN、JC、AA、PA、RA 的预测准确度随时间的变化趋势。

依据 Ranking Score 值的折线变化趋势，CN、AA、PA、RA 这四种表现较好的相似度方法在 2000 年到 2017 年间的预测准确度总体上是随着年份的增长在逐渐提高的。这意味着 2000 年后的国际原油贸易市场稳定，国家趋于与拥有较多共同贸易伙伴的国家之间建立原油贸易关系，共同贸易伙伴是对贸易关系的有力保证。

而在 CN、AA、PA、RA 表现优异的同时， JC 对国际原油贸易网络的预测的准确度明显低于其他方法。图 4.5 中 JC 的 Ranking Score 值在 0.17 以上，而其他指标的 Ranking Score 值在 0.15 以下。JC 是建立在公共邻居基础上考虑所有邻居的数量，对比 JC 与 CN 的折线，JC 的 Ranking Score 值明显高于 CN，且折线的变化趋势明显不同。这表明国际原油贸易中，若国家间的共同贸易伙伴数量稳定，在国家贸易伙伴数量出现变化时，国家间的潜在贸易联系受到的影响较小。

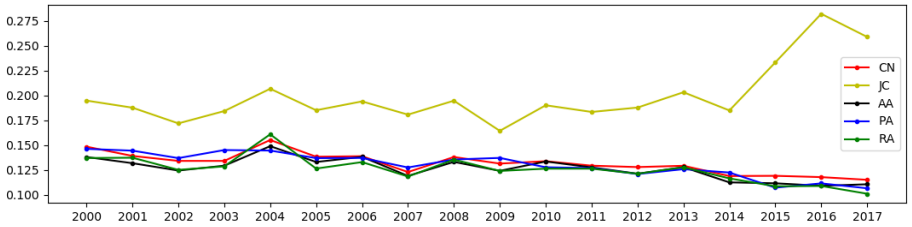

图 4.5 相似度方法的 Ranking Score 值折线

同时 JC 的 Ranking Score 值明显高于 PA 的 Ranking Score 值，但 PA 的 Ranking Score 值和变化趋势与 CN、AA、RA 的 Ranking Score 值和变化趋势是

基本一致的。这说明了贸易伙伴和共同贸易伙伴都是影响两个国家之间原油贸易关系的重要因素，但是共同贸易伙伴占国家所有贸易关系的比例不是影响国际原油贸易关系的重要因素。一个拥有较少贸易伙伴的国家 A 是更倾向于与拥有较多贸易伙伴和较多共同贸易伙伴的国家 B 之间建立贸易联系。而不是与同样拥有较少贸易伙伴但是有大部分贸易伙伴是共同贸易伙伴的国家 C 之间建立贸易联系，这样的两个贸易国家 A、C 可能是有着类似国情的两个国家，例如 A、C 可能都是富原油小国，贸易伙伴不多，但大多是共同贸易伙伴。同样也说明了一个拥有大量贸易伙伴的大国，其贸易伙伴中的大部分国家之间建立贸易关系的可能性是较小的，与 AA、RA 指标在国际原油贸易网络的优异表现相符。

4.3.3 时间近因分析

图 4.3 至图 4.7 分别展示 CN、JC、AA、PA、RA 五中相似度度量方法结合节点近因和边近因的变化。

图 4.3 至图 4.7 中五种相似度度量方法的折线与各自结合近因的相似度度量方法的折线变化趋势是基本一致，表明近因对于几种相似度度量方法的度量是有效的。

从节点近因的方面可以看到，结合节点近因的相似度方法的黑色折线和绿色折线与原方法的红色折线非常接近，说明节点近因对于相似度度量方法的优化很小。这说明在目标国际原油贸易网络中，节点活跃度低，节点的活跃度对于节点对之间连边的影响很小。这可能有两方面的原因，一是国际贸易相对稳定，在国际原油贸易中卖方、买方相对稳定，各国政府为了规避风险会维持己方的贸易关系，整体的节点活跃度较低，所以加入节点活跃度对相似度算法的优化较为微弱。二是国际贸易网络中出现节点度的变化的节点，度的变化对其共同邻居的影响小，所以没有影响到节点对之间的潜在联系。

整体上节点活跃性低，表明 2000 年后国际原油贸易市场处于稳定发展期，国家的贸易伙伴和贸易联系不会轻易产生波动。所以在稳定的国际原油贸易市场中，结合边的近因对于相似度度量方法的优化是明显的。并且，在折线图中可以看到蓝色折线的 Ranking Score 值整体上明显低于黄色折线，其对于相似度方法的优化最高，说明边近因在国际原油贸易网络的相似度度量中的影响较强。表明在 2000 年以后国际原油贸易网络中，国际原油贸易市场稳定，两个国家之间有了贸易联系后会继续保持这个贸易联系。过去的贸易联系是影响国家之间原油贸易的重要影响因素，国家在考虑未来的原油贸易时会优先考虑曾经的伙伴。边近因对于国际原油贸易的明显优化也表明了时间是影响动态网络链路预测准确度

的重要影响因素，选择合适的方法结合时间因素对动态网络的链路预测有着较为明显的积极影响。

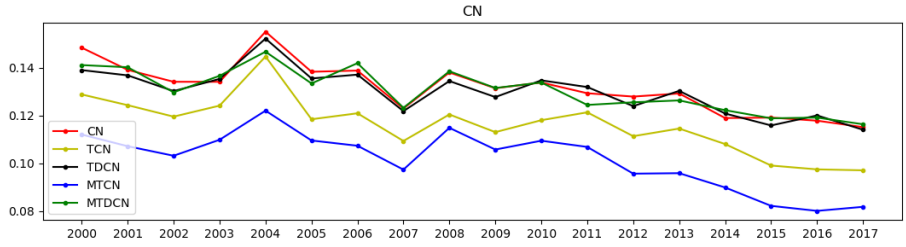

图 4.6　基于 CN 的近因优化

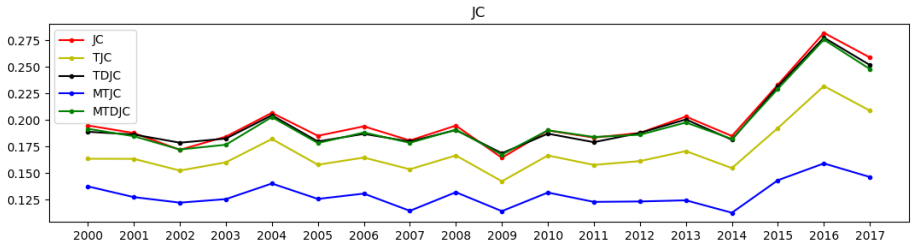

图 4.7　基于 JC 的近因优化

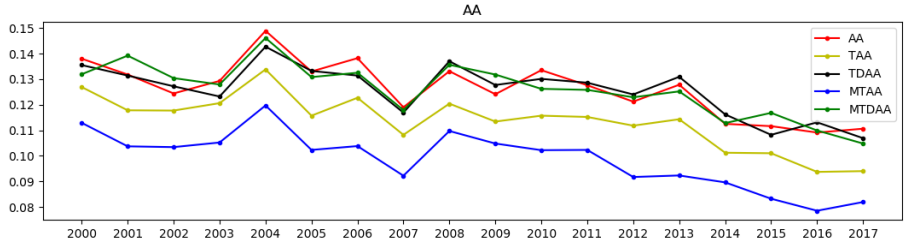

图 4.8　基于 AA 的近因优化

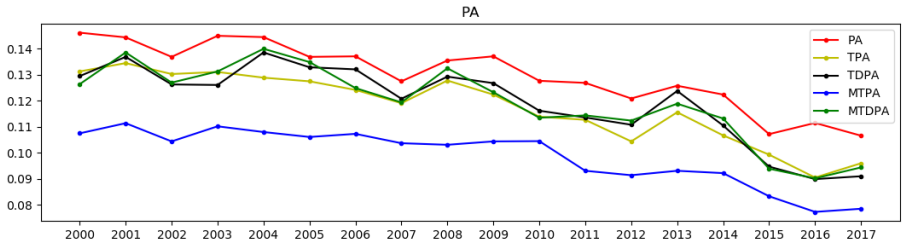

图 4.9　基于 PA 的近因优化

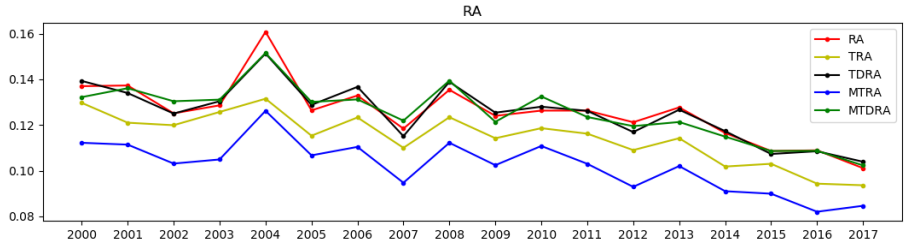

图 4.10　基于 RA 的近因优化

第五章 结论

对于动态复杂网络，其影响因素不是单方面的，而是多方面因素的综合作用。在复杂网络的链路预测中，要提高预测的准确度，可用不同的相似度方法来进行链路预测。结合不同的方法对比分析，可以有效的提高链路预测的准确度，选择更合适的相似度方法，并且更合理的推断网络的演化机制。

在本文的国际原油贸易网络中，国家的贸易伙伴和共同贸易伙伴数量都是对国家贸易关系的建立有积极影响的重要因素，但是贸易网络不同于通信网络，在通信网络中两个孤僻但是有着共同好友的人是有较大可能建立起通信，但是在贸易网络中两个即使是有着共同贸易伙伴但是贸易伙伴数量较少的国家之间建立贸易联系的可能性较低。

同时，在这个国际形势和平发展的时代，国际贸易的形势也是平稳发展的。过去时间点中有过贸易关系或者保持着贸易关系的国家之间在未来重新产生贸易关系或继续保持贸易关系的可能性较高。

本文在研究中用基于节点局部信息的相似性进行国际原油贸易网络的预测分析，但是并没有考虑国际原油贸易网络中进口国与出口国的角色问题和贸易流量问题。在国际原油贸易网络的进一步研究中，引入出度，入度和权重，不仅能进一步分析国际原油贸易网络的演化机制，而且预测结果更精确，更有实用价值。

并且活跃节点和活跃连边使用第一类定义能更精确的描述一段时间内节点和连边的活跃性，以此定义近因并结合相似度方法来计算节点相似性能进一步提升预测的准确度。

参考文献

[1] An H, Zhong W, Chen Y, et al. Features and evolution of international crude oil trade relationships: A trading-based network analysis[J]. Energy, 2014, 74: 254-259.

[2] Helpman E, Melitz M, Rubinstein Y. Estimating trade flows: Trading partners and trading volumes[J]. The quarterly journal of economics, 2008, 123(2): 441-487.

[3] Ji Q, Zhang H Y, Fan Y. Identification of global oil trade patterns: An empirical research based on complex network theory[J]. Energy conversion and management, 2014, 85: 856-865.

[4] Zhang Z, Lan H, Xing W. Global Trade Pattern of Crude Oil and Petroleum Products: Analysis Based on Complex Network[C]//IOP Conference Series: Earth and Environmental Science. IOP Publishing, 2018, 153(2): 022033.

[5] Fagiolo G, Reyes J, Schiavo S. The evolution of the world trade web: a weighted-network analysis[J]. Journal of Evolutionary Economics, 2010, 20(4): 479-514.

[6] Duan W. Research on the measurement and evolution model of world trade networks[J]. 2011.

[7] Fagiolo G, Reyes J, Schiavo S. World-trade web: Topological properties, dynamics, and evolution[J]. Physical Review E, 2009, 79(3): 036115.

[8] Zhong W, An H, Gao X, et al. The evolution of communities in the international oil trade network[J]. Physica A: Statistical Mechanics and its Applications, 2014, 413: 42-52.

[9] Yang Y, Poon J P H, Liu Y, et al. Small and flat worlds: A complex network analysis of international trade in crude oil[J]. Energy, 2015, 93: 534-543.

[10] Zhang H Y, Ji Q, Fan Y. Competition, transmission and pattern evolution: A network analysis of global oil trade[J]. Energy Policy, 2014, 73: 312-322.

[11] Foschi R, Riccaboni M, Schiavo S. Missing links in multiple trade networks[C]//2013 International Conference on Signal-Image Technology & Internet-Based Systems. IEEE, 2013: 585-590.

[12] Guan Q, An H, Gao X, et al. Estimating potential trade links in the international crude oil trade: A link prediction approach[J]. Energy, 2016, 102: 406-415.

[13] 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.

[14] 吕琳媛. 复杂网络链路预测[J]. 电子科技大学学报, 2010, 39(5): 651-661.

[15] Liu H. Uncovering the network evolution mechanism by link prediction[J]. Scientia Sinica, 2011, 41(7):816-823.

[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] 王林, 商超. 无标度网络中的链路预测问题研究[D]. , 2012.

[18] Lin D. An information-theoretic definition of similarity[C]//Icml. 1998, 98(1998): 296-304.

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

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

[21] Barabási A L, Albert R. Emergence of scaling in random networks[J]. science, 1999, 286(5439): 509-512.

[22] Zhou T, Lü L, Zhang Y C. Predicting missing links via local information[J]. The European Physical Journal B, 2009, 71(4): 623-630.

[23] Aghabozorgi F, Reza Khayyambashi M. A new study of using temporality and weights to improve similarity measures for link prediction of social networks[J]. Journal of Intelligent & Fuzzy Systems, 2018, 34(4): 2667-2678.

[24] Potgieter A, April K A, Cooke R J E, et al. Temporality in link prediction: Understanding social complexity[J]. Emergence: Complexity & Organization (E: CO), 2009, 11(1): 69-83.

[25] Mohan S, Subramanian M. A New Method of Identifying Individuals' Roles in Mobile Telecom Subscriber Data for Improved Group Recommendations[C]//International Conference, MISNC. Springer, Berlin, Heidelberg, 2014: 213-227.

[26] 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.

[27] Herlocker J L, Konstan J A, Terveen L G, et al. Evaluating collaborative filtering recommender systems[J]. ACM Transactions on Information Systems (TOIS), 2004, 22(1): 5-53.

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

致　　谢

论文写作是大学最后一个学期的最重要的一次经历。在这次论文写作中，感谢几个月以来指导老师李老师在论文选题，查阅文献资料，实验到写作对我的指导与意见。同时也感谢所有大学四年来所有的老师的教导。

论文（设计）成绩

导师评语 建议成绩 _____ 指导教师（签字）_____
答辩小组意见 答辩委员会负责人（签字）_____
成绩 _____ 学院（盖章）_____ 年 月 日