



北京大学

# 硕士研究生学位论文

题目：面向高速路网的关键路段识别  
研究与实现

姓 名： 刘丹萌

学 号： 1401214385

院 系： 北京大学

专 业： 计算机科学与技术(智能科学与技术)

研究方向： 数据仓库与数据挖掘

导 师： 宋国杰

2017 年 5 月



---

## 版权声明

任何收存和保管本论文各种版本的单位和个人，未经本论文作者同意，不得将本论文转借他人，亦不得随意复制、抄录、拍照或以任何方式传播。否则一旦引起有碍作者著作权之问题，将可能承担法律责任。



## 摘要

交通问题是当今世界关注的热点问题。随着人们生活水平的提高、交通系统的发展，社会对交通的需求也日益增大。交通与环境、交通与能源、交通与需求之间的矛盾日益加剧，交通事故和堵塞给人们带来了巨大的效率、能源和生命上的损失。在高速公路的建设中，如何选取关键路段，并对关键路段进行针对性建设已经成为一个热点问题。

在传统交通关键路段研究中，研究方法主要有①针对交通网络拓扑结构的研究；②针对交通网络统计数据进行分析；③在微观层面下，针对路段的特性进行研究。这些方法都有各自的局限性。方法①只针对了路网的拓扑结构，没有考虑节点之间的信息交流；方法②基于统计学方法，该方法只能静态分析路网关键路段，无法分析关键路段随着时间／路网流量的变化规律；方法③主要研究交通的微观领域特性，没有考虑路网的宏观变化规律。在此我们提出一种基于宏观高速公路网络的目标模型。

本文从智能交通的实际应用角度出发，针对现有关键路段挖掘方法的不足，提出两个研究方法：①提出一种基于贪心算法的关键路段挖掘模型，并证明贪心算法的可行性；②提出一种基于复杂网络社群划分的关键路段挖掘模型，实现动态挖掘关键路段。

集成上述成果，实现了一个基于 B/S 架构的高速公路关键路段挖掘系统，并且在真实应用场景下初步验证了原型系统的可靠性和适用性。

**关键词：**复杂网络，智能交通，关键路段，社群划分



## Test Document

Test (Some Major)

Directed by Prof. Somebody

### ABSTRACT

Traffic problem is a hot issue in the world today. With the improvement of people's living standard and the development of the traffic system, the demand of social traffic is increasing day by day. The contradiction between traffic and environment, transportation and energy, traffic and demand increasing, traffic accidents and congestion brings efficiency, energy and life in the great loss, simple traffic control technology has been unable to meet the demand. The study of the traditional intelligent transportation is based on the research of the single spatial location in the road network, which is based on the theory of dynamics, statistics, simulation and machine learning. With the continuous development of the traffic system, traffic system gradually presents the network situation, the close connection between each node in the network, the macro characteristics of single point position is not enough to describe the whole Expressway network. With the traffic accident has become the bottleneck of the traffic system, the research of intelligent transportation system has new demand: how to find the key nodes in the traffic system, by processing the key nodes, in order to reduce the traffic paralysis rate, increase the operation stability of road network to.

In the intelligent highway system network, the research focused on 1 key nodes) study on traffic network topology; 2) for statistical research on

traffic network information; 3) using the propagation dynamics, the research of micro site. Our purpose is to find the key nodes in the network, improve the efficiency in the whole network, method 1) only for the topology of the network, do not consider the exchange of information between nodes; 2) based on the traditional statistical methods, and on the basis of statistics, using data mining method of key nodes, and the change in the flow of network traffic change over time, this method can only static analysis of network key nodes, to analysis of key nodes changes with time and network flow method; 3) focus on the traffic characteristics of the micro field, the whole road network is of little significance. In this paper, we propose a target model based on macroscopic Expressway network. In order to deeply explore the key nodes of freeway, we propose a more complex probabilistic model. This model has high time complexity and can not meet the requirement of real-time application of intelligent transportation system. Based on the small world characteristics of complex networks, this paper introduces the clustering algorithm of expressway network, which is divided into four parts

Therefore, this article from the intelligent transportation according to the practical requirement, aiming at the key nodes of mining lack of research methods, in-depth study of two aspects: the function model between the key nodes and highway traffic conditions put forward a description of expressway, aiming at the limitations of existing key nodes of the complex network, put forward the objective function from the macro level, combined with real-time traffic highway, on key nodes in the expressway real-time position; proposes a classification method of highway network based on community. The resolution limit and extreme for the existing community division method in the degradation characteristics, combined with the

characteristics of the highway network, community partition model is established by the highway, to a certain extent to solve the resolution limitation of the traditional method and extreme degradation characteristics.

The main contributions of this paper are as follows:

- (1) In this paper, a new model of key nodes in highway network is proposed, which breaks through the limitations of existing complex networks
- (2) In this paper, a new model of community partition based on expressway network is proposed, which can reduce the complexity of the key nodes and make the method practical.

**KEYWORDS:** Complex network, Intelligent Transportation, Key road, Community partition



# 目录

<b>第一章 绪论</b>	<b>1</b>
1.1 研究背景 . . . . .	1
1.2 研究内容 . . . . .	4
1.3 研究意义 . . . . .	5
1.4 论文结构 . . . . .	6
<b>第二章 相关研究</b>	<b>7</b>
2.1 高速公路关键路段/节点挖掘相关研究 . . . . .	7
2.2 复杂网络关键路段/节点挖掘相关研究 . . . . .	8
2.2.1 基于节点临近 . . . . .	8
2.2.2 基于路径临近 . . . . .	10
2.2.3 基于特征向量的排序方法 . . . . .	17
2.2.4 基于节点移除和收缩的排序方法 . . . . .	21
2.2.5 节点重要性排序方法的评价标准 . . . . .	25
2.3 复杂网络社群划分相关研究 . . . . .	28
<b>第三章 基于贪心的高速公路关键路段识别模型</b>	<b>33</b>
3.1 引言 . . . . .	33
3.2 问题定义 . . . . .	34
3.2.1 问题定义 . . . . .	34
3.2.2 模型定义 . . . . .	35
3.2.3 次模性分析 . . . . .	36
3.2.4 基于贪心法的关键路段求解 . . . . .	38
3.3 实验及结果 . . . . .	41
3.3.1 实验数据 . . . . .	41

3.3.2 实验结果 . . . . .	42
3.3.3 时间分析 . . . . .	45
3.4 本章小结 . . . . .	46
<b>第四章 基于社群划分的关键路段识别方法</b>	<b>47</b>
4.1 引言 . . . . .	47
4.2 问题定义 . . . . .	48
4.2.1 问题定义 . . . . .	48
4.2.2 模型定义 . . . . .	51
4.2.3 基于社群划分的关键路段求解 . . . . .	53
4.3 实验及结果 . . . . .	56
4.4 本章小结 . . . . .	62
<b>第五章 原型系统的设计与实现</b>	<b>63</b>
5.1 系统功能 . . . . .	63
5.2 系统架构 . . . . .	63
5.3 界面功能展示 . . . . .	64
5.4 本章小结 . . . . .	64
<b>总结与展望</b>	<b>67</b>
5.4.1 主要工作 . . . . .	67
5.4.2 未来工作展望 . . . . .	68
<b>参考文献</b>	<b>69</b>
<b>附录 A 附件</b>	<b>73</b>
<b>致谢</b>	<b>75</b>

## 插图

1.1 中国高速公路网络示意图 . . . . .	1
1.2 国家科技支撑计划 . . . . .	4
3.1 关键路段挖掘：以 1h 为区间 . . . . .	43
3.2 关键路段挖掘：以 1d 为区间 . . . . .	43
3.3 不同方法求得的关键路段结果图 . . . . .	44
3.4 关键路段挖掘：03: 00 . . . . .	45
3.5 关键路段挖掘：09:00 . . . . .	45
4.1 高速公路车辆跳数分布图 . . . . .	49
4.2 高速公路社群特性 . . . . .	49
4.3 基于模块化函数的社群划分方法 . . . . .	57
4.4 基于模块化函数的社群划分方法 . . . . .	58
4.5 结合物理路网特性的社群挖掘 . . . . .	58
4.6 结合路段距离的变权社群挖掘结果 . . . . .	59
4.7 对比实验：以 1h 为区间 . . . . .	60
4.8 对比实验：以 1d 为区间 . . . . .	60
4.9 枚举求得关键路段 . . . . .	60
4.10 贪心求得关键路段 . . . . .	60
4.11 基于社群划分的关键路段 . . . . .	60
4.12 基于统计学的关键路段 . . . . .	60
5.1 智能高速系统架构 . . . . .	64
5.2 逻辑流程图 . . . . .	65
5.3 系统分群结果图 . . . . .	65

5.4 系统路段选取结果图 . . . . .	65
-------------------------	----

## 表格

3.1 算法结果集 . . . . .	43
3.2 不同算法的运行时间 . . . . .	45
4.1 不同社群划分方法效果对比 . . . . .	59
4.2 路网通行效率提升量误差分析 . . . . .	61
4.3 关键路段选取误差分析 . . . . .	61
4.4 不同方法运行效率分析 . . . . .	61



# 第一章 絮論

## 1.1 研究背景

高速公路是支撑国家经济发展、服务群众生活、保障国家安全的战略资源和设施。截止 2016 年底，我国公路通车总里程达到 457 万公里，其中高速公路 12 万公里，2017 年将新增 4500 公里，是世界上规模最大的高速公路（如图 1.1 所示）。



图 1.1 中国高速公路网络示意图

随着中国经济的快速发展，人们生活水平的不断提高，居民的出行和货物运输的数量也在逐渐增加。交通出行是人类活动不可缺少的一部分。据估计，每天平均有 40% 的人口在路上花费至少 1 小时。近几年来，人们变得越来越依赖于交通系统，对于交通管理人员来说，机遇和挑战共存。首先，交通拥堵已成为一个日益严重的问题。全球范围内的道路上的车辆增加，根据调查，截止至 2016 年初，北京共有 544 万辆车，比 2014 年初增加了 50 万辆。这些激增的车辆会对道路系统产生严重的压力，极大的增加拥堵以及拥堵后的损耗。拥堵会导

致燃油消耗增加，空气污染，以及实施公共交通计划的困难。车流流量过多时，交通事故风险与交通运输系统中的膨胀增加，交通事故之后的恢复时间与恢复代价也会急剧增加。在中国，2009年的交通事故死亡人数约有7万人，在2015年达到9万人。美国联邦公路管理局公布的报告显示，发生在城市的交通事故约占所有拥堵延误的50% - 60%。一个国家的技术竞争力、经济实力和生产能力，在很大程度上取决于其交通系统性能。交通管理资源有限，中国高速公路正在逐渐走向免费通行，难以对交通系统进行全面建设；同时，国家运输系统的有效性也依赖于一个国家的处理紧急情况的能力，如交通事故发生以后的大规模疏散。随着高速公路的不断发展，各类维护高速公路的需求也都被一一提出：

1) 人员分配问题。最典型的是今年五一，交通部联合路政大队、高速交警大队、项目部制定组有针对性的应急预案，根据历史的交通信息，负责做好恶劣天气、旅游高峰车流量饱和、突发事件引起交通阻塞的应急处理。安排值班人员，落实机械设备和应急物资准备，一旦发生突发事件，迅速启动，切实做好节日期的保畅工作。

2) 安全管理。面对节日期间可能出现的状况，高速公路养护所需要在节日期前组织开展道路安全隐患检查活动：一是对管段路段进行安全隐患排查，发现问题立即落实整改；二是加强春季防火工作的管理，及时清理桥涵构造物下的易燃物品，对边坡、中央分隔带进行打草、苗木修剪，对匝道圈进行专人看护，安排养护员负责所辖路段的防火报警工作。

3) 基础设施建设。交通的基础设施建设主要包括轨道交通设施、停车场设施建设等。交通基础设施包括为交通系统保障安全正常运营而建设的公路、轨道、隧道、高架道路、车站、通风亭、机电设备、供电系统、通信信号、道路标线等设施。

4) 高新设施建设。目前越来越多的新技术涌现，高新技术层出

不穷 [1]。这些新技术对高速公路稳定性的提高具有重要意义，但是，新技术普遍造价高昂，无法直接大面积使用，需要交通管理者进行合理分配。

上述的交通需求中有一个共同问题，那就是如何找出高速公路网络中最重要的路段。针对高速公路中的重要路段，交通管理者可以进行针对性管理。高速公路的变化日新月异，传统的研究方法很难准确的预估关键路段，亟需提出一套适用于大规模路网的高速公路的关键路段挖掘识别系统。在准确识别高速公路关键路段的前提下，我们可以进行有效的人员分配，避免资源浪费；可以针对高危路段进行安全管理，减少事故风险；可以针对核心路段进行基础设施建设，提升高速公路稳定性，节省预算；可以研究道路施工问题，合理分配施工路段，将施工造成的影响减到最小。高速公路关键路段挖掘问题是交通研究领域的重要基础问题。

本文提出面向高速公路的关键路段挖掘模型，以高速网络整体运行效率为评价标准，分析高速公路的网络特性，给出贪心算法的可行性证明，提出基于贪心算法的高速公路关键路段挖掘模型；在贪心算法基础上，提出一种优化策略，给出一种基于复杂网络社群划分的关键路段挖掘模型，以优化算法效率，实现关键路段动态挖掘。

本文针对高速公路进行关键路段挖掘，主要基于高速公路的收费数据，并得到如下项目的支持：

1) 山西/安徽省智能交通系统

实验室曾经开发过面向山西/安徽省的智能交通系统，在这个系统里，我们开发了比较完善的底层数据存取框架，采用 mysql 数据库加 infobright 数据仓库的数据存储格式，Linq to sql 数据存取方式，集成数据存取逻辑接口，为本研究提供数据基础。

2) 国家科技支撑计划：高速公路网运行状态智能检测与安全服务保障关键技术研发及系统集成

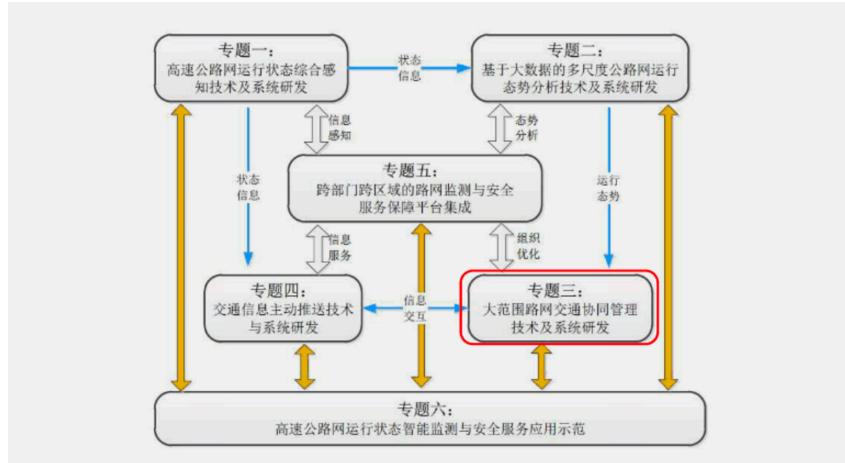


图 1.2 国家科技支撑计划

图1.2为项目专题间关系图，我们主要负责专题三：大范围路网交通协同管理技术以及系统研发。专题一主要研究高速公路的网络监测设施的部署问题，这些设施将用于改善路网运行状态以及获取路网运行数据；专题二主要用于分析和处理这些数据，将这些数据转化为流量、速度、密度和 O-D 信息。专题三负责提出决策模型，基于前面专题获取的数据，输出目前路网状态下最合理的绕行路线和限流建议；专题四负责研究如何设置设施，将这些建议分发出去；专题五和专题六负责将整个系统进行集成。可以看出，本课题中的专题一、专题三、专题四都和高速公路重要路段挖掘方法有关：专题一中的监测设施主要部署在重要路段中，优化监测效果；专题三中的决策模型可以基于路段的重要程度，进行导流决策；专题四中可以根据重要路段的分布信息，划定信息发布点，优化信息发布行为。

## 1.2 研究内容

本文从交通领域的诸多实际问题出发，研究高速公路关键路段挖掘问题。针对现有复杂网络关键路段挖掘技术的不足，提出以下两个研究思路：

- 1) 提出一种基于贪心算法的关键路段挖掘模型，并给出次模性

证明。在高速路网上建设基础设施、部署警力，增加维护成本，都可以归类于对高速公路中路段的资源投放问题。本文提出关键路段挖掘模型，以每条路段都具有一定的损毁概率为基础，定义当路段获得投资后，损毁概率下降；结合路段损毁概率矩阵，计算用户的出行效率，最终获得高速公路的通行效率。本文还分析了该模型的性质，证明了模型的次模性，并给出了贪心求解步骤。

2) 提出一种基于复杂网络社群划分算法的关键路段挖掘模型。本文提出的关键路段挖掘模型，本质上可归类于概率规划问题，时间复杂度达指数级。在小范围路网中，模型计算时间较少，可以实现静态路网中关键路段的挖掘。但是在大范围路网中，算法的时间消耗会急剧增长，而且算法对于需要实时动态分析关键路段的需求无法满足。本节分析了以往复杂网络社群划分的局限性，给出一种结合高速公路的网络特性的社群划分方法。根据高速公路路段特性，结合模拟退火和变权算法，解决传统社群划分方法中的低分辨率问题 [2] 和极端退化问题 [2]。

### 1.3 研究意义

研究意义主要包括两个方面：

1) 应用实践价值：从应用的角度来看，随着我国高速公路网络规模的逐渐扩大，路网结构日益复杂，人们的出行需求逐年增加，高速公路遇到的挑战不断增加，高速公路的稳定越来越重要。高速公路监测设施不断完善，可以实时监测高速公路中的车流信息。本文结合高速公路的特性，解决传统关键路段挖掘方法中的不足，优化时间复杂度，使得模型可以在可承受时间复杂度内求解，最终实现静态、动态挖掘路网的关键路段。

2) 理论研究价值：本文提出的“面向高速公路的关键路段挖掘模型”具有一定的理论创新性。将一些高速公路建设上的具体问题抽象

成一类基础问题，并给出一个可行解。传统高速公路关键路段研究，大多是研究高速公路的统计特性（如流量的大小，路段周围城市的重要程度，路段周围环境的变化等），本文利用数学模型描述各类高速公路现象，从宏观角度给出并求解一种具有普适性的高速公路关键路段挖掘模型。

## 1.4 论文结构

第一章为绪论，介绍了本文的研究背景，提出了本文的研究内容。第二章介绍了相关研究，主要介绍了高速公路关键路段研究、复杂网络关键路段研究的相关工作，结合交通问题的特点分析了现有方法的优势与不足；同时对复杂网络社群划分方法及其相关研究进行了介绍，通过对现有社群划分方法的分类对比，分析了它们的优势与不足；第三章论述本文的主要研究内容，提出了一种复杂网络关键路段挖掘模型，分析模型的贪心可解性，并给出了贪心解法与实验。第四章提出了一种基于复杂网络社群划分的关键路段挖掘方法，给出了高效的优化算法和详尽的理论分析，并在多个数据集下的进行了验证。第五章给出了本文实现的一个原型系统，并且已经在实际数据上运行。第六章给出了全文的总结与未来工作展望。

## 第二章 相关研究

### 2.1 高速公路关键路段/节点挖掘相关研究

高速公路的关键路段挖掘问题研究较少，主要分为基于统计的研究方法和基于路网拓扑结构的研究方法。

2016 年，Yip 等人基于高速公路统计学方法 [3]，研究了高速公路路段的重要程度。该文章从高速公路拥堵情况出发，讲述了高速公路关键路段挖掘的意义，并且基于弗吉尼亚的交通管控系统，获取路段的拥堵概率，构建概率模型，根据拥堵率来选择关键路段。2016 年 8 月，Kerner 提出了一种基于微观道路信息的关键路段挖掘模型 [4]，该模型结合道路的集合形状，考虑驾驶员的视线等因素，利用路段安全性度量函数来挖掘关键路段。2017 年，Yacine 结合路段滑坡敏感性，给出了君士坦丁公路中路段的敏感性挖掘方法 [5]。2014 年，Ren 等人基于节点的重要程度进行链接位置的优化，他们用节点的地理指标和集群特性来判定节点的重要程度。2011 年，Song 等人基于因子分析法，结合 k-means 算法对每个节点的重要程度进行进一步的划分 [6]。2011 年，Wang 等人基于节点的度、介数和交通量进行关键路段研究，同时结合 C-means 聚类方法进行进一步排序 [7]。2013 年，他们又根据节点删除法来进行关键节点的识别 [8]。2010 年，Chen 等人认为影响节点重要程度的因素有很多，而且它们的权值固定不变，基于这个思想，提出了一种基于 Matlab 聚类分析的解点重要性划分方法。现有的高速公路的关键路段研究较少，且大都是基于微观角度、基于统计学特性、基于路网结构求解，有各自的局限性。基于微观角度的研究方法虽然可以模拟真实情况下的路网状况，但是这类研究单纯的研究细节层面的高速公路网络规律，忽视了宏观层面的整体情况。

基于统计学的研究方法虽然可以简单直观的根据历史先验经验，快速总结出规律，挖掘出路网的关键路段，但是这类基于经验的方法效果无法得到保证。基于网络拓扑结构虽然是研究各种复杂网络特性的经典方法，但是高速公路网络属于稀疏网络，他的网络复杂程度低，整个路网的运行主要还是受到流量的影响。

高速公路本身是一种复杂网络，在意识到传统高速公路关键路段挖掘方法的不足后，我对复杂网络关键路段挖掘进行了研究。

## 2.2 复杂网络关键路段/节点挖掘相关研究

复杂网络的重要节点是指在网路中，相比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特殊节点。近年来，复杂网络中节点重要性排序研究受到越来越广泛的关注，不仅因为其重大的理论研究意义，更因为其强大的实际应用价值。由于不同类型的网络中节点的重要性评价方法各有侧重，且应用领域极广，研究者们从不同的实际问题出发，设计出各种各样的方法。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复杂网络。网络中的节点代表系统的各种构成要素（如高速公路收费站），节点间的连边表示要素之间的联系（如连接两个收费站的路段）。本文最核心的研究问题就是如何识别这些网络中的重要的节点。

在传统的复杂网络关键路段研究中，主要用空间分布、平均距离、连通性、聚类系数、度相关性等参数来度量节点的重要程度。用网络的抗毁性、传播、同步、控制等数据来测试网络的稳定性和完备性。基于复杂网络的关键节点挖掘算法研究较多，分类介绍。

### 2.2.1 基于节点临近

基于节点临近法是最简单直观的复杂网络关键路段挖掘方法，主要包含度中心性、半局部中心性、 $k$ -壳分解法等方法。

度中心性的主要观点是：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。直观说来就是一个节点的邻居数目越多，他的影响力就越大 [9]。假设  $v_i$  是高速公路中的一个节点， $k_i$  即为该节点的度，即为与该节点相连的节点的数目。在含权网络中，节点度定义为与节点相连的边的权重之和。度中心性刻画的是节点的直接影响力，一个节点的度中心性越大，证明该节点能够影响的邻居就越多，改节点就越重要。定义度中心性  $L(i)$ :

$$L(i) = \frac{k_i}{n - 1}$$

式中的  $k_i$  代表节点的度，分母  $n - 1$  代表整体路网的度之和。基于度中心性的关键路段挖掘方法具有简单、直观、计算复杂度低等特点。但是，他仅仅考虑了网络中节点的局部的信息，没有考虑对网络整体的拓扑结构、网络各个节点之间的深入联系。同时也缺乏对宏观层面的考虑。Chen 等人提出了半局部中心性的想法 [10]，首先定义  $N(v_i)$  为节点  $v_i$  的两层邻居度，其值等于从  $v_i$  出发 2 跳（在路网中，直接相连的节点之间距离为 1 跳）内可到达的邻居的数目， $V_i^1$  是距离节点  $v_i$  小于等于 1 跳的解点集合。然后定义  $L(i)$ :

$$L(v_i) = \sum_{v_i \in V_i^1} N(v_i)$$

$v_i$  的局部中心性定义为：

$$F(v_i) = L(L(v_i))$$

半局部中心性将度中心性由 1 跳扩展为 4 跳，不仅考虑了邻居节点的数量，还考虑了网络的聚类影响。在算法上达到了一定的提升。然而，研究表明节点在网络中的位置也是影响节点重要程度的重要因

素。在复杂网络中，一个节点如果处于网络的核心位置，即使它的度中心性非常差，这个节点也往往具有很高影响力；处在边缘的大度数节点影响力往往有限。Kitsak 等人提出一种  $k$ -壳分解法 [11]。这个方法的思路是利用拓扑排序思路，将外围的节点层层剥去，节点存留时间越长，节点的重要性越大。 $k$ -壳分解法计算复杂度低，当网络规模较大时，可以有效的分析网络的层级结构。然而，改方法不能应用于规则网络如树形图、星形图中。同时该方法的排序结果粒度太粗，节点的区分度不大。同时完全不考虑节点的度，显然不合理。Zeng 等人提出了在每一次迭代过程中，剥去最外层节点的同时，考虑节点剩余的邻居数  $x_i$  和节点已经移除的邻居数  $y_i$  的方法 [12]：定义节点  $v_i$  的混合度为  $x_i + z * y_i$ ，不断计算新的混合度值，对网络分层。这种方法能够很好地区分树形图中不同节点的影响力，提高了节点传播能力的区分度。但是这个方法依然只局限于微观层面的结构特征，没有考虑宏观层面的影响。本文研究的是关键路段挖掘，关键路段的方法大都不能直接使用，但是可以给出一定的参考价值。

### 2.2.2 基于路径临近

在通信、交通、社交网络等网络结构中存在着一些很重要的边，这些边是连接几个区域的桥梁，它们在信息包和交通流的传递中担任重要的角色。此时，刻画节点重要性就需要考察网络中节点对信息流的控制力，这种控制力往往与网络中的路径密切相关。基于最短路径的排序方法假设网络中的信息流只经过最短路径传输，而真实的通信网络中必须考虑负载平衡，容错机制，服务水平协议 (SLA) 等 [13]。除了路径长度，路径上的中间节点个数对传播也有不可忽视的影响。一对节点的中间节点会增加这两个节点之间进行互动所需要的消耗。第一，中间节点越多，一对节点之间互动所需要的时间就越长；第二，中间节点相当于在一对进行互动的节点之间引入了“第三方”，这会使

传递的信息失真或者延迟传递。另一方面，从提高网络的可靠性和抗毁性角度看，任意节点对之间的路径数目越多，网络的鲁棒性就越高。此外类似于“桥节点”，程学旗等人提出了刻画网络边重要性的指标用来寻找“桥链路”，相关讨论参见文献 [14]。

在连通网络中，定义  $d_{ij}$  为节点  $v_i$  与  $v_j$  之间的最短路径长度，也称最短距离，一个节点  $v_i$  的离心中心性 (Eccentricity) 为它与网络中所有节点的距离之中的最大值 [15]，即：

$$ECC(i) = \max_j(d_{ij})$$

网络直径定义为网络  $\mathbf{G}$  中所有节点的离心中心性中的最大值，网络半径定义为所有节点的离心中心性值中的最小值。显然，网络的中心节点就是离心中心性值等于网络半径的节点，一个节点的离心中心性与网络半径越接近就越中心。要强调的是，网络直径在复杂网络研究中还有多种不同的定义，例如 Albert 等人在研究万维网的时候定义网络直径为网络中所有节点对的最短路径的平均值 [16]。离心中心性的缺点是极易受特殊值的影响，如果一个节点与大部分节点的距离都很小，只与极小部分节点的距离很大，这个节点的离心中心性仍然会取其中的最大值。接近中心性则采取距离平均值的方式克服了这一缺点。接近中心性 (closeness centrality) 通过计算节点与网络中其他所有节点的距离的平均值来消除特殊值的干扰。一个节点与网络中其他节点的平均距离越小，该节点的接近中心性就越大。接近中心性也可以理解为利用信息在网络中的平均传播时长来确定节点的重要性。平均来说，接近中心性最大的节点对于信息的流动具有最佳的观察视野。对于有  $n$  个节点的连通网络，可以计算任意一个节点  $v_i$  网络中其他节点的平均最短距离：

$$d_i = \frac{1}{n-1} \sum_{j \neq i} d_{ij}$$

$d_i$  越小意味着节点  $v_i$  更接近网络中的其他节点，于是把  $d_i$  的倒数定义为节点  $v_i$  的接近中心性，即：

$$CC(i) = \frac{1}{d_i}$$

上面定义的缺点是仅能用于连通的网络中，文献 [17] 在研究网络效率时对上式进行了改进，使其能够用于非连通网络中，即：

$$EFF(i) = \sum_{j=1}^n \frac{1}{d_{ij}}$$

如果节点  $v_i$  和  $v_j$  之间没有路径可达则定义  $d_{ij}$ ，即  $\frac{1}{d_{ij}} = 0$ 。接近中心性利用所有节点对之间的相对距离确定节点的中心性，在研究中应用非常广泛，但时间复杂度比较高。与接近中心性不同，Katz 中心性不仅考虑节点对之间的最短路径，还考虑它们之间的其他非最短路径 [18]。Katz 中心性认为短路径比长路径更加重要，它通过一个与路径长度相关的因子对不同长度的路径加权。一个与  $v_i$  相距有  $p$  步长的节点，对  $v_i$  的中心性的贡献为  $s^p (s \in (0, 1)$  为一个固定参数)。设  $l_{ij}^{(p)}$  为从节点  $v_i$  到  $v_j$  经过长度为  $p$  的路径的数目。显然  $A^2 = l_{ij}^{(2)} = (\sum_k a_{ik} a_{kj})$ ，其中元素  $l_{ij}^{(2)}$  即从节点  $v_i$  到  $v_j$  经过的边数为 2 的路径的数目，同理我们可以得到  $A^3 A^4 \dots A^p \dots$  将这些值赋予不同权重然后相加，便可以得到一个描述网络中任意节点对之间路径关系的矩阵：

$$K = sA + s^2A^2 + \dots + s^pA^p = (I - sA)^{-1} - I$$

其中， $I$  为单位矩阵。 $K$  矩阵中第  $i$  行  $j$  列对应的元素  $k_{ij}$  实际上就是我们所熟知的节点  $v_i$  和  $v_j$  的 Katz 相似性 [19]。为保证  $K$  可写成公式 (8) 右侧的矩阵形式，要求参数  $s$  小于邻接矩阵的最大特征值的倒数。由此可定义一个节点  $v_j$  的 Katz 中心性为矩阵  $K$  第  $j$  列元素的和：

$$Katz(j) = \sum_i k_{ij}$$

Katz 中心性使用矩阵求逆的方法虽然比直接数路径数目简单，但时间复杂度依然比较高。另一方面，在考虑所有路径长度时，如果节点  $v_i$  与  $v_j$  之间存在长度为  $p$  的路径，在使用  $K$  矩阵计算节点间长度为  $p$  的奇数倍的路径时，这条路径会被重复计算多次。衰减因子  $s$  的引入正好削弱了这些由于重复计算产生的对中心性值的影响，特别是当  $s$  很小时，高阶路径的贡献就非常小了，使 Katz 指标的排序结果接近于局部路径指标。Katz 中心性主要用在规模不太大，环路比较少的网络中。受到 Katz 中心性指标的启发，我们还可以应用其他刻画节点间相似性的指标 [19] 来定义节点中心性。信息指标 (information indices)[20] 通过路径中传播的信息量来衡量节点重要性。该方法假定信息在一条边上传递的时候存在一定的噪音，路径越长噪音就越大。一条路径上的信息传输量等于该路径长度的倒数。一对节点  $(v_i, v_j)$  间能够传输的信息总量就等于它们之间所有路径传输的信息量之和，记为  $q_{ij}$ 。值得注意的是，如果我们把网络看成一个电阻网络，每条边的电阻记为 1，则  $1/q_{ij}$  相当于以 2 个节点  $v_i$  和  $v_j$  为两端点的电阻值 ( $q_{ij}$  相当于电导)[21]，于是我们可以通过计算矩阵  $R = (r_{ij}) = (D - A + F)^{-1}$  获得  $q_{ij}$ ，其中  $D$  是  $n$  阶对角矩阵，对角线元素都是对应节点的度值，非对角线元素为 0， $F$  是每个元素均为 1 的  $n$  阶方阵。由此可得该网络中每一对节点  $(v_i, v_j)$  间通过所有路

径能够传播的信息总量为

$$q_{ij} = (r_{ii} + r_{jj} - 2r_{ij})^{-1}$$

最后，用调和平均数的方法定义节点  $v_i$  的中心性指标（有时也采用算术平均数）[22]：

$$INF(i) = \left[ \frac{1}{n} \sum_j \frac{1}{q_{ij}} \right]$$

信息指标考虑了所有路径，并可通过电阻网络简化繁复的计算过程。该方法可以很容易地扩展到含权网络，也适用于非连通的网络。可见，无论是接近中心性、Katz 中心性还是信息指标，它们的思路是一致的。如果用一个矩阵  $M=(m_{ij})$  来表示网络中所有节点之间的关系， $M$  的每一个元素  $m_{ij}$  刻画了节点  $v_i$  和  $v_j$  之间的某种联系，这个联系既可以是它们之间的距离（如接近中心性），也可以是某种相似性，于是一个节点  $v_i$  的重要性可表示为  $\text{Centrality}(i) = \sum_j m_{ij}$ 。由此可见，只要我们能够给出一种刻画节点关系的方式，就能够基于这个方法定义一个节点的中心性。通常提到的介数中心性 (betweenness centrality) 一般指最短路径介数中心性 (shortest path BC)，它认为网络中所有节点对的最短路径中（一般情况下一对节点之间存在多条最短路径），经过一个节点的最短路径数越多，这个节点就越重要。介数中心性刻画了节点对网络中沿最短路径传输的网络流的控制力。节点  $v_i$  的介数定义为

$$BC(i) = \sum_{i \neq s, i \neq t, s \neq t} \frac{g_{st}^i}{g_{st}}$$

其中， $g_{st}$  为从节点  $v_s$  到  $v_t$  的所有最短路径的数目， $g_{st}^i$  为从节点  $v_s$  到  $v_t$  的  $g_{st}$  条最短路径中经过  $v_i$  的最短路径的数目。显然，当一

个节点不在任何一条最短路径上时，这个节点的介数中心性为 0，比如星形图的外围节点。对于一个包含  $n$  个节点的连通网络，节点度的最大可能值为  $n-1$ ，节点介数的最大可能值是星形网络中心节点的介数值：因为所有其他节点对之间的最短路径是唯一的并且都会经过该中心节点，所以该节点的介数就是这些最短路径的数目，于是得到一个归一化的介数：

$$BC'(i) = \frac{2}{(n-1)(n-2)} \sum \frac{g_{st}^i}{g_{st}}$$

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。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 [23]：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 [24, 25]。介数中心性仅考虑网络流通过最短路径传输。Yan 等人 [26] 的研究指出，如果选择最短路径来运输网络流，很多情况下反而会延长出行时间、降低出行效率。把一对节点之间的每条路径看作一条单独的管道，一条管道能够传输一个单位的网络流，从源节点  $v_s$  到目标节点  $v_t$  的最大流量是指  $v_s$  与  $v_t$  之间所有管道可同时运输的网络流的总和（实际上，这种假设没有实际意义，多条路径往往有重合的部分，重合部分的流量就会超过假设的情况）。基于这样的假设，流介数中心性 (flow betweenness centrality) [1] 认为网络中所有不重复的路径中，经过一个节点的路径的比例越大，这个节点就越重要。由此得到节点  $v_i$  的流介数中心性为

$$FBC'(i) = \sum_{s < t} \frac{g_{st}^i}{g_{st}}$$

介数中心性和流介数中心性考虑的是两个极端，前者只考虑最短路径，后者考虑所有路径并认为每条路径作用相同，接下来介绍两种介于两者之间的介数中心性算法。首先介绍随机游走介数中心性：从源节点  $v_s$  到目标节点  $v_t$  的随机游走的过程中当  $i=s$  或者  $t$  的时候， $I_{st}^s = I_{st}^t = 1$ 。该方法计算复杂度较高。路由介数中心性：计算机网络中，每个路由器都有一个包含很多行记录的路由表，每行记录存储着要到达的目标地址及下一跳地址。显然，每个路由器只记录了局部的网络结构信息。对网络中的每一对节点  $(v_s, v_t)$ ，将分布在各个路由器中的信息聚合，可形成一个关于这一对节点的有向无环图  $R(s, t)$ 。定义  $p(s, u, v, t)$  为有向无环图  $R(s, t)$  中节点  $v_u$  转发给节点  $v_v$  一个从源节点  $v_s$  到目标节点  $v_t$  的信息包的概率。如果  $p(s, u, v, t) > 0$ ，则在  $R(s, t)$  中存在一条从  $v_u$  指向  $v_v$  的有向边。用  $k_{s,t}^{(u)}$  表示信息包从  $v_s$  到  $v_t$  的传递过程中，经过节点的  $v_u$  概率，显然  $k_{s,t}^{(s)} = k_{s,t}^{(t)} = 1$ ，用  $Pred_{s,t}^{(v)}$  表示。那么有向无环图  $R(s, t)$  中经过任意一个节点  $v_v$  的概率可由下式得出：

$$K_{s,t}(v) = \sum_{u \in \text{Pred}_{s,t}(v)} K_{s,t}(u) p(s, u, v, t)$$

我们考虑经过节点的路径为一个封闭环的时候，就可以定义子图中心性 (subgraph centrality)[27]。该方法从全局的视野考察了网络中所有可达的邻居对节点中心性的增强作用，并且认为增强作用会随距离的增加而衰减。与图论中的概念有所不同，这里一个子图特指从一个节点开始到这个节点结束的一条闭环回路。一个节点  $v_i$  的子图数目就是以该节点为首尾的闭环回路的个数。子图中心性认为闭环回路

的路径长度越小，回路信息交流越便利，节点之间的联系越紧密，对节点的中心性贡献越大，其定义为

$$SC(i) = \sum_{t=1}^{\infty} \frac{a_{ii}^t}{t!}$$

其中  $a_{ii}^t$  为网络的邻接矩阵  $A$  的  $t$  次幂的第  $i$  个对角线元素。 $t=1$  时， $a_{ii}^1 = 0$ ； $t=2$  时， $a_{ii}^2$  为节点  $v$  的度值，即  $a_{ii}^2 = k$ ，此时，子图中心性就等价于度中心性； $t > 2$  时， $a_{ii}^t = k$  表示从点  $v_i$  开始，经过  $t$  条边又回到  $v_i$  的路径的数目。子图中心性赋予较短的回路较高的权重，使得节点的度在其中发挥较大作用的同时，还考虑了高阶回路。在实际应用时，根据具体计算需求， $t$  可以取到任意值截断。子图中心性用邻接矩阵特征值和特征向量可表示为

$$SC(i) = \sum_{t=1}^{\infty} \frac{a_{ii}^t}{t!} = \sum_{j=1}^N (K_i^j)^2 e^{l_j}$$

其中， $l_j$  为邻接矩阵  $A$  的特征值， $K_j$  是  $l_j$  所对应的特征向量， $K_i^j$  表示特征向量的第  $i$  个元素。有些情况下，度中心性，接近中心性以及介数中心性都不能区分网络中某些节点谁更重要时，可用子图中心性来对这些节点进行更加细致地区分 [27]。另外，子图中心性的方法还能够应用于网络中模体的检测 [27]。

### 2.2.3 基于特征向量的排序方法

前面介绍的方法都是从邻居的数量上考虑对节点重要性的影响，基于特征向量的方法不仅考虑节点邻居数量还考虑了其质量对节点重要性的影响。下面将详细介绍 7 种方法。其中前两种方法，即特征向量中心性和累计提名方法一般用在无向网络中，后者收敛更快。后面五种方法可看成特征向量中心性在有向网络中的应用。PageRank

算法和 LeaderRank 算法通过模拟用户上网浏览网页的过程，使节点的分值沿着访问路径增加，用于识别网页重要性。实验结果显示，LeaderRank 表现好于 PageRank 算法。HITS 算法、自动信息汇集算法，SALSA 算法中考虑节点的双重角色：权威性和枢纽性，并认为两者相互影响。本类方法在理论和商业上都受到了极大的关注，很有借鉴意义。

特征向量中心性 (eigenvector centrality)[9] 认为一个节点的重要性既取决于其邻居节点的数量 (即该节点的度)，也取决于每个邻居节点的重要性。记  $x_i$  为节点 v 的重要性度量值，则：

$$EC(i) = x_i = c \sum_{j=1}^n a_{ij} x_j$$

特征向量中心性更加强调节点所处的周围环境 (节点的邻居数量和质量)，它的本质是一个节点的分值是它的邻居的分值之和，节点可以通过连接很多其他重要的节点来提升自身的重要性，分值比较高的节点要么和大量一般节点相连，要么和少量其他高分值的节点相连。从传播的角度看，特征向量中心性适合于描述节点的长期影响力，如在疾病传播、谣言扩散中，一个节点的 EC 分值较大说明该节点距离传染源更近的可能性越大，是需要防范的关键路段。特征向量法完全用与某节点相连接的其他节点的信息来评价该节点的重要性。Bonacich 等人 [28] 认为节点的重要性还可能受到不依赖于节点连接信息的一些来自外部的信息的影响。例如在微博上有人喜爱转发其他人发布的信息 (依赖于网络连接的内部信息)，有的却比较热衷于发布原创信息或从其他网站转发一些信息 (不依赖于网络连接的外部信息)。由此 Bonacich 等人提出阿尔法中心性 (Alpha-centrality)，即  $x = \alpha Ax + e$ ，其中  $\alpha$  为刻画来自网络内部连接影响的内因参数， $e$  为刻画那些不受网络连接影响的外因参数。不失一般性， $e$  可以设

置为一个所有元素都等于 1 的向量，此时阿尔法中心性与 Katz 中心性一致。当网络中有一些度特别大的节点的时候，特征向量中心性会出现分数局于化现象 (Localization)，即大多数分值都集中在大度节点上，使得其他节点的分值区分度很低。为了避免这一现象，Martin 等人 [29] 对特征向量中心性进行改进，提出在计算节点  $v_i$  的分值时，求和中其邻居的分值不再考虑节点  $v_i$  的影响。特征向量中心性中，一个节点的打分值完全由邻居决定，收敛过程缓慢。此外，当不存在一个正的自然数  $t$ ，使得转移矩阵的  $t$  次幂所有元素都是正的时，节点打分值会出现周期性循环，不能收敛。为了使打分值能够收敛并且快速收敛，累计提名 (cumulative nomination) [30] 方法在每次迭代过程中，同时考虑邻居节点和自身的打分值。设  $p_i^t$  为节点  $v_i$  在时刻  $t$  时得到的提名次数，假设  $t=0$  时每个节点都获得 1 次提名 (即  $p_i^0 = 1$ )，每个时间步每个节点从所有有相邻的节点处获得新增的提名，新增的提名数为邻居节点已有的提名数的总和。于是定义节点在  $t+1$  时刻的累积提名数为

$$p_i^{t+1} = p_i^t + \sum_j a_{ij} p_j^t$$

如果所有节点归一化后的提名次数不再变化，则停止迭代。稳态时每个节点的提名次数占所有节点的提名次数的比例就是其重要性权值。特征向量中心性算法在每次迭代的时候，一个节点  $v_i$  的中心性值完全等于邻居的中心性值之和，而累计提名算法则保留了节点  $v_i$  上一步的中心性值，实验结果显示累积提名相比原始的特征向量中心性收敛速度更快。累积提名和 Alpha 中心性在数学形式上非常相似，但 Alpha 中心性中的  $e$  是固定值，即每次迭代的时候不变，而累积提名中添加的是上一时间步的打分值，这个打分值会随着每步更新变化。

特征向量中心性及其变体应用广泛，例如网页排序领域中最著名

的 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  $v_i$  在 t 时刻的 PR 值为

$$PR_i(t) = \sum_{j=1}^n a_{ji} \frac{PR_j(t-1)}{k_j^{out}}$$

迭代到每个 PR 值都达到稳定时为止。公式的缺陷在于 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：

$$PR_i(t) = (1 - c) \sum_{j=1}^n a_{ji} \frac{PR_j(t-1)}{k_j^{out}} + \frac{c}{n}$$

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 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[31] 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 [32] 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[33] 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 [34]、对社交网络上用户的排序 [35]、对风投公司 (VC) 的排序 [36]、对科学论文的排序 [37] 以及科学家影响力排序 [38] 等。

#### 2.2.4 基于节点移除和收缩的排序方法

节点 (集) 的移除和收缩方法与系统科学中确定一个系统的核心的思路暗合，其最显著的特点是在重要节点排序的过程中，网络的结构会处于动态变化之中，节点的重要性往往体现在该节点被移除之后对网络的破坏性。从衡量网络的健壮性角度看，一些节点一旦失效或移除，网络就有可能陷入瘫痪或者分化为若干个不连通的子网。实际生活中的很多基础设施网络，如输电网、交通运输网、自来水-天然气供应网络等，都存在“一点故障，全网瘫痪”的风险。为了预防风险，研究人员提出了很多方法来研究节点收缩或者移除之后网络的

结构与功能的变化，从而为新系统的设计与建造提供依据。比较典型的是系统的“核与核度”理论。许进等人在定义规则网络图的核概念基础上，提出了核度的测量方法，研究了网络核度与节点数、边数的关系，并根据它们之间的关系设计了规则网络构造定理；李鹏翔等人认为直接的联系往往是间接联系的必经之路，在评估节点重要性的过程中更加重要，用节点集被删除后形成的所有不直接相连的节点对之间的最短距离的倒数之和来反映节点删除对网络连通的破坏程度；陈勇等人分析了通信网络，考察去掉节点（集）及其相关边后所得到的图的生成树的数目，数目越小，表明该节点（集）越重要；谭跃进等人用收缩节点方法替代删除节点法，综合考虑了节点的度以及经过该节点的最短路径的数目，将节点收缩后网络的聚集度作为节点重要性评估的标准。系统科学的方法给我们提供了新的视角，但由于计算复杂度较高，目前这类方法还仅限于小规模的网络实验。此外，Restrepo 等人 [39] 提出通过考察网络最大特征值在移除节点后的变化来衡量节点重要性的方法，该方法还可以应用于刻画网络连边的重要性。

破坏性反映重要性。节点删除的最短距离法 [40] 认为一个节点移除后的破坏性与所引起的距离变化有关：移除一个节点（集）会引起网络分化，并形成若干个连通分支，网络中节点对之间较短距离的变化越大，被移除的节点就越重要。该算法区别对待不同长度的路径，认为“相对直接的、近距离的联系所造成的破坏性大于相对间接的、远距离的联系所造成的破坏性”[40]。具体地，在连通图中一个节点被删除之后，对网络的整体状况的影响体现在两个方面：直接损失和间接损失。直接损失是指被删除的节点与其他剩余的节点之间不再存在通路，如果连通网络中共有  $n$  个节点，删除一个节点后产生的不连通节点对的数目为  $n-1$ 。如果删除的是节点集，直接损失还应该包括删除的节点集内节点之间的不再连接的损失。间接损失是指删除一个节点造成剩余节点之间不连通而引发的损失：用  $N_k$  ( $k=1, 2, \dots, s$ ) 表

示一个节点  $v_i$  被删除后，网络分化成的  $s$  个连通子图中第  $k$  个连通子图的节点数，则该节点被删除后所形成的不再连通的节点对的数目为  $\sum_{t=1}^s \sum_{r=t+1}^s N_t N_r$ ，记由于删除节点  $v_i$  造成的不再相连的节点对表示为集合  $E$ （包括直接损失和间接损失两部分），那么节点  $v_i$  的重要性等于集合  $E$  中节点对之间的最短距离的倒数之和，即：

$$DSP(i) = \sum_{(j,k) \in E} \frac{1}{d_{jk}}$$

$d_{jk}$  为删除节点  $v_i$  之前  $v_j$  与  $v_k$  间的最短距离。注意，当  $j$  或  $k=i$  的时候，相当于直接损失；当  $j \neq k \neq i$  的时候，相当于间接损失。节点删除的最短距离法在衡量一些节点集的重要性方面优势比较突出。在实际的大规模网络中，仅删除一个节点时网络的拓扑图一般不会分化为几个连通子图，网络的间接损失为 0，节点删除的最短距离法效果并不明显。而如果同时删除多个节点，则很容易使网络不再连通，这时该方法的优越性就显现出来了。

在通信网络中，节点删除后网络中节点对之间最短距离会发生变化，但一般对网络时延影响不大，用最短距离法不一定准确。这时可通过考察节点删除后网络拓扑图的生成树个数来衡量节点的重要性。在图论中，一个图的树是该图的一个连通的无环子图，一个图的生成树定义为拥有该图的所有顶点的树。节点删除的生成树法 [41] 认为一个节点删除后对应的网络的生成树的数目越少，该节点越重要。给定一个无向连通图，其邻接矩阵为  $A$ ，网络拉普拉斯矩阵  $L=D-A$ （将矩阵  $A$  主对角线上的元素  $a_{ii}$  替换为节点  $v_i$  的度值，非对角线上的元素值全部乘以-1）。那么，这个连通无向图的生成树个数  $t_0$  为矩阵  $L$  的任意一个元素  $l_{pq}$  的余子式  $M_{pq}$  的行列式，即： $t_0 = \|M_{pq}\|$ 。删除任意一个节点  $v_i$ ，网络的邻接矩阵变为  $A_{-i}$ ，然后用上面的方法计算网络的生成树个数为  $t_{-i}$ 。由此可定义节点  $v_i$  的中心性指标为

$$DST(i) = 1 - \frac{t-i}{t_0}$$

在节点的移除对网络的连通性影响不大的网络中，节点删除的生成树法优于最短距离法。但节点删除的生成树法有一些缺点，例如，只能用在连通网络中。若一个节点删除后网络变得不再连通，这些节点的重要性就难以判断了，这时可采用节点收缩法评估节点的重要性。

节点收缩就是将一个节点和它的邻节点收缩成一个新节点 [42]。如果  $v_i$  是一个很重要的核心节点，将它收缩后整个网络将能更好地凝聚在一起。最典型的就是星形网络的核心节点收缩后，整个网络就会凝聚为一个大节点。从社会学的角度讲，社交网络中人员之间联系越方便（平均最短路径长度  $d$  越小），人数越少（节点数  $n$  越小），网络的凝聚程度就越高。因此定义网络的凝聚度为

$$A[G] = \frac{1}{nd} = \frac{1}{\sum_{i \neq j} d_{ij}} = \frac{n-1}{n \frac{\sum_{i \neq j} d_{ij}}{n(n-1)}}$$

可见，节点收缩法中节点的重要程度由节点的邻居数量和节点在网络路径中的位置共同决定。由于每次收缩一个节点，都要计算一次网络的平均路径长度，时间复杂度比较高，不适于计算大规模网络。

为了研究网络的抗毁性，Dangalchev[43] 提出了残余接近中心性 (residual closeness centrality)，用来衡量节点的移除对网络带来的影响。残余接近中心性认为若一个节点的删除使得网络变得更加脆弱，该节点就越重要。文献 [43] 对接近中心性的改进使得接近中心性应用的范围从连通图扩展到了非连通图。该方法对接近中心性进行了改进，分母取以 2 为底的指数，相当于提升了短路径的影响力，同时会使本算法更易计算和扩展 (文献 [43] 给出了将几个图合并为一个图计算接

近中心性的详细算法)。在移除一个节点  $v_i$  之后, 定义其残余接近中心性为

$$RCC(i) = \sum_j \sum_{k \neq j} \frac{1}{2^{d_{jk}(-i)}}$$

其中  $d_{jk}^{(-i)}$  为删除节点  $v_i$  之后, 节点  $v_j$  与  $v_k$  的最短距离。残余接近中心性在测度网络的脆弱性方面比图坚韧度 (graph toughness)、离散数 (scattering number)、节点完整度 (vertex intergrity) 等方法表现要好。基于该方法可以定义出边的残余接近中心性和节点集、边集的残余接近中心性。

### 2.2.5 节点重要性排序方法的评价标准

根据评价标准的不同又分为用网络的鲁棒性和脆弱性评价排序算法、用传播动力学模型评价排序算法。网络科学研究的早期, 所关注的网络中节点数目较少, 典型的有同性恋接触网络 [40, 44]、女生用餐伙伴选择网络 [45]、空手道俱乐部网络 [20] 等, 对于这些小规模网络, 可以通过调查问卷等方式对每个节点的重要性进行打分, 然后将实际的调查结果作为标准与其他算法结果进行比较, 分析各种方法的表现和优劣。随着科技的发展和进步, 大数据时代已经来临, 现在我们所面对的网络规模迅速增长, 想要得到一个对所有节点的重要性的较为客观的评价标准极为困难。目前评价各种排序算法优劣的主要思路是: 将排序算法得出的重要节点作为研究对象, 通过考察这些节点对网络某种结构和功能的影响程度、对其他节点状态的影响程度来判断排序是否恰当。例如, 如果一个排序算法得出节点  $v_i$  比  $v_j$  更重要, 单独考察  $v_i$  比  $v_j$  发现前者对网络的结构功能或对其他节点的影响程度更大, 就说明这种排序算法比较符合实际。常用来评价各排序算法的方法有基于网络的鲁棒性和脆弱性方法以及基于网络的传播

动力学模型的方法。下面分别对这两类方法进行简单的介绍。

### 用网络的鲁棒性和脆弱性评价排序算法

本类方法着重考察网络中一部分节点移除后网络结构和功能的变化，变化越大移除的节点越重要。用某一种重要节点挖掘方法将网络中所有节点按重要性进行排序，然后按重要性从大到小的顺序，将一部分节点从网络中移除，用  $k(i/n)$  表示移除  $i/n$  比例的节点后，网络中属于巨片 (giant component)[46] 的节点数目的比例，网络的鲁棒性 (robustness) 可用 R-指标刻画 [47]:

$$R = \frac{1}{n} \sum_{i=1}^n k(i/n)$$

显然，不论对何种算法，星形图中，R 取最小值  $(1/n - 1/n^2)$ ，完全图中 R 取最大值  $(1-1/n)/2$ ，当 n 比较大时  $R \in (0, 1/2)$ 。可定义  $V=1/2-R$  来表示网络对于所实施的移除方法的脆弱性 (vulnerability)，可见，V-指标越大表示采用该方法进行攻击的效果越好。V-指标和 R-指标可从整体上反应各种重要节点挖掘方法的有效性。另外也可画出  $i/n$  与  $k(i/n)$  在二维坐标上的曲线，对节点移除的影响进行详细分析。例如文献 [48] 中考察了在无标度网络中使用 4 种排序方法移除节点后对网络最大连通集的影响，这 4 种方法包括度中心性、介数中心性、接近中心性和特征向量中心性，并和随机移除节点的方法进行比较。用于实验的无标度网络节点数为  $n=10000$ ，平均度为 4，移除节点时采用同时移除的方法。

### 用传播动力学模型评价排序算法

复杂网络传播研究的对象极广 [49]，比如通信网络中的病毒传播 [50]、社会网络中的信息传播 [51]、电力网络中的相继故障 [1]、经济网络中的危机扩散等 [52]。在评价各种节点重要性挖掘方法时广泛采

用的是传染病模型，主要包括 SIS 模型 [11] 和 SIR 模型 [53]。在 SIS 模型中一个节点的传播能力被定义为稳态下该节点被感染的概率；在 SIR 模型中，一个节点的传播能力被定义为该节点的平均传播范围。下面简要介绍 SIR 模型及一个应用的例子。SIR 假设网络中的节点有三个状态：易染态 S (susceptible，可被处于感染态的邻节点感染)，感染态 I (infected，处于 I 态的节点一定时间后会变为免疫态)，免疫态 R (recovered，免疫态的节点不会被感染，也不会传播病毒)。SIR 模型有单点接触和全接触两种 [54]，前者指在每一时间步内，处于 I 态的节点感染其邻居的时候将随机选择一个 S 态的邻居，然后以概率  $p$  使其由 S 态变为 I 态；后者指处于 I 状态的节点感染邻居的时候选择的是所有 S 态的邻居，每个 S 状态的邻居都有机会以概率  $p$  转变为 I 态。设置一个(组)节点为初始感染节点(即处于 I 态)，观察每一时间步网络中感染过的节点数目和最终稳定态时(没有 I 态的节点时)感染过的节点数目，可通过病毒的传播速度和范围两个方面来考察节点的真实影响力。要对比两种重要节点挖掘方法的优劣，可分别用这两种方法对网络中的节点按重要性进行排序，取相同数目的最重要的节点设为初始感染态，用 SIR 模型在网络上进行实验，如果一个排序方法的结果使得网络传播地又快又广，则说明该重要节点排序方法优于其他方法。例如文献 [55] 中应用 SIR 模型比较了 LeaderRank 算法和 PageRank 算法的排序结果。图 5 显示了使用两种方法获得的前 20 个(图 5(a))最重要的节点中，以不同的节点为初始感染源进行 SIR 传播的过程。可见，以 LeaderRank 获得的节点为初始感染源的传播又快又广，说明 LeaderRank 算法比 PageRank 算法更能够识别网络中传播影响力高的节点。图 5(b) 为考虑前 50 个节点的情况。需要注意的是，网络中信息传播和病毒传播有很大的不同。文献 [50] 深入比较了信息传播与病毒传播的不同，提出了网络中的信息传播模型。文中还全面总结了影响网络流在网络中传播速度和快慢的 7 种因素，比如边

的强度、信息内容、传播者的角色、记忆效应、时间延迟效应等。因此，在评价节点信息传播影响力的时候，例如社交网站上意见领袖挖掘，应该考虑更加符合实际传播方式的模型。

### 2.3 复杂网络社群划分相关研究

现有的研究主要分布于普通社区挖掘方法和重叠社区挖掘方法。2002 年 Girvan 和 Newman 引提出社区挖掘的概念。现实世界中的许多复杂系统或以复杂网络的形式存在、或能被转化成复杂网络。例如：社会系统中的人际关系网、科学家协作网和流行病传播网，生态系统中的神经元网、基因调控网和蛋白质交互网，科技系统中的电话网、因特网和万维网等等。复杂网络普遍存在着一些基本统计特性，如反映复杂网络具有短路径长度和高聚类系数之特点的“小世界效应”；又如表达复杂网络中结点之度服从幂率分布特征的“无标度特性”；再如描述复杂网络中普遍存在着“同一社区内结点连接紧密、不同社区间结点连接稀疏”之特点的“社区结构特性”[]。目前，关于复杂网络基本统计特性的研究已吸引了不同领域的众多研究者，复杂网络分析已成为最重要的多学科交叉研究领域之一。

随着应用领域的不同，社区结构具有不同的内涵。譬如，社会网中的社区代表了具有某些相近特征的人群、生物网络中的功能组揭示了具有相似功能的生物组织模块、Web 网络中的文档类簇包含了大量具有相关主题的 web 文档、交通网络中的集群区段等等。近 10 年来，已有很多复杂网络社区挖掘方法被提出，它们分别采用了来自物理学、数学和计算机科学等领域的理论和技术，就其依据的原理可分为基于划分、基于模块性优化、基于标签传播、基于动力学和基于仿生计算的方法等。2002 年，Girvan 和 Newman 提出了最著名的社区挖掘方法 GN(Girvan Newman)。该算法采用的启发式规则为：社区间链接的边介数 (edge betweenness) 应大于社区内链接的边介数，其中

每个链接的边介数被定义为“网络中经过该链接的任意两点间最短路径的条数”。算法 **GN** 通过反复计算边介数，识别社区间链接，删除社区间链接，以自顶向下的方式建立一棵层次聚类树 (dendrogram)。该算法最大的缺点是计算速度慢。2003 年，Tyler 等人将统计方法引入算法 **GN** 中，提出一种近似的 **GN** 算法。他们的策略是：采用蒙特卡洛方法估算出部分链接的近似边介数，而不去计算全部链接的精确边介数。2004 年，Radicchi 等人提出了用链接聚类系数 (link clustering coefficient) 取代算法 **GN** 中链接的边介数。他们认为：社区间链接应该很少出现在短回路 (如三角形或四边形) 中，否则短回路中的其他多数链接也会成为社区间链接，从而显著增加社区间的链接密度。2004 年，Newman 和 Girvan<sup>[J]</sup> 提出了一个用于刻画网络社区结构优劣的量化标准，被称之为模块性函数 **Q**。该算法中候选解的搜索策略为：选择并合并两个现有的社区。初始化时，候选解中每个社区仅包含一个结点；在每次迭代时，算法 **FN** 选择使函数 **Q** 值增加最大 (或减小最少) 的社区对进行合并；当候选解只对应一个社区时算法结束。通过这种自底向上的层次聚类过程，算法 **FN** 输出一棵层次聚类树 (denogram)，然后将对应的函数 **Q** 值最大的社区划分作为最终聚类结果。2005 年，Guimera 和 Amaral<sup>[K]</sup> 胡提出了基于模拟退火的模块性优化算法 (simulated annealing, **SA**)。该算法首先随机生成一个初始解；在每次迭代中，在当前解的基础上产生一个新的候选解，由函数 **Q** 判断其优劣，并采用模拟退火策略中的 Metropolis 准则决定是否接受该候选解。**SA** 算法产生新候选解的策略是：将结点移动到其他社区、交换不同社区的结点、分解社区或合并社区。该算法具有非常好的聚类质量，但其缺点是运行效率低。2006 年，Newman<sup>[L]</sup> 朝将谱图理论引入模块性优化中。2008 年，Blondel 等人<sup>[M]</sup> 提出了快速模块性优化方法 (fast unfolding algorithm, **FUA**)。该算法结合了局部优化与多层次聚类技术。2007 年，Raghavan 等人<sup>[N]</sup> 提出了著名的标签传播算

法 (label propagation algorithm, LPA). 该算法的流程为：初始化时，为每个结点赋一个唯一标签；每次迭代中，每个结点采用大多数邻居的标签来更新自身标签；当所有结点的标签都与其多数邻居的标签相同时，算法结束。2008年，Tib61y 等人 [1] 发现标签传播算法 LPA 等价于最小化哈密尔敦函数，2009年，Leung 等人 [2] 朝将算法 LPA 作为分析大规模在线社会网的工具。他们通过研究算法 LPA 的优势和限制，讨论了其扩展和优化方面的一些问题，进而对算法 I. PA 进行了修正。2009年，Barber 等人 [3] 朝将算法 LPA 等价为一个优化问题，并给出对应的目标函数 20lo 年，I. iu 等人 [26] 发现算法 LPAm 得到的社区划分具有“每个社区内结点的度之和都相似”的特性，就是说该算法有陷入局部最优解的倾向。为跳出局部最优解，他们给出一种多步层次贪婪算法 (muhistep greedy agglomerative algorithm, MSG)，每次可合并多个社区对。进而他们将算法 LPAm 与 MSG 相结合，提出了一个基于模块性优化、层次化标签传播算法 I. PAm+，使标签传播类算法的聚类性能得到进一步改善。2000年，van Dongen 比刊提出了 Markov 聚类算法 (Markov cluster algorithm, MCI. )。该算法主要是基于 Markov 动力学理论，通过改变和调节 Markov 链呈现出网络社区结构。2007年，杨博等人 [30] 针对符号网络社区挖掘问题(包括正负权值的网络)，提出了基于 Markov 随机游走模型的启发式社区挖掘算法 (finding and extracting communities, FEC)。2008年，Rosvall 等人 [1] 提出了映射平衡算法 infomap。该方法基于最小描述长度 (MDL) 原理 [14]，通过信息传播扩散技术探测网络社区结构。2011年，Morfirescu 等人 [1] 副研究了一类离散时间的多 agent 系统，基于信任度衰减的观点建立动力学模型。他们将复杂网络视为一个 agent 网络，其中每个 agent 拥有一个信念值。2012年，杨博等人 [34] 给出了一个采用 Markov 转移矩阵的特征值来评估亚稳态之进出时间的方法，揭示了网络内在属性与社区结构的数学联系，提出了分析复杂网络社区结构的谱理论。

基于此，定义了 3 个刻画社区结构的量，分别为社区之间的分离度、每个社区的凝聚度和刻画社区结构的谱特征。2007 年，Liu 等人 [1] 基于每个蚂蚁个体的行为，提出了一个用于探测邮件社会网社区结构的蚁群聚类算法。2009 年，Sadi 等人 [71] 采用蚁群优化技术发现网络中的团，并将这些团视为新结点而构建一个简化网络，然后通过传统社区挖掘算法来探测社区结构。2010 年，刘大有等人 [2] 列从仿生角度出发提出一个基于 Markov 随机游走的蚁群算法 (ant colony optimization based on random walk, RWACO)。RWACO 将蚁群算法框架作为基本框架。以 Markov 随机游走模型作为启发式规则，通过集成学习的思想将蚂蚁的局部解融合为全局解，并用其更新信息素矩阵。通过“强化社区内连接，弱化社区间连接”这一进化策略逐渐呈现出网络的社区结构。



## 第三章 基于贪心的高速公路关键路段识别模型

本章主要分四个部分。第一节讲述了关键路段挖掘的意义，分析了现有研究的不足；第二节定义了高速公路关键路段挖掘模型，并给出次模性证明；第三节针对不同的挖掘方法进行实验，从多个角度分析本文方法的效果；第四节对本章内容做了一个小结。

### 3.1 引言

对于交通运输 [56]、水利传输 [57]、能源和通信等基础设施系统，在遭遇自然灾害或者人为灾害时，会对整个系统的性能造成显著的影响，带来重大的经济损失。所以在发生事故或者自然灾害的时候，维护这些网络的完整性至关重要。

灾难管理是一个多阶段的过程，从防灾减灾和准备，着眼于长期消除或降低风险的措施，延伸到灾后响应、恢复与重构。投资基础设施系统在缓解中起着至关重要的作用活动，它可以增强链接的稳定性。但是，将所有的路段稳定性都增强到坚不可摧，在管理人员看来是十分浪费的。本章节主要研究如何在有限的资源下，找到可以最大化网络通行效率的关键路段集合。将资源布置到这个关键路段集合中，最大化提升关键路段通行效率，增加路网稳定性，实现事故前的预防，事故后的快速恢复。

本章主要研究如何对高速公路关键路段挖掘问题进行建模，并且围绕安徽、山西、北京的收费站车辆数据，进行真实数据集上的实验。

## 3.2 问题定义

### 3.2.1 问题定义

高速公路具有成网性，给定一个有向图  $G = \{V, E\}$ ，其中  $V$  代表收费站（节点）的集合； $E$  表示高速公路中路段的集合。对于经过高速公路的车辆，定义  $O$  为车辆的出发节点， $D$  作为车辆的目标节点。定义  $P_e$  ( $0 \leq P_e \leq 1$ ) 为路段的损毁率，这个概率通过历史上的高速公路路段损毁事件得到，同时可以随着交通管理者对路段进行管理而减小。定义管理者的决策向量  $y=\{y_1, y_2, \dots, y_n\}$ ， $y$  是一个  $n$  维向量，每一维  $y_i$  的数值取 0 或 1，1 表示这条路段属于关键路段集合，管理者会对其进行维护和投资管理，改善路段状况；0 表示非关键路段。基于每一条路段都有一定概率损毁，定义  $C_{e_i}$  表示第  $i$  个路段的状态。当  $C_{e_i}$  等于 1 时，路段保持完好，当  $C_{e_i}$  等于 0 时，路段因为事故损毁。定义  $c=\{C_{e_1}, C_{e_2}, \dots, C_{e_n}\}$ ， $c$  表示路网的某一种拓扑结构， $C$  表示路网的所有拓扑结构的集合。对于行驶在高速公路上的车辆，定义车辆的出行时间为  $X_i$ ，这个出行时间由车辆的路径选择、出行时途径路径的车流密度决定。定义当高速公路路段断裂严重，车辆无法抵达目的地时，车辆的出行时间定为常量  $M$ 。 $M$  的大小在一定程度上代表了路网连通性的权重， $M$  越大，高速网络的连通性就越重要。为了更好的求解目标函数，在此提出两个假设：

1) 路段之间的损毁概率相互独立：假设处于静态模型下，所有路段都有一定的概率损毁，这些损毁率之间没有相互影响。传统研究网络可靠性的相关文献 [3] 都是基于这个假设所做的研究。

2)  $M > \text{Max}(X_i)$ :  $M$  必须要大于连通路网中的最大出行代价。当车辆在路段中找不到一条可以抵达终点的路径时，定义车辆此次出行的代价要绝对大于路段仍然连通情况下的任意时间。即默认断裂对路网造成的影响大于路段仍然连通的情况。

根据高速公路的历史事故数据，通过结构分析和统计调查 [23]，确定路段损毁的概率。使用该概率作为本研究的先验概率。高速公路建设管理者可以通过在高速路段上建立基础设施，投放人力资源等方式管理路段，增强路段的稳定性。假设路段  $e$  以概率  $P_e$  损毁，以概率  $(1 - P_e)$  保持完好。基于路段的损毁率，我们可以计算路网拓扑结构概率矩阵  $Z$ ：

$$\begin{matrix} C_{e_1}^1 & C_{e_2}^1 & \cdots & C_{e_n}^1 & P^1 \\ C_{e_1}^2 & C_{e_2}^2 & \cdots & C_{e_n}^2 & P^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ C_{e_1}^m & C_{e_2}^m & \cdots & C_{e_n}^m & P^m \end{matrix}$$

矩阵中， $C_{e_i}$  表示第  $i$  条路段的状态，0 表示遭遇事故，已经损毁，1 表示完好无损；对于每一行来说，前  $n$  项  $C^j = \{C_{e_1}^j, C_{e_2}^j, \dots, C_{e_n}^j\}$  表示路网的拓扑结构，全 0 表示全部路段断裂，路网瘫痪；全 1 表示路网完全连通。第  $n+1$  项  $P^j$  表示高速公路网络拓扑变成这个拓扑结构的概率。其中  $P^j = \prod_{i=1}^n (P_{e_i} C_{e_i}^j + (1 - P_{e_i})(1 - C_{e_i}^j))$ 。在交通管理者选取关键路段，并进行一定的决策处理后，路段的损毁概率发生变化，进而概率矩阵  $Z$  也同时会发生变化。

### 3.2.2 模型定义

在此提出关键路段挖掘模型：

$$L(\mathbf{y}) = -E(T(\mathbf{c}|\mathbf{y})) \quad (3.1)$$

其中， $T(\mathbf{c}|\mathbf{y})$ ：

$$T(\mathbf{c}|\mathbf{y}) = P(K|\mathbf{c}) \sum_{k \in K} X_k \quad (3.2)$$

$\mathbf{y}$  是一个  $n$  维向量，表示关键路段集合，即管理者想要投资维护的路段。 $T(\mathbf{c}|\mathbf{y})$  表示当关键路段集合为  $\mathbf{y}$ ，路网拓扑结构为  $\mathbf{c}$  的时候，高速公路的整体通行时间。式3.1中对路网通行时间的期望取负，转化为通行效率。模型的目标是选取出关键路段，对这些关键路段增加投入，使得在同样的投入情况下，整个路网的通行效率得到最大的提升。结合式3.1，式3.2，得到展开式：

$$\text{Max}(L(\mathbf{y})) = -\text{Min}_{\mathbf{y}} \sum_{\mathbf{c} \in C} P(\mathbf{c}|\mathbf{y})P(K|\mathbf{c}) \sum_{k \in K} X_k \quad (3.3)$$

式中  $\mathbf{y}$  表示关键路段集合，假设高速公路网络的路段数量为  $n$ ，则  $\mathbf{y}$  为  $n$  维向量，对于  $\mathbf{y}$  的第  $i$  个维度，如果值为 0，则表示第  $i$  个路段不是关键路段，反之表示第  $i$  个路段是关键路段； $\mathbf{c}$  表示路网的拓扑结构， $C$  是高速公路网络所有拓扑结构的集合； $P(\mathbf{c}|\mathbf{y})$  表示当关键路段集合为  $\mathbf{y}$  时，高速路网的拓扑结构为  $\mathbf{c}$  的概率； $k$  表示第  $k$  个车辆的出行路径， $K$  表示所有车辆的出行路径集合； $P(K|\mathbf{c})$  表示当路网拓扑结构为  $\mathbf{c}$  时，高速公路车辆出行路径集合为  $K$  的概率； $X_k$  表示当车辆的行驶路径为  $k$  时，车辆的行驶时间。这个时间可以用交通动力学理论求解 [10]。

### 3.2.3 次模性分析

#### 单调性证明

定义  $\mathbf{y}$  是关键路段集合， $e_i$  代表路段  $i$ ，定义  $\mathbf{y}^1 = \mathbf{y} + e_i$ ， $\mathbf{y}^1$  所代表的关键路段集合比  $\mathbf{y}$  多出关键路段  $e_i$ 。定义  $\Delta L$ ：

$$\Delta L = L(\mathbf{y}^1) - L(\mathbf{y})$$

结合式3.2，得到：

$$\Delta L = -\left(\sum_{c \in C} P(c|\mathbf{y}^1)P(K|c)X_k - \sum_{c \in C} P(c|\mathbf{y})P(K|c)X_k\right)$$

化简，得到：

$$\Delta L = -\left(\sum_{c \in C} \Delta P(c|e_i = 1)P(K|c)X_k - \sum_{c \in C} \Delta P(c|e_i = 0)P(K|c)X_k\right)$$

式中， $\Delta P(c|e_i)$  表示在  $\mathbf{y}^1$  和  $\mathbf{y}$  两种投资方式中，路网拓扑结构概率差值的绝对值。易知当路网结构只有  $e_i$  不同时， $\Delta P(c|e_i = 0) = \Delta P(c|e_i = 1)$ 。公式化简为：

$$\Delta L = -\left(\sum_{c \in C} \Delta P(c|e_i = 1)P(K|c)(X_k^{e_i=1} - X_k^{e_i=0})\right)$$

式中， $X_k^{e_i=1} - X_k^{e_i=0} \leq 0$ 。所以  $\Delta L \geq 0$ ，函数单调性得证。

## 次模性证明

次模函数 (submodular function) 是一种具有“边际效应递减”效应的函数，即对于一个集合函数，如果  $S \subseteq V$ ，那么在  $V$  中增加一个元素所增加的收益要小于等于在  $S$  的子集中增加一个元素所增加的收益。形式化表述就是：对于函数  $f$  而言，若  $A \subseteq B \subseteq V$ ，且  $\varepsilon \in V - B$ ，则  $f(A \cup \{\varepsilon\}) - f(A) \geq f(B \cup \{\varepsilon\}) - f(B)$ ；或者若  $A \subseteq \Omega$   $B \subseteq \Omega$ ，则  $f(A) + f(B) \geq f(A \cup B) + f(A \cap B)$ ；或者对于任意  $X \subseteq \Omega$   $x_1, x_2 \in \Omega$ ，下面的式子一定成立： $f(X \cup x_1) + f(X \cup x_2) \geq f(X \cup x_1, x_2) + f(X)$ 。满足这三个条件中的任意一个，函数  $f$  即满足次模性。

假设  $\varepsilon$  是某一条路段， $\mathbf{y}$  和  $\mathbf{Y}$  都是关键路段的集合， $\Omega$  是关键路段集合的全集空间， $\mathbf{y} \subseteq \mathbf{Y} \subseteq \Omega$ 。 $\varepsilon \in \Omega - Y$ 。 $\{y + \varepsilon\}$  表示对于关键路段集合  $\mathbf{y}$ ，将  $\varepsilon$  作为新的关键路段加入，形成新的关键路段集合。

定义：

$$I = L(\mathbf{y} + \boldsymbol{\varepsilon}) - L(\mathbf{y}) - (L(\mathbf{Y} + \boldsymbol{\varepsilon}) - L(\mathbf{Y})) \quad (3.4)$$

不妨假设  $\mathbf{Y} = \mathbf{y} + \boldsymbol{\varepsilon}_2$ , 公式 3.4 转化为:  $I = L(\mathbf{y} + \boldsymbol{\varepsilon}_1) - L(\mathbf{y}) - (L(\mathbf{y} + \boldsymbol{\varepsilon}_1 + \boldsymbol{\varepsilon}_2) - L(\mathbf{y} + \boldsymbol{\varepsilon}_2))$

令  $J = L(\mathbf{y} + \boldsymbol{\varepsilon}_1) - L(\mathbf{y})$ , 要证明  $I \geq 0$ , 即证  $J$  单调非增。

$J$  属于有限离散函数, 对  $J$  进行求导化简 [3], 得到:  $\frac{dy}{dx} = \sum_{c_1|\mathbf{y}+\boldsymbol{\varepsilon}} (\sum_{c_1|\mathbf{y}+\boldsymbol{\varepsilon}} P(c_1) - \sum_{c_2|\mathbf{y}} P(c_2)) X_k$ 。显然  $\sum_{c_1|\mathbf{y}+\boldsymbol{\varepsilon}} P(c_1) * X_k$  具有单调非减性, 导数恒大于等于 0。模型的次模性得到证明

由于对于具有次模性的模型, 贪心求解的精度误差不会超过  $\frac{1}{e} * OPT$ , 所以模型可以用贪心方法近似求解。

### 3.2.4 基于贪心法的关键路段求解

#### 贪心法求解

贪心算法主要用于优化算法的复杂度, 采用逐步获取局部最优解的方式, 不断循环, 直到得到最终解集。局部最优解求解思路是: 遍历  $n$  条路段, 计算每一条路段被选为关键路段之后对高速公路通行效率的影响, 选出影响最大的那条边作为关键路段。

定义路段损毁概率矩阵  $P = \{p_1, p_2, \dots, p_n\}$ , 其中  $p_i$  代表路段  $i$  出现事故的概率。定义  $U = \{u_1, u_2, \dots, u_n\}$  为路段概率变化矩阵,  $u_i$  表示管理者对路段  $i$  采取措施后, 路段  $i$  事故率的变化量。 $Y = \{y_1, y_2, \dots, y_n\}$  表示关键路段集合, 其中  $y_i$  取值 0 或 1, 表示路段  $i$  是否属于关键路段, 1 表示属于, 0 表示不属于。定义管理者处理关键路段后, 路段的损毁概率矩阵  $P^1 = P + U^T * Y$ 。易知在贪心求解过程中,  $|Y| = 1$ , 定义贪心方法的目标函数:

$$\underset{|Y|=1}{\operatorname{Max}}(L(Y))$$

使用上式，求得第一条关键路段  $k$ ，更新  $P^1 = P + U^T * Y_{y_{k=1}}$ ， $P = P^1$ ，将  $k$  记录为关键路段，并将  $y_k$  恒置为 0，循环搜索下一条关键路段，直到关键路段数量达到预算值。伪代码如下：

---

**Algorithm 1** 贪心算法求解模型

**Require:** 高速车辆 O-D 数据，高速公路网络拓扑结构，关键路段数量，路段损毁率

**Ensure:** 高速公路关键路段集合

```

1: function GREEDY(ODMatrix G = V, E B Pe)
2:   res  $\leftarrow 0$ 
3:   Array  $\leftarrow \text{Null}$ 
4:   k  $\leftarrow 0$ 
5:   l  $\leftarrow 0$ 
6:   while len(Array)  $\leq B do
7:     for i  $\in E - \text{Array} do
8:       if L(Array.Append(i)) > k then
9:         k  $\leftarrow L(\text{Array.Append}(i))$ 
10:        l  $\leftarrow i$ 
11:      end if
12:    end for
13:    res  $\leftarrow k$ 
14:    Array  $\leftarrow \text{Array.Append}(l)$ 
15:  end while
16:  return Array
17: end function$$ 
```

---

为验证贪心算法的效果，在此引入对比方法：

算法2使用枚举方法，获取最优解。从  $n$  条路段中，利用枚举方法，枚举计算所有  $C_n^B$  种关键路段情况，计算每一种情况下的高速公路通行效率的变化，从中选取最优解。

算法3利用高速公路网络拓扑结构，抽取关键路段。算法中的  $Z(i)$  是计算路段  $i$  的中心性函数 [9]，该方法引用经典路段中心性来度量关键路段。

算法4基于统计学方法，计算路段重要程度，获取关键路段。结合式中  $f_i$  表示路段  $e$  的流量。：

**Algorithm 2** 枚举

**Require:** 高速车辆 O-D 数据, 高速公路网络拓扑结构, 关键路段数量

**Ensure:** 高速公路关键路段集合

```

1: function ENUMERATION(ODMatrix G = V, E B Pe)
2:   res  $\leftarrow 0$ 
3:   Array  $\leftarrow Null$ 
4:   k  $\leftarrow 0$ 
5:   for l  $\in \Omega$  and len(l)  $\leq B$  do
6:     if L(l)  $> k$  then
7:       k  $= L(l)$ 
8:       Array  $= l$ 
9:     end if
10:   end for
11:   return Array
12: end function
```

**Algorithm 3** 拓扑中心性

**Require:** 高速公路网络拓扑结构, 关键路段数量

**Ensure:** 高速公路关键路段集合

```

1: function ENUMERATION(ODMatrix G = V, E B)
2:   res  $\leftarrow 0$ 
3:   Array  $\leftarrow Null$ 
4:   k  $\leftarrow Null$ 
5:   for i  $\in E$  do
6:     k.Append({i, Z(i)})
7:   end for
8:   SortByValue(k)
9:   Array  $\leftarrow k[0 : B]$ 
10:  return Array
11: end function
```

## 复杂度分析

最优解法为直接枚举, 即穷举所有可能的路段组合, 一一计算当这些路段被选为关键路段时, 路网通行效率的提升量。选出对路段通行效率提升最大的关键路段集合。对关键路段的选取时间复杂度为  $O(C_n^B) = O(n^B)$ , 复杂度属指数级别, 随着路网规模  $n$  与关键路段规模  $B$  的增大而急剧增加。

贪心法选取关键路段的时间复杂度是  $O(n * B)$ 。通过  $B$  轮循环, 选取关键路段, 每次循环中, 选取节点的时间复杂度为  $n$ 。

**Algorithm 4** 统计**Require:** 高速公路网络拓扑结构, 关键路段数量, 高速公路路段损毁概率**Ensure:** 高速公路关键路段集合

```
1: function ENUMERATION( $G = V, E, B, P_e$ )
2:    $res \leftarrow 0$ 
3:    $Array \leftarrow Null$ 
4:    $k \leftarrow Null$ 
5:   for  $i \in E$  do
6:      $k.Append(\{i, f_i * P_i\})$ 
7:   end for
8:    $SortbyValue(k)$ 
9:    $Array \leftarrow k[0 : B]$ 
10:  return  $Array$ 
11: end function
```

---

### 3.3 实验及结果

本节针对各种方法在真实的交通数据集中进行实验, 通过对比已有的关键路段挖掘方法, 评估模型的效果。实验环境为: Windows Server 2008, 64GB RAM, Inter(R)Xeon(R) CPU E7-4830 2.13GHz 2.13GHz (2 处理器), 后续章节的实验均在相同的实验环境下进行。特别地, 实验中采用了两个国内高速公路网的数据: 安徽省和山西省高速公路网数据。

#### 3.3.1 实验数据

本节的实验数据来自于安徽省和山西省的高速公路路网, 其中的主要数据为高速路网中车辆的行驶 O-D 数据。路网中包含 142 个出口位置和 142 个入口位置。为了方便研究, 将车辆的 O-D 数据整合为出行 O-D 矩阵 ODMatrix:

$$\begin{matrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{matrix}$$

其中,  $a_{ij}$  表示以收费站 i 为起点 O, 以收费站 j 为终点 D 的车辆数量。

高速公路路段损毁概率通过统计历史的路段损毁次数获得。这些数据一部分数据库已有, 一部分通过新闻抓取获得。路段的损毁包括交通事故损毁, 重大自然灾害损毁, 严重堵车等。

### 3.3.2 实验结果

图3.1, 3.2给出了在不同时间区段下, 几种方法的最终结果比较。图3.1是基于 2010 年 10 月 30 日一天的实验结果, 纵坐标代表路网整体通行效率 (路网整体通行时间取负) 的绝对值, 横坐标代表一天内的不同时间段, 本实验中以 1 小时为一个时间段, 采样八个时间点  $[0 - 1, 3 - 4, 6 - 7, 9 - 10, 12 - 13, 15 - 16, 18 - 19, 21 - 22]$ 。由图3.1可以发现, 在整体上贪心算法明显优于统计算法, 同时统计算法又比直接基于高速公路拓扑结构获取关键路段有效, 原因是高速公路整体网络结构比较简单, 路网拓扑结构的某些性质体现的不明显。在不同的时间段, 高速公路的流量在不断变化, 不同方法的效果之间的差异也在变化。在高速公路车流最少的午夜, 几种方法差异达到最小, 从早上六点开始, 到流量最高的中午, 三种方法之间的差异逐渐增大, 这体现了高速公路流量对关键路段选取结果的影响, 流量越大, 关键路段维护后造成的效益越大; 流量越大, 选取关键节点的误差造成的影响就越大。图3.2是基于从 2010 年 10 月 10 日开始, 到 2010 年 10 月 16 日为止一周数据的实验, 纵坐标和图3.1一样, 表示网络整体的通行时间。纵坐标以一天为一个时间段, 从当天 0 点采样到当天 24 点, 采样七天 (从周日到下一个周六)。可以发现, 在以一整天的 O-D 矩阵为数据集进行研究时, 不同天之间的路网通行效率变化较小, 不同方法之间的差异也趋于平稳。这证明了高速公路具有一定的稳定性和规律性, 即可以通过研究历史数据, 获取静态关键路段。这些关键路

段具有普适性。

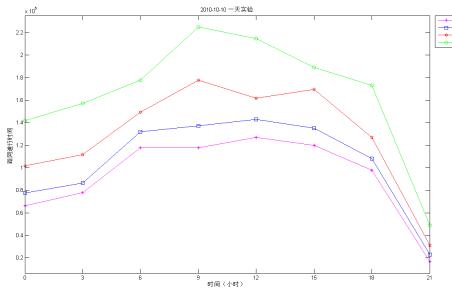


图 3.1 关键路段挖掘：以 1h 为区间

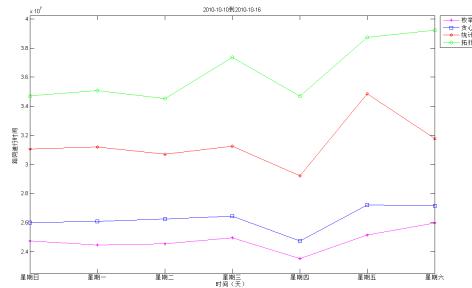


图 3.2 关键路段挖掘：以 1d 为区间

表3.1给出了不同方法求得的路网通行效率。表格中的数值是  $L(\mathbf{y})$  的绝对值（路网整体通行时间），可以看出贪心算法最接近最优解，而且误差在可接受范围内。

	枚举	贪心	统计	拓扑
一天	926030.06	1053575.26	1287439.55	1660243.55
一周	21674024.80	22989458.02	27510044.42	31790488.20

表 3.1 算法结果集

图3.3给出了关键路段在路网中的分布图，图3.3(a)是基于贪心算法求解的关键路段集合，图3.3(b)是基于高速公路统计方法获得的路段集合。图3.3(c)是基于枚举所得的最优解集，图3.3(d)是基于路网拓扑结构选取的关键路段集合。对比图3.3(a)和图3.3(b)可以发现，直观上重要的点（承载流量较大的路段，事故多发路段等）并不一定在路网中属于关键路段集合，关键路段集合需要经过计算才能求出；直接枚举的路段集合与贪心算法求得的路段集合十分接近，平均有 80% 以上的相似度，而统计和度中心性方法求得的关键路段和枚举方法差距较大。

在实验过程中，我们还发现当时间区间缩短到 1h 时，高速公路网络中的关键路段具有随着时间和流量变化而变化的特性。图3.4是凌晨 3 点时的高速公路关键路段集合，图3.5是早上九点时的高速公

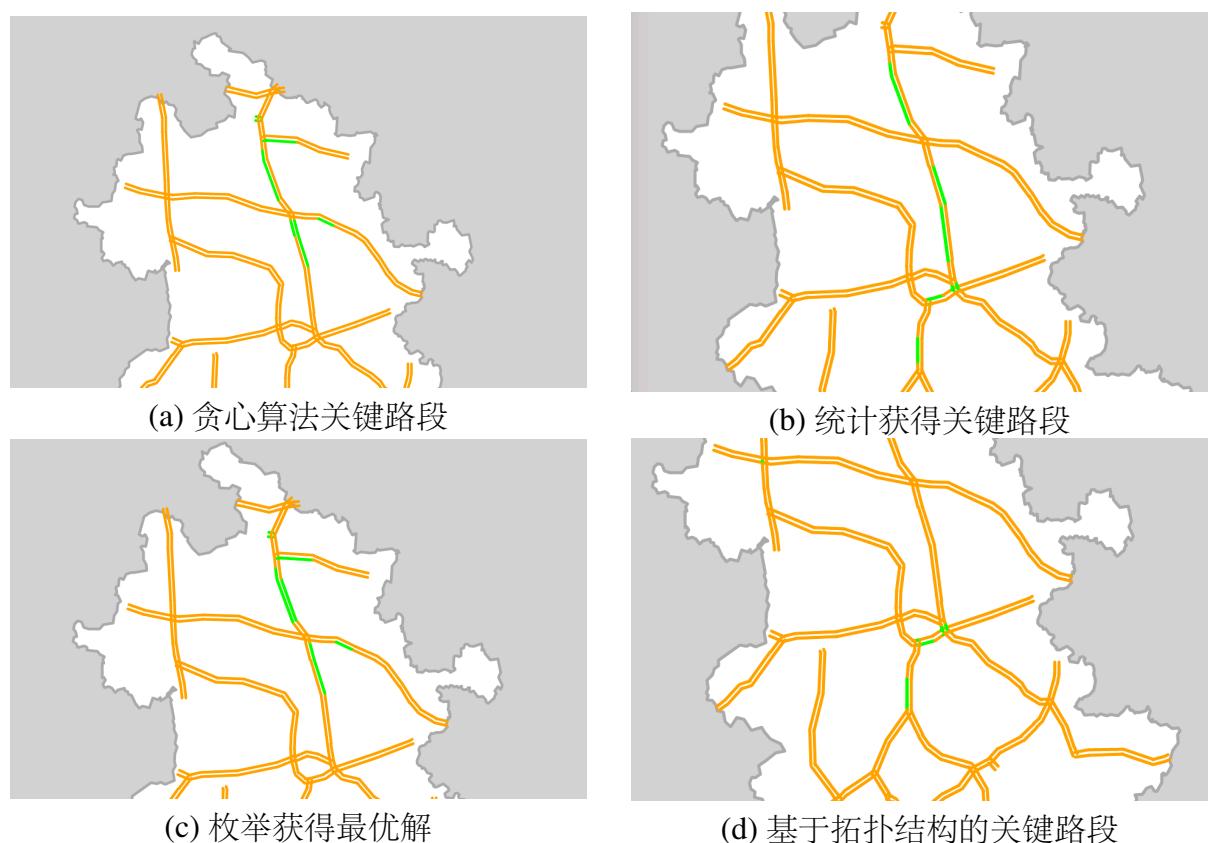


图 3.3 不同方法求得的关键路段结果图

路关键路段集合。这一发现证明了高速公路的关键路段具有动态变化特性，即高速公路的关键路段不是一成不变的，而是会随着整个路网的流量的变化而变化。

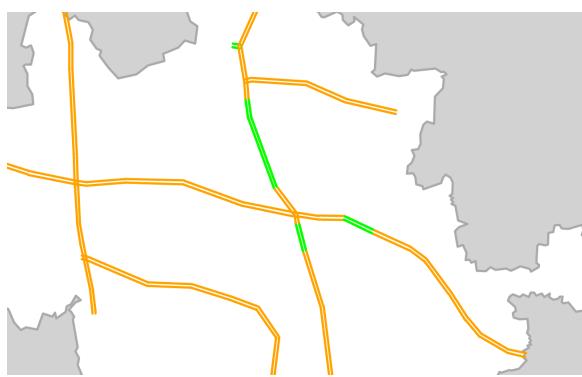


图 3.4 关键路段挖掘: 03: 00

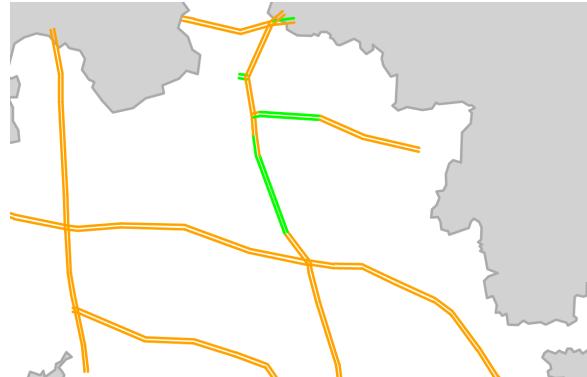


图 3.5 关键路段挖掘: 09:00

### 3.3.3 时间分析

基于暴力枚举方法的时间复杂度:  $O(n^B * 2^n)$

基于贪心算法的时间复杂度:  $O(n * B * 2^n)$

基于统计路段重要性方法的时间复杂度:  $O(n * \log(n))$

基于路网拓扑结构方法的时间复杂度:  $O(n * \log(n))$

实验时间如表4.2所示，第一行代表实验的方法，第一列代表实验数据的范围，表格内的数值是实验的平均运行时间。

	枚举	贪心	统计	拓扑
一小时	1day	30min	1min	1min
一天	6day	2h	2min	1min
一周	7day	3h	5min	1min
一月	7day	3h	8min	1min

表 3.2 不同算法的运行时间

由表格可以看出，在以一个省为数据集的基础上，枚举方法已经处于一种较大的时间复杂度；贪心算法在一定程度上解决了算法过慢的情况，并且在精度上有一定的保证，可以应用于静态路网关键路段

识别问题，但是对于动态实时应用仍旧不够；基于统计领域的路段重要性排序方法、基于路网拓扑结构的关键路段挖掘方法虽然在时间上效率较高，但是在精度上达不到要求。一周和一月的时间差距较小的原因是：对数据进行预处理，O-D 相同的用户被归为一类。在时间区段增大到一定程度时，O-D 类数量不再增加，算法运行时间增长较小。

### 3.4 本章小结

本章提出了一种面向高速公路网络的关键路段挖掘模型，同目前高速公路关键路段已有的挖掘方法相比，该方法的优势是结合高速公路的特性，考虑高速公路上的车流流量、路段事故率，从宏观角度提出一个整体的优化模型。针对上述模型，本章分析证明了模型的次模性，并给出基于贪心方法的关键路段挖掘方法。特别的，本文通过枚举方法，在较低的时间效率下计算高速公路中的最优解。结果表明该模型的贪心算法解可以很好地逼近真实解，并且在时间复杂度上有了一定规模性的优化，证明了贪心算法的可行性。然而，即使贪心算法可以在一定规模上优化整体的时间复杂度，并且可以在实际应用中运行良好，但是这是基于目前的研究目标是静态关键路段挖掘，同时高速公路也只有部分路段产生过断流等重大事故的情况下达成的。当任务环境更为复杂时（扩大到全国高速公路网络），当管理者需要更加迅速得到实时反馈的时候，上述方法受到算法计算规模的约束，无法达到预期的效果。下一章将针对高速公路的网络特性，给出相应的解决手段。

## 第四章 基于社群划分的关键路段识别方法

本章主要分四小节。第一节讲述了模型的性质，分析了贪心算法的不足；第二节给出了基于社群划分的关键路段挖掘方法模型；第三节是实验结果；第四节总结了本章内容。

### 4.1 引言

上一章介绍了面向高速公路的关键路段挖掘模型，并给出了贪心算法。然而根据模型的定义，就算进行简化，认为关键路段已经选出，计算对关键路段进行维护之后的整体网络通行效率也需要  $2^n$  的时间复杂度。贪心算法的实际时间复杂度是

$$O(B * n * 2^t)$$

式中  $B$  表示关键路段数量， $n$  表示网络中路段的数量， $t$  表示网络中可能出现损毁的路段数量。虽然在实际应用中，高速公路中路段规模不大 ( $O(10^3)$ )，大部分路段的损毁概率是 0，贪心算法对于静态高速公路中的关键路段挖掘可适用性高，但是贪心算法的时间复杂度仍旧属于指数级别，当高速公路网络规模变大后，复杂度指数上升，对于有实时性要求的动态关键路段挖掘方法并不适用。本节给出基于复杂网络社群划分的关键路段挖掘方法。

模型需要从输入的代表关键路段的离散 0-1 向量  $y$ ，求得高速公路网络通行效率的期望。这种输入为整数或整数向量，并且内部具有概率事件的问题，属于随机整数规划问题。在数学优化领域，随机规划是一个涉及不确定性优化问题的框架。比如说两阶段线性规划。决策者在第一阶段采取一些行动，之后发生随机事件影响第一阶段决策

的结果。不断调整第一阶段的决策，使得整体期望收益达到最大。

现有的随机整数规划问题大都是基于班德斯分解方法（Benders Decomposition）进行研究，然而班德斯优化方法要求有两层模型，且两层模型之间互不影响。本研究中，第一层的决策变量  $\mathbf{y}$  会直接影响到第二层里面的路网拓扑结构概率，对于这种相互依赖的随机规划问题，现有的研究没有找到适用的优化方法。

## 4.2 问题定义

### 4.2.1 问题定义

高速公路网络除具有绝大多数复杂网络的特征外，作为空间网络还具有不同于抽象网络的特性，这些特性决定了高速公路网络的拓扑性质。具体可以归纳为：高速公路交通网络的节点存在于真实物理空间，高速公路网络中的边是一种实体联接，具有明确的空间意义。高速路网中的边与抽象网络中的边不同，高速公路交通网络中节点的边权与交通距离直接相关，这一特性直接影响着高速公路网络出现小世界行为的可能性 [1]；高速公路交通网络中单一节点所能联接的边的数目受到物理空间的限制，这种限制会影响到网络的复杂程度。

在高速公路项目研究中，我们发现低跳数的用户占大多数。如图 4.1，可以发现在高速公路中，低跳数的车辆占了大多数，10 跳以下的车辆占所有车辆总数的 90% 以上。再结合高速公路的异质性，复杂网络的社群性，我们认为高速公路网络应该也具备社群性质，即存在一个个社群，这些社群各自包含一些收费站和高速公路路段，高速公路中的车辆大都从社区内部的节点出发，在同一个社区的另一个节点驶离。社区之间的车辆交流尽量小。

为此，抽取某一天的高速公路 O-D 数据，将有 O-D 交流的收费站之间连线，流量越多，线的颜色越深，流量越少，线的颜色越浅。

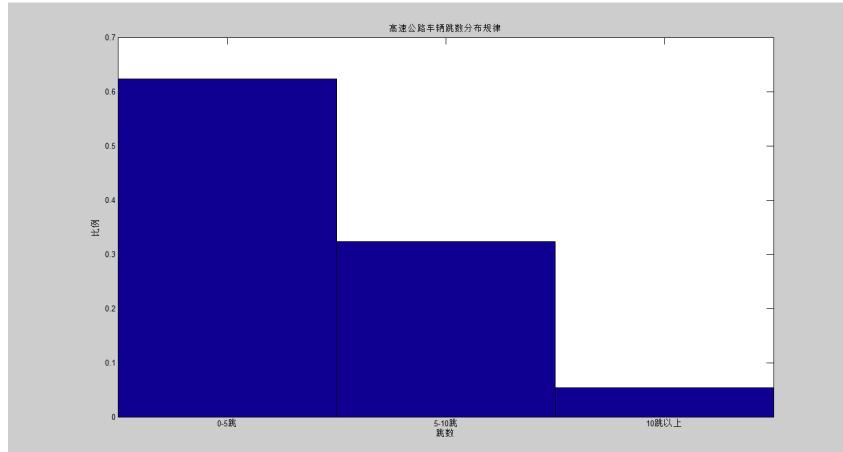


图 4.1 高速公路车辆跳数分布图

如图4.2，可以较直观的看出高速公路的社群特性。

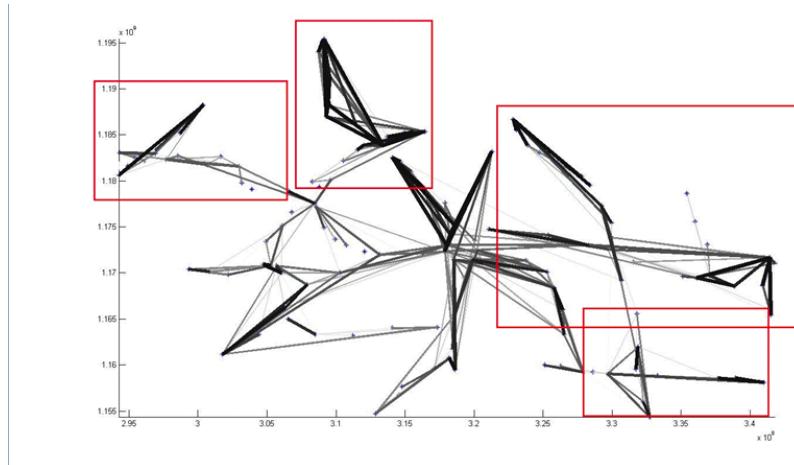


图 4.2 高速公路社群特性

高速公路社群划分的目的是将整个高速公路拓扑结构分成一个个社区，使得社区内部交流尽量多，社区之间的交流尽量少，最终在各自社群分别计算关键路段，分治计算，最后进行合并，达到优化算法效率的目的。在此引入基于模块性优化的社区划分方法。

复杂网络具有社群特性，高速公路属于复杂网络的一种。给定高速公路有向图  $G = \{V, E\}$ ，其中  $V$  代表收费站（节点）的集合； $E$  表示边的集合。定义社群  $c = \{v_1, v_2, \dots, v_m\}$ ，其中  $v_i$  是网络中的节点，即收费站或者交叉路口；社群集合  $C = \{c_1, c_2, \dots, c_u\}$ ；其中  $v_i \in V$ ，

$$c_i \cap c_j = \emptyset, \sum_{i=1}^u \sum_{v \in c_i} v = V.$$

基于高速公路社群划分的关键路段挖掘算法主要采用分治思想，将一个难以直接解决的复杂问题，分割成一些规模较小的问题，逐个计算，分而治之。本文主要将路网分成一个个子路网，在子路网中分别计算关键路段，之后再用一定方法进行合并。在此需要解决两个问题：

- 1) 如何分群
- 2) 分群求解后，如何合并

传统的复杂网络社群划分方法中，大都是针对虚拟网络（如社交网络）进行研究。高速公路网络和虚拟网络有很大的不同。在虚拟网络中，两个点之间只要有交流，那就代表有边相连；在高速公路中，我们认为只要两个收费站有流量交流，即 O-D 不为 0，那么这两个收费站之间就有边连接（不同于上一章的路网定义）。但是这个边和其他的复杂网络如社交网络不同，社交网络中两个节点之间的空间距离定义为 1 跳，但是对于物理网络来说，两个节点之间的边具有实体距离。高速公路中路段之间的影响也会根据物理距离的变化而变化，这些都是传统方法中没有考虑到的。

2004 年，Newman 和 Girvan[56] 提出了一个用于刻画网络社区结构优劣的量化标准，被称作模块化函数。简单的带权模块化函数定义如下：

$$Q = \frac{1}{2m} \sum_{ij} [A_{ij} - \frac{k_i k_j}{2m}] \delta(c_i, c_j) \quad (4.1)$$

式4.1 中， $A_{ij}$  表示节点  $i$  和节点  $j$  之间的边权； $k_i = \sum_j A_{ij}$  表示所有与节点  $i$  相连的边的边权和； $c_i$  是指  $i$  所属的社群编号；如果  $c_i = c_j$ ，那么  $\delta(u, v) = 1$ ，否则等于 0； $m = \frac{1}{2} \sum_{ij} A_{ij}$ 。

模块化函数主要用于度量社群划分结构的优劣，现有的基于模块化函数的分群算法都没有考虑高速公路的特性 [2]，并且在高速公路网络上出现了低分辨率特性和极端退化特性 [58]。Newman 提出了一种社群挖掘方法 [59]：初始的时候没个节点都是一个社群，之后进行迭代，每次迭代时都选择使目标模块函数在  $Q$  增加最大的社群进行合并。这个方法虽然在时间效率上很高，但是没有解决社群划分的极端退化特性。Guimera 提出一种基于模拟退火的模块性优化方法：初始解是随机生成的社团集合，在每次迭代过程中，采用一定策略，结合当前解生成新的解集，用模块化函数  $Q$  判断解集的优劣，最后用模拟退火中的 Metropolis 准则来决定是否采用该解。这个方法虽然在一定程度上解决了极端退化特性，但是他有一个很严重的问题，在相通的输入集合上，生成的最终结果往往不同，不符合稳定性要求，而且时间复杂度大，求解效率低。Blondel[60] 提出快速模块优化方法，他认为首先在局部使用局部模块化函数  $f$  获得局部社团，然后再对这些局部社团作为一种超级节点，再进行合并，不断迭代，直到模块化函数  $Q$  不再增加为止。这个聚类方法存在聚类社团过大的情况，不符合本文中缩小节点量级，优化算法时间复杂度的目的。针对现有研究的不足，结合高速公路的路网特性，在此提出一种新的面向高速公路的社群划分模型。

#### 4.2.2 模型定义

首先定义模块化函数  $Q$ :

$$\Delta Q = \left[ \frac{\sum_{in} C + 2k_{i,in}}{2m} - \left( \frac{\sum_{tot} C + k_i}{2m} \right)^2 \right] - \left[ \frac{\sum_{in} C}{2m} - \left( \frac{\sum_{tot} C}{2m} \right)^2 - \left( \frac{k_i}{2m} \right)^2 \right] - L(i) \quad (4.2)$$

公式4.2用于判断当节点从一个社区转移到另一个社群的时候，整

体路网的社群化结构的变化。根据变化的大小决定节点的社群归属。式中， $\Sigma_{in} C$  表示社群  $C$  内部的所有边的权重和； $\Sigma_{tot} C$  表示所有与社群  $C$  中的节点相连的边的权重和； $k_{i,in}$  表示  $i$  到  $C$  中所有节点之间的连线的权重和； $k_i$  表示所有和节点  $i$  直接相连的边的权重和； $m$  是路网中所有边的权重之和； $L(i)$  是模型罚项，代表  $i$  转移社群后，不同社区之间交通流的变化。

$L(i)$ :

$$L(i) = \frac{k_{i,c_1} - k_{i,c_2}}{k_{c_1,c_2}} \quad (4.3)$$

式4.3中， $k_{i,c_1}$  表示路段  $i$  流向社群  $c_1$  的流量， $k_{i,c_2}$  代表路段  $i$  流向社群  $c_2$  的流量， $k_{c_1,c_2}$  表示社群  $c_1,c_2$  中所有节点之间的流量和。

本文提出的模型中，边的权重不止与两个节点之间的流量有关，还与两个节点之间的物理距离有关。和传统复杂网络不同，节点之间的距离不再由节点之间的最短跳数决定，而是由节点之间的最短物理距离  $L$  决定：

$$L_{ij} = \sum_{e \in E_{ij}} e \quad (4.4)$$

式4.4中， $E_{ij}$  是节点  $i$  和节点  $j$  之间的最短路径中路段的集合。定义边权重：

$$W_{ij} = \frac{f_{ij}}{L_{ij} * T} \quad (4.5)$$

为了解决传统社群划分中的低分辨率问题，本文中的社群划分方法也采用自底向上的聚类思想，首先定义每一个节点都是一个社群，社群集合  $C = \{c_1, c_2, \dots, c_u\}$ ，路段集合  $E = \{e_1, e_2, \dots, e_n\}$ 。初始情况下， $u = n$ 。在每次迭代过程中，遍历路段  $E$ ，利用模块化函数  $\Delta Q$ ，

判断节点  $e_i$  所属的社群:

$$\underset{c_j \in C}{\operatorname{Max}} (\Delta Q)$$

式中,  $\Delta Q$  表示当路段  $e_i$  由原来的社群划分到社群  $c_j$  时, 模块化函数的改变量。路段  $|E|$  经过一遍遍历后, 形成一个新的社群集合  $C^1 = \{c_1^1, c_2^1, \dots, c_u^1\}$ , 结合社群集合和路段集合, 进行下一轮遍历, 直到收敛。

传统方法中, 社群划分具有极端退化特性, 即最终结果无法收敛到某一个确定的最优解, 而是会收敛到一个解集合中, 在一定规模的解集中循环。比如说, 假设社群划分已经收敛, 在第  $i$  次迭代过程中, 得到了社群集合  $C^i$ 。 $C^{i+1} \neq C^i$ , 但是在经过  $k$  次迭代后,  $C^{i+k} = C^i$ 。实验表明, 在高速公路上进行这样的社群划分, 最终结果具有空间交叉特性, 即不同的社群之间存在物理空间上的交叉(如图4.3), 这种情况下不同社群之间的相互影响比较大。为了解决社群划分的空间交叉与极端退化问题, 我们采用多变权值的思想, 初始权值为  $W_{ij} = \frac{f_{ij}}{L_{ij}*2}$ , 目标模型收敛后, 记录本次迭代解集的规模  $k$ , 变化权值  $W_{ij} = \frac{f_{ij}}{L_{ij}*(2-0.1*k)}$  ( $l$  为迭代次数), 通过加大路段距离的权重, 不断减少解集规模, 改善社群之间空间交叉情况; 为了加速收敛, 结合模拟退火思想, 定义退火温度为  $T$ , 当轮迭代的解集数量为  $k_i$ , 上轮迭代的解集数量为  $k_{i-1}$ 。当  $k_i < k_{i-1}$  时, 以概率  $\frac{(k_{i-1}-k_i)}{T(k_{i-1})}$  结束迭代。分群算法伪代码见 Algorithm 5。

#### 4.2.3 基于社群划分的关键路段求解

基于已经分群的高速公路网络, 在此提出关键路段挖掘方法。

分治法的核心是分而治之, 首先分割社群, 将每个社群看作独立的路网。在每个社群里, 用前一章提出的贪心算法选出各个社群中的关键路段, 并且计算出每个关键路段被选出后对路段通行效率的

**Algorithm 5** 高速公路社群划分方法

**Require:** 高速车辆 O-D 数据, 高速公路网络拓扑结构, 最大社群节点数量

**Ensure:** 高速公路社群划分结果

---

```

1: function COMMUNITY(ODMatrix G = V, E B)
2:   res  $\leftarrow [[\{0|0\} \{1,1\} \cdots \{n,n\}]]$ 
3:   tmp  $\leftarrow [\{\}]$ 
4:   pre  $\leftarrow [[\{0|0\} \{1,1\} \cdots \{n,n\}]]$ 
5:   k  $\leftarrow 0$ 
6:   l  $\leftarrow 0$ 
7:   T  $\leftarrow 100$ 
8:   while  $|len(res) - len(pre)| \leq T$  do
9:     res  $= res[-1]$ 
10:    pre  $= res$ 
11:    while res $[-1] \notin res[0:-1]$  do
12:      tmp  $\leftarrow res[-1]$ 
13:      for i  $\in E$  do
14:        for ( do C  $\in tmp \& |C| \leq B)
15:          if  $\Delta Q > k$  then
16:            l  $\leftarrow C$ 
17:            k  $\leftarrow \Delta Q$ 
18:          end if
19:        end for
20:        tmp $[l] \leftarrow i$ 
21:      end for
22:      res add tmp
23:    end while
24:    T $-$ 
25:  end while
26:  return res
27: end function$ 
```

---

增量  $\Delta L$ 。忽略不同社团的关键路段之间的相互影响, 把分治法的合并问题归类于投资问题。投资问题的定义如下: 定义资产总额为  $B$ , 总共有  $\{X_1, X_2, \dots, X_u\}$  种货物, 每种货物可以投资  $[0 - B]$  份资源。 $g(y * X_i)$  表示当货物  $X_i$  投资量为  $y$  时, 它所带来的收益。本节问题中, 关键路段的数量归类于资产总额  $B$ , 每个社团看作一种货物  $X_i$ , 每个社团中, 改造关键路段造成的影响用  $g(t * X_i)$  计算。在此提出目标模型:

$$Q = \text{Max} \left( \sum_{i=1}^u g(y * X_i) \right) \quad (4.6)$$

$$\text{Subject to. } \begin{cases} \sum_{i=1}^u y \leq B \\ x_i \geq 0 \end{cases} \quad (4.7)$$

4.6 属于投资问题，可以利用动态规划，在多项式时间内求解。求解步骤如下：

定义  $f_k(x)$  为前  $k$  个社团投入  $x$  份资源时，高速公路通行效率的最大提升量，首先赋初始值： $f_0(x) = 0$ ;  $f_k(0) = 0$ ;  $f_1(x) = g(x * X_1)$ 。递推公式：

$$f_i(j) = \underset{0 \leq y \leq j}{\text{Max}} (CMATRIX[i][y] + f_{i-1}(j - y))$$

公式中，CMATRIX 矩阵表示当社团  $i$  投资  $y$  条路段时，路网通行效率的增加量； $y$  表示本社团中投资路段的数量，根据  $y$  和贪心算法结果，可以推算出具体路段。当  $k = 2$  时，CMATRIX 矩阵已知， $f_1(x)$  全部已知，据此可以推算出所有  $f_2(x)$  的值。递推，依次得到  $f_3(x), \dots, f_u(x)$  的值，结合  $f_u(x)$  中  $y$  的值，根据贪心算法，获取关键路段，反向递推，得到最终关键路段的集合。动态规划伪代码：

其中，路段收益矩阵 CMATRIX：

$$\begin{matrix} c_{11} & c_{12} & \cdots & c_{1B} \\ c_{21} & c_{22} & \cdots & c_{2B} \\ \vdots & \vdots & \ddots & \vdots \\ c_{u1} & c_{u2} & \cdots & c_{uB} \end{matrix}$$

矩阵中， $c_{ij}$  表示第  $i$  个社群中，选取  $j$  条关键路段进行资源投放

**Algorithm 6** 关键路段挖掘方法求解

**Require:** 每个社团中选取不同路段的收益，高速公路网络社团结构，最大社群节点数量

**Ensure:** 高速公路关键路段集合

```

1: function COMMUNITY(CMatrix C = c1, c2, ..., cu B)
2:   定义  $f_k(x)$ : 当前  $k$  个社团投入  $x$  份资源时，最大的通行效率提升量
3:    $f_0(x) \leftarrow 0$ 
4:    $f_k(0) \leftarrow 0$ 
5:    $f_1(x) \leftarrow CMatrix[1][x]$ 
6:    $i = j = 1$ 
7:   while  $i \leq u$  do
8:     while  $j \leq B$  do
9:        $f_i(j) = \text{Max}_{0 \leq y \leq j} (CMatrix[i][y] + f_{i-1}(j-y))$ 
10:       $j = j + 1$ 
11:    end while
12:     $i = i + 1$ 
13:  end while
14:  return res
15: end function
```

后，高速公路网络通行效率的提升量。这一数据由贪心算法在每个社群分别求得。

### 4.3 实验及结果

本章节出了针对每一种方法的有效性做出实验，并将基于高速公路社群划分方法的实际效果与通过枚举得到的最优解进行对比。

基本的社群划分存在分辨率限制和极端退化特性。分辨率限制是指社群划分方法无法发现小于一定规模的社群，极端退化特性是指最终的社群划分结果会收敛于指数数量级的高分解决方案，而不是指向一个或少量最优解。Newman[59] 采用一种方法解决低分辨率问题：初始化时，将每一个节点看作一个独立的社群，之后根据模块化函数不断循环修正节点的所属社群。这个方法用在高速公路上时，虽然解决了低分辨率社群无法发现的问题，但是最终会产生一系列孤立点（如图4.3），这不符合社群划分的初衷。而且最终结果也没有避开极端

退化特性，最终的社群划分结果在一个非常大的解空间中循环。

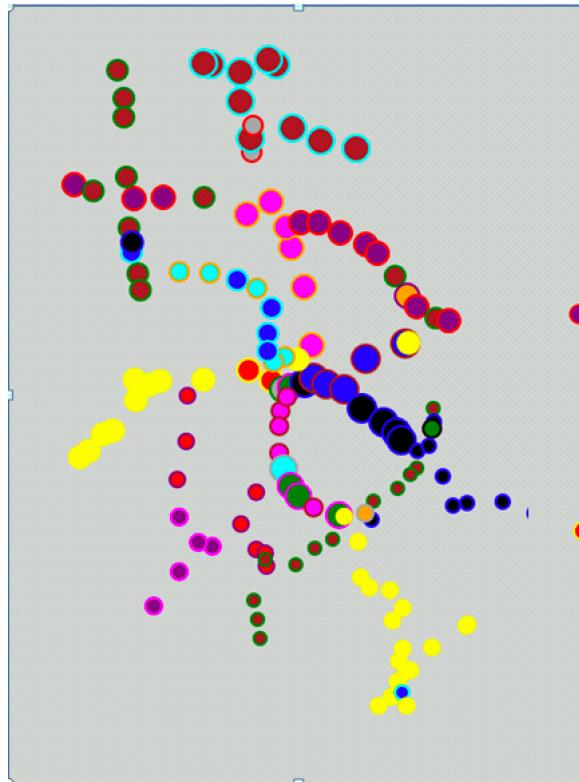


图 4.3 基于模块化函数的社群划分方法

图4.3给出了基本的基于模块化函数的分群结果，首先需要指出：使用基础方法的分群结果收敛于一个具有一千多个解的解集合。最终分群结果会在这些解集合内循环。由图我们可以看出两个问题：

- 1) 存在很多未被分群的孤立点。
- 2) 很多社群存在物理意义上的交叉收费站。

孤立点的产生原因有两个，一是这个收费站本身流量较小，与其他站点交流不多；二是这个站点与其他站点之间的交流较为平均，站点不断流动于不同的社团中。图4.5是基于公式4.2的社群划分结果，该图由几百个社群划分组成的解集中选出，由图可以看出加入物理路经长度的情况下，可以在一定程度上消除孤立点，并且将高速公路划分成较为清晰的几类。但是我们发现仍旧有少量社群，存在物理层面的相互交叉情况，而这种情况不符合高速公路这种物理网络的社群

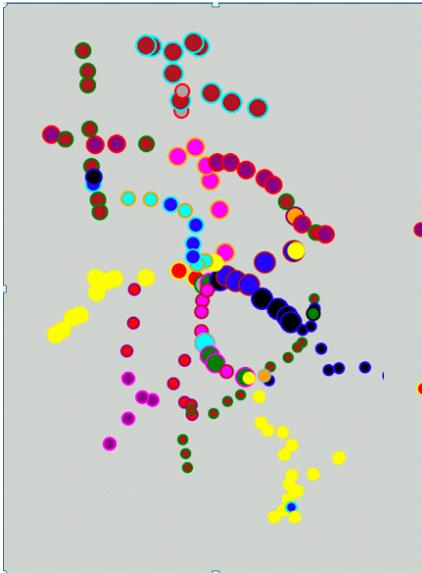


图 4.4 基于模块化函数的社群划分方法

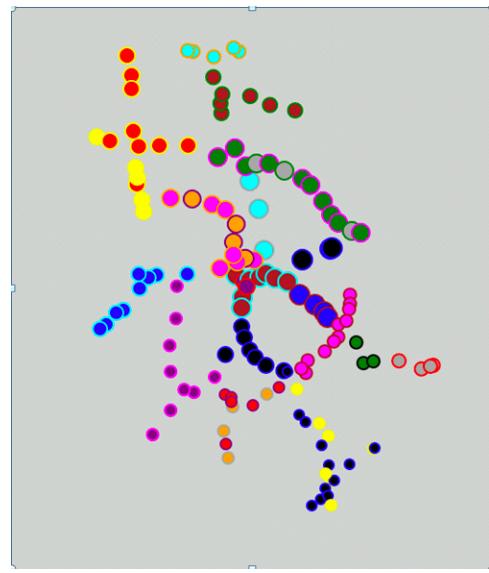


图 4.5 结合物理路网特性的社群挖掘

划分特点。

经过数据分析，出现图4.5中不同社群内部的节点之间存在物理上的交叉情况的原因是——不同社群节点之间的流量差远大于节点之间的距离差。直接将具有交叉节点的社区进行合并虽然简单有效，但是不具有更大规模的适应性，这种方法得到的社群划分效果得不到保证，而且有可能出现过大的社团，不符合社群划分的目的。图4.6给出了基于模拟退火方法的迭代分群方法结果，该结果最终收敛于由 5 个结果组成的结果集，基本消除所有孤立点与社群交叉节点。

根据公式4.1给出的模块化函数  $Q$ ，表4.1给出了不同社群方法模块化的效果。可以看出将边权与物理距离结合考虑后，模块化效果得到了显著提高；虽然模拟退火方法的时间消耗较大，但是它提供了符合物理网络的分群结果，减少不同社群之间的交叉节点，将不同社群之间节点的相互影响降到最低。

图4.7给出了一天时间内，基于分群算法和简单贪心方法的对比试验；图4.8给出了在一周时间内两种方法的对比试验。和上一章节一样，横坐标表示时间，纵坐标表示路网通行效率的绝对值（路网通行

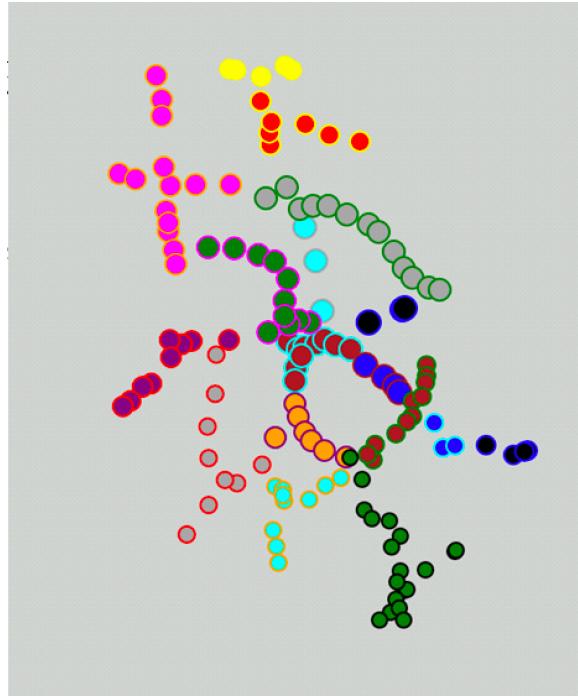


图 4.6 结合路段距离的变权社群挖掘结果

	基于流量划分	基于流量/距离划分	基于变化距离的模拟退火
模块化效果	-1321.21	-1025.50	-1182.84
算法效率	1min	30s	1.5min
收敛度	$10^3$	$10^2$	0 – 10

表 4.1 不同社群划分方法效果对比

时间)。由图可以看出，简单贪心算法和基于分群算法的关键路段挖掘算法之间的误差较为平稳，并且一直维持在一个较低的水平线上。由图可以看出，分群算法可以在和统计算法相似的时间复杂度上，得到比统计算法优秀的解集。

下图给出不同方法选出的关键路段集合，图4.9给出了枚举方法选出的关键路段集合，图4.10给出了简单贪心算法给出的关键路段集合，图4.11给出了结合社群划分的关键路段识别算法的结果，图4.12给出了基于统计学的关键路段集合。观察图4.9和图4.10，发现两者选取的关键路段具有很强的相似性。

误差分析：

表4.2描述了枚举方法和直接贪心方法之间的误差，直接贪心和

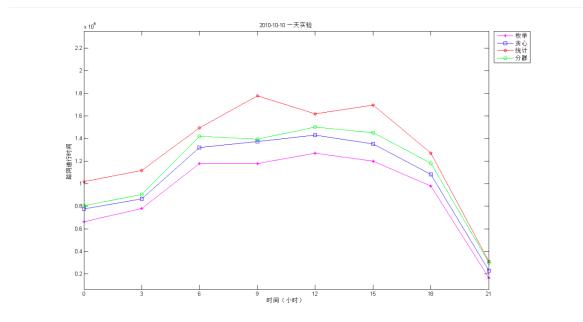


图 4.7 对比实验：以 1h 为区间

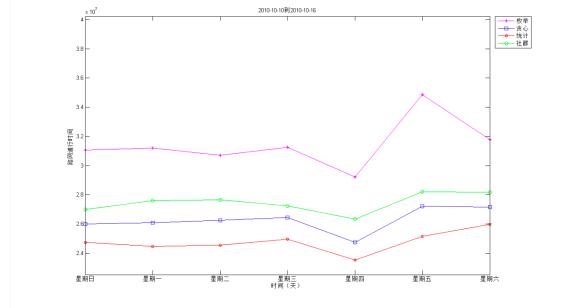


图 4.8 对比实验：以 1d 为区间

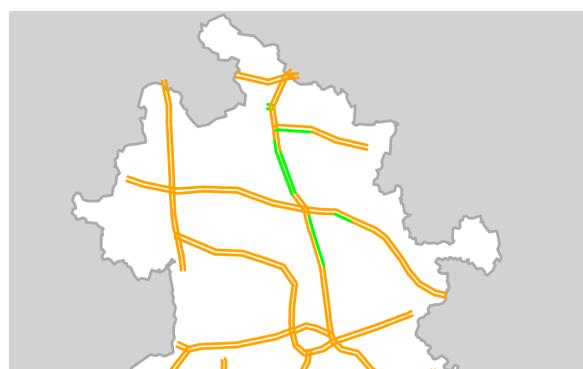


图 4.9 枚举求得关键路段

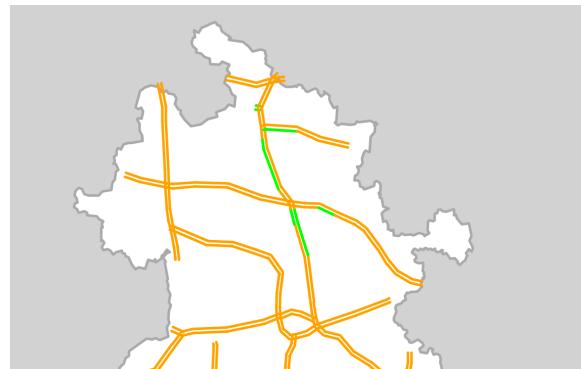


图 4.10 贪心求得关键路段

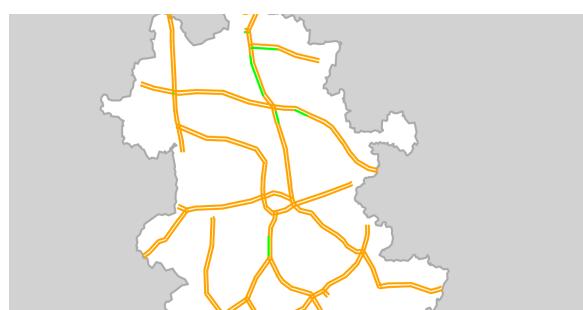


图 4.11 基于社群划分的关键路段

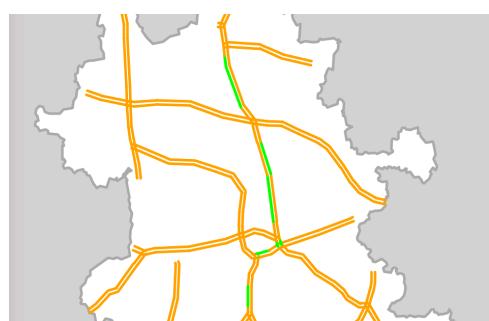


图 4.12 基于统计学的关键路段

	枚举-直接贪心	直接贪心-基于社群划分
一小时	14.63%	12.89%
一天	13.25%	13.26%
一周	13.10%	15.61%
一月	12.99%	11.59%

表 4.2 路网通行效率提升量误差分析

基于社群划分方法之间的误差。误差由高速路网的通行效率计算，可以看出误差在允许范围内。

关键路段选取误差分析：

	枚举-直接贪心	枚举-基于社群划分
一小时	0.18%	0.25%
一天	0.14%	0.20%
一周	0.15%	0.19%
一月	0.14%	0.18%

表 4.3 关键路段选取误差分析

表4.3分析了关键路段选取情况的误差，采用欧式距离来刻画区别。可以看出，随着数据集的扩大，基于社群划分方法的关键路段准确率逐步上升。

运行效率分析：

	枚举	直接贪心	基于社群划分	基于统计
一小时	1day	30min	2min	1min
一天	6day	2h	5min	2min
一周	7day	3h	6min	5min
一月	7day	3h	7min	8min

表 4.4 不同方法运行效率分析

由表4.4可以看出，基于社群划分方法可以将整个算法的时间复杂度再降一个数量级，而结合表4.2来看，精度误差处于可接受范围( $1/e$ )。

#### 4.4 本章小结

本章提出了面向高速公路的社群划分方法，首先分析了传统方法的局限性，然后结合高速公路的独有特性，采用多变权值-模拟退火结合的方法，实现符合高速公路网络特点的社群划分方法。最后结合动态规划方法，实现多项式时间求解。

## 第五章 原型系统的设计与实现

### 5.1 系统功能

系统主要有如下功能：1) 社群划分。根据输入的时间信息，结合数据库中的高速公路收费信息，进行社群划分。2) 关键路段挖掘。根据输入的时间信息，结合数据库中的高速公路收费信息，进行关键路段挖掘。3) 静态挖掘。根据历史信息，稳定挖掘关键路段，给管理者提供参考。4) 动态挖掘。根据实时数据，实现动态挖掘关键路段。

### 5.2 系统架构

高速公路关键路段识别算法是为了解决高速公路管理者解决高速公路资源配比问题而开发的，整个系统基于 B/S 架构，从高层划分为前端和后端两部分，具体系统架构见图 5.1，后端为 Windows 服务，主要包括实时车辆数据的处理以及数据仓库的存储。本文使用了该系统的架构，通过编写“业务逻辑层”和“展现层”，实现原型系统。

系统逻辑如图 5.2 所示，首先对交通数据进行处理，剔除噪音数据，如：①车辆在半小时时间内横跨安徽省；②车辆丢失入口数据；③车辆丢失出口数据等，用筛选后的数据组成车辆 O-D 矩阵；从数据库内抽取重大交通事故信息，结合新闻中的高速公路事故信息，计算并输入路段的损毁概率。当选取贪心算法时，系统直接根据当前输入数据，进行贪心计算，注意本系统中将关键节点维护后的损毁率下降设为 0.1，具体数值可以根据实际应用调整；当选取基于社群划分的关键路段挖掘策略时，首先根据 O-D 流量信息计算社群分布，并给出划分结果（图 5.3）、网络模块化程度，之后对每一个社群内部进行贪心求解，这个求解过程并行执行，最后结合动态规划，求得最终解

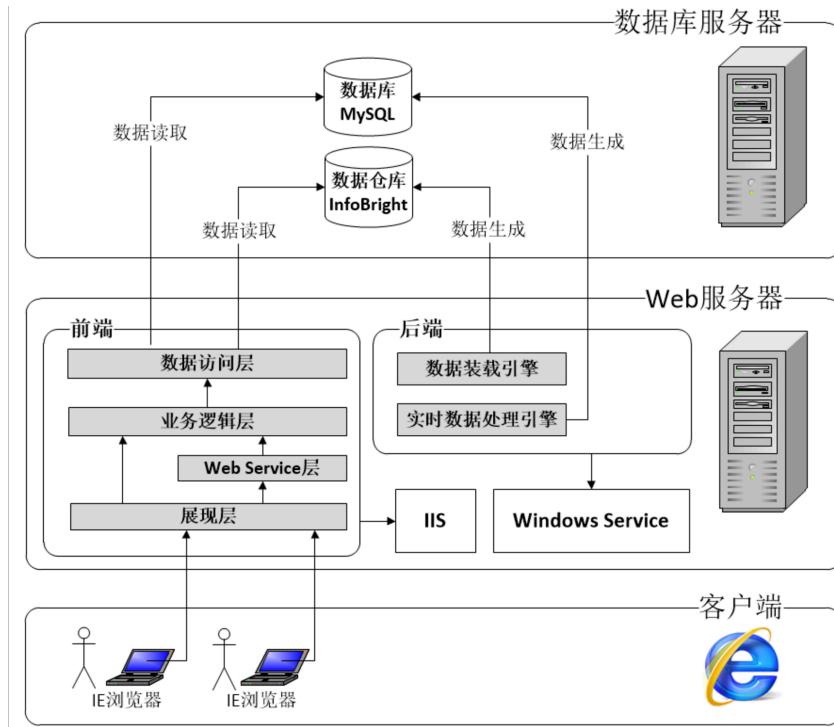


图 5.1 智能高速系统架构

集合（图5.4）。

### 5.3 界面功能展示

系统的输入有时间，时间区间，路段损毁概率，管理者对关键路段的维护效果。

对于静态关键路段挖掘，直接输出关键路段结果（如图5.4）；对于动态关键路段挖掘，系统输出两层信息，第一层是关键路段分群效果（图5.3）；第二层是关键路段选取结果（图5.4）。

目前系统只适用于安徽高速系统，但是可以扩展到任何具有收费站数据的中国高速公路上。

### 5.4 本章小结

本章介绍了一个面向高速公路的关键路段挖掘原型系统，给出了系统的流程图和样例图。该系统现在已经在安徽省高速公路网络上完

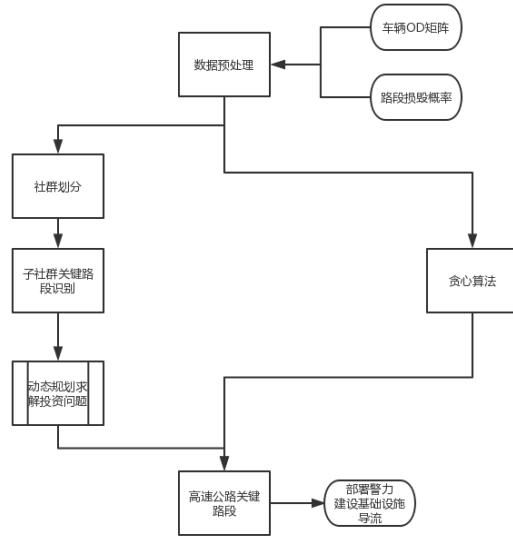


图 5.2 逻辑流程图



图 5.3 系统分群结果图

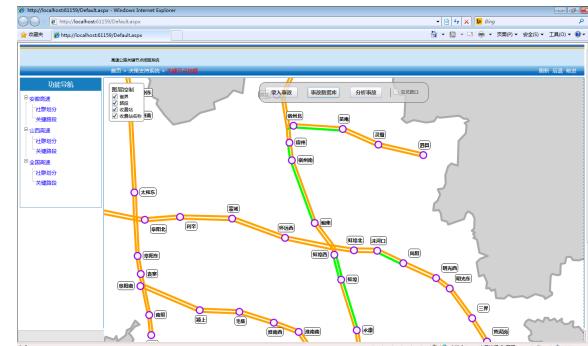


图 5.4 系统路段选取结果图

全实现。该系统可扩展性强，所以可以很快的复用到其他省份乃至全国高速公路网络中。



## 总结与展望

交通系统中的关键节点识别非常重要。在高速公路系统中，关键节点的损毁会对整个系统的性能造成显著的影响，带来重大的经济损失。所以识别关键节点，在发生事故或者自然灾害之前进行维护巩固，在发生事故后进行快速修复，维护网络完整性非常重要。

### 5.4.1 主要工作

本文的主要工作和创新点如下：

1) 结合高速公路特性，提出高速公路关键路段挖掘模型，证明模型的贪心可解性，给出贪心解法。

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；网络的“小世界特性”和“无标度特性”的发现掀起了网络科学持续 10 多年至今丝毫没有降温的研究热潮。网络科学研究的热点逐渐从早期发现跨越不同网络的宏观上的普适规律转变为着眼于从中观（社团结构、群组结构）和微观层面（节点、链路）去解释不同网络所具有的不同特征。重要节点的挖掘研究也逐渐转为以微观研究为主。然而，从管理者的角度来说，过于微观的研究又无法体现高速公路系统的宏观特性。所以在此本文提出一种结合宏观（目标函数基于宏观理念）微观（基于路段损毁的随机规划问题）的高速公路关键路段挖掘模型，有效的解决了微观研究无法很好的顾及整体的问题。

2) 结合高速公路的复杂网络，提出一种基于复杂网络社群划分的关键路段挖掘方法。

传统的复杂网络社群划分与高速公路不同，首先高速公路是一种

相当稀疏的复杂网络，网络中的拓扑结构特性不是很复杂；其次高速公路和普通的复杂网络不同，他的不同的节点之间具有物理空间距离，和其他复杂网络如社交网络中的距离概念不同。所以传统的复杂网络社群划分方法已经不再适用。在此引入可变权值方法，有效的解决了传统分群算法中的低分辨率特性以及极端退化特性。同时采用模拟退火思想加速模型收敛过程。

#### 5.4.2 未来工作展望

本文工作虽然具有一定的创新性和实用性，但仍然存在一些局限和不足，需要在今后的研究中进一步探讨和完善，主要包括以下几个方面：

- 1) 由于时间限制，目前只在安徽路网上做完了原型系统。下一步的工作是将系统复用到全国的高速公路网络。
- 2) 目前分群算法结合高速公路的物理空间特性，在社群划分效果上可能有一定的损失。随着中国交通建设的不断完善，高速公路监测数据也越来越丰富。希望之后可以结合新的高速数据集，进一步完善高速公路关键路段挖掘方法。

## 参考文献

- [1] 姚寿康, 曹小军 and 刘杰. “浅析高速公路养护技术”. 工程技术: 全文版, **2017**, (2): 00155–00155.
- [2] A Khadivi, Rad A Ajdari and M Hasler. “*Network community-detection enhancement by proper weighting.*” *Physical Review E Statistical Nonlinear & Soft Matter Physics*, **2011**, 83(2): 894–901.
- [3] Christian Yip, Phillip Fiorenzo, Kil Do Jung *et al.* “*A network-based congestion management model for Safety Service Patrol vehicle deployment*”. **2016**: 26–31.
- [4] Boris S. Kerner, Micha Koller, Sergey L. Klenov *et al.* “*The physics of empirical nuclei for spontaneous traffic breakdown in free flow at highway bottlenecks*”. *Physica A Statistical Mechanics & Its Applications*, **2015**, 438: 365–397.
- [5] Yacine Achour, Abderrahmane Boumezbeur, Riheb Hadji *et al.* “*Landslide susceptibility mapping using analytic hierarchy process and information value methods along a highway road section in Constantine, Algeria*”. **2017**, 10.
- [6] Xinsheng Song, Xiaoxiao Wang, L. I. Aizeng *et al.* “*Node Importance Evaluation Method for Highway Network of Urban Agglomeration*”. *Journal of Transportation Systems Engineering & Information Technology*, **2011**, 11(2): 84–90.
- [7] Li Wang, Pei Zhou Yan, Ying Hong Li *et al.* “*Signal sub-control-area division of traffic complex network based on nodes importance assessment*”. **2011**: 5606–5609.
- [8] Zhengwu Wang, Aiwu Kuang and Hejie Wang. “*Calculating Node Importance Considering Cascading Failure in Traffic Networks*”. *Research Journal of Applied Sciences Engineering & Technology*, **2013**, 5(1): 264–269.
- [9] Phillip Bonacich. “*Factoring and weighting approaches to status scores and clique identification*”. *The Journal of Mathematical Sociology*, **1972**, 2(1): 113–120.
- [10] Duanbing Chen, Linyuan Lü, Ming Sheng Shang *et al.* “*Identifying influential nodes in complex networks*”. *Physica A Statistical Mechanics & Its Applications*, **2012**, 391(4): 1777–1787.
- [11] Maksim Kitsak, Lazaros K. Gallos, Shlomo Havlin *et al.* “*Identification of influential spreaders in complex networks*”. *Nature Physics*, **2010**, 6(11): 888–893.
- [12] An Zeng and Cheng Jun Zhang. “*Ranking spreaders by decomposing complex networks*”. *Physics Letters A*, **2012**, 377(14): 1031–1035.
- [13] Shlomi Dolev, Yuval Elovici and Rami Puzis. “*Routing betweenness centrality*”. *Journal of the Acm*, **2010**, 57(4): 1–27.
- [14] Xue Qi Cheng, Fu Xin Ren, Hua Wei Shen *et al.* “*Bridgeness: A Local Index on Edge Significance in Maintaining Global Connectivity*”. **2010**, 10(10): 595–685.
- [15] Per Hage and Frank Harary. “*Eccentricity and centrality in networks*”. *Social Networks*, **1995**, 17(1): 57–63.

- [16] Miro Pu?nik. “*The diameter of the world wide web*”. *Knji?nica: revija Za Podro?je Bibliotekarstva in Informacijske Znanosti*, **1997**, 32(2): 155–159.
- [17] Vito Latora and Massimo Marchiori. “*Efficient Behavior of Small-World Networks*”. *Physical Review Letters*, **2001**, 87(19): 198701.
- [18] Leo Katz. “*A new status index derived from sociometric analysis*”. *Psychometrika*, **1953**, 18(1): 39–43.
- [19] Linyuan Lü and Tao Zhou. “*Link prediction in complex networks: A survey*”. *Physica A Statistical Mechanics & Its Applications*, **2011**, 390(6): 1150–1170.
- [20] Karen Stephenson and Marvin Zelen. “*Rethinking centrality: Methods and examples* ”. *Social Networks*, **1989**, 11(1): 1–37.
- [21] Michael Altman. “*Reinterpreting network measures for models of disease transmission* ”. *Social Networks*, **1993**, 15(1): 1–17.
- [22] R. Poulin, M. C Boily and B. R Mâsse. “*Dynamical systems to define centrality in social networks*”. *Social Networks*, **2000**, 22(3): 187–220.
- [23] K. I. Goh, B Kahng and D Kim. “*Universal behavior of load distribution in scale-free networks*”. *Physical Review Letters*, **2001**, 87(27 Pt 1): 278701.
- [24] Ulrik Brandes. “*A Faster Algorithm for Betweenness Centrality*”. *The Journal of Mathematical Sociology*, **2001**, 25(2): 163–177.
- [25] 周涛, 刘建国 and 汪秉宏. “*Notes on the Algorithm for Calculating Betweenness*”. *Chinese Physics Letters*, **2006**, 23(8): 2327.
- [26] G. Yan, T. Zhou, B. Hu et al. “*Efficient routing on complex networks*”. *Physical Review E Statistical Nonlinear & Soft Matter Physics*, **2006**, 73(2): 046108.
- [27] E Estrada and J. A. Rodríguezvelázquez. “*Subgraph centrality in complex networks.*” *Physical Review E Statistical Nonlinear & Soft Matter Physics*, **2005**, 71(5 Pt 2): 056103.
- [28] Phillip Bonacich and Paulette Lloyd. “*Eigenvector-like measures of centrality for asymmetric relations*”. *Social Networks*, **2001**, 23(3): 191–201.
- [29] T Martin, X. Zhang and M. E. Newman. “*Localization and centrality in networks.*” *Physical Review E*, **2015**, 90(5-1): 052808–052808.
- [30] Daijun Wei, Xinyang Deng, Xiaoge Zhang et al. “*Identifying influential nodes in weighted networks based on evidence theory*”. *Physica A Statistical Mechanics & Its Applications*, **2013**, 392(10): 2564–2575.
- [31] Sung Jin Kim and Ho Lee Sang. “*An Improved Computation of the PageRank Algorithm*”. In: *Proceedings of the 24th BCS-IRSG European Colloquium on IR Research: Advances in Information Retrieval*, **2002**: 73–85.
- [32] Li Zhang, Tao Qin, Tie Yan Liu et al. “*N-Step PageRank for Web Search*”. **2007**, 1(4): 1.
- [33] B. Sergey and P. Lawrence. “*The anatomy of a large-scale hypertextual Web search engine*”. *Computer Networks*, **1998**, 56(18): 3825–3833.

- [34] Peter Jacso. “Grim tales about the impact factor and the h-index in the Web of Science and the Journal Citation Reports databases: reflections on Vanclay’s criticism”. *Scientometrics*, **2012**, 92(2): 325–354.
- [35] Jianshu Weng, Ee Peng Lim, Jing Jiang *et al.* “TwitterRank: finding topic-sensitive influential tweeters”. In: *International Conference on Web Search and Web Data Mining, WSDM 2010, New York, Ny, Usa, February*, **2010**: 261–270.
- [36] Harish S Bhat and Bryan Sims. “InvestorRank and an Inverse Problem for PageRank”. *Electronic Theses & Dissertations*, **2012**.
- [37] A. M. Petersen, F. Wang and H. E. Stanley. “Methods for measuring the citations and productivity of scientists across time and discipline.” *Physical Review E Statistical Nonlinear & Soft Matter Physics*, **2010**, 81(2): 036114.
- [38] Ying Ding, Erjia Yan, Arthur Frazho *et al.* “PageRank for ranking authors in co-citation networks”. *Journal of the Association for Information Science and Technology*, **2009**, 60(11): 2229–2243.
- [39] Juan G. Restrepo, Edward Ott and Brian R. Hunt. “Characterizing the Dynamical Importance of Network Nodes and Links”. *Physical Review Letters*, **2006**, 97(9): 094102.
- [40] 李鹏翔, 任玉晴 and 席酉民. “网络节点(集)重要性的一种度量指标”. *系统工程*, **2004**, 22(4): 13–20.
- [41] 陈勇, 胡爱群 and 胡啸. “通信网中节点重要性的评价方法”. *通信学报*, **2004**, 25(8): 129–134.
- [42] 谭跃进, 吴俊 and 邓宏钟. “复杂网络中节点重要度评估的节点收缩方法”. *系统工程理论与实践*, **2006**, 26(11): 79–83.
- [43] Chavdar Dangalchev. “Residual closeness in networks”. *Physica A Statistical Mechanics & Its Applications*, **2006**, 365(2): 556–564.
- [44] Kiyotaka Ide, Loganathan Ponnambalam, Akira Namatame *et al.* *Risk Analysis and Quantification of Vulnerability in Maritime Transportation Network Using AIS Data*, **2015**: 139–151.
- [45] J. L. Moreno. “The sociometry reader.” *Revue Française De Sociologie*, **1961**, 2(4): 331.
- [46] Steffen Dereich and Peter Mörters. “Random networks with sublinear preferential attachment: The giant component”. *Annals of Probability*, **2013**, 41(1): 329–384.
- [47] C. M. Schneider, A. A. Moreira, Andrade Js Jr *et al.* “Mitigation of malicious attacks on networks.” *Proceedings of the National Academy of Sciences of the United States of America*, **2011**, 108(10): 3838.
- [48] Swami Iyer, Timothy Killingback, Bala Sundaram *et al.* “Attack Robustness and Centrality of Complex Networks”. *Plos One*, **2013**, 8(4): e59613.
- [49] 周涛, 傅忠谦, 牛永伟 *et al.* “复杂网络传播动力学研究综述”. *自然科学进展*, **2005**, 15(5): 513–518.
- [50] Linyuan Lü, Duan Bing Chen and Tao Zhou. “Small world yields the most effective information spreading”. *New Journal of Physics*, **2011**, abs/1107.0429(12): 825–834.
- [51] C. D Brummitt, R. M D’Souza and E. A Leicht. “PNAS Plus: Suppressing cascades of load in interdependent networks”.

- [52] X. L. Peng, X. J. Xu, X. Fu *et al.* “Vaccination intervention on epidemic dynamics in networks”. *Physical Review E Statistical Nonlinear & Soft Matter Physics*, **2013**, 87(2): 022813.
- [53] Phillip Bonacich. “Factoring and Weighting Approaches to Clique Identification”. *Journal of Mathematical Sociology*, **1972**, 2(1): 113–120.
- [54] Rui Yang, Bing Hong Wang, Jie Ren *et al.* “Epidemic spreading on heterogeneous networks with identical infectivity”. *Physics Letters A*, **2006**, 364(3): 189–193.
- [55] Linyuan Lü, Yi Cheng Zhang, Ho Yeung Chi *et al.* “Leaders in Social Networks, the Delicious Case”. *Plos One*, **2011**, 6(6): e21202.
- [56] M. E. Newman and M Girvan. “Finding and evaluating community structure in networks.” *Physical Review E Statistical Nonlinear & Soft Matter Physics*, **2004**, 69(2 Pt 2): 026113.
- [57] Fortunato and Santo. “Community detection in graphs”. *Physics Reports*, **2010**, 486(3–5): 75–174.
- [58] Chuanxiang Ren, Fasheng Liu and Li Zhou. “The Position of Connected-way for Highway and Urban Road Based on Node Importance Analysis Method”. *Journal of Shandong University of Science & Technology*, **2014**.
- [59] M. E. J. Newman. “Fast algorithm for detecting community structure in networks.” *Physical Review E Statistical Nonlinear & Soft Matter Physics*, **2004**, 69(6 Pt 2): 066133.
- [60] Zhi Wang and Jianzhi Zhang. “In Search of the Biological Significance of Modular Structures in Protein Networks”. *Plos Computational Biology*, **2007**, 3(3): e107.

## 附录 A 附件

### *pkuthss 文档模版最常见问题：*

在最终打印和提交论文之前,请将 pkuthss 文档类选项中的 **colorlinks** 替换为 **nocolorlinks**, 因为图书馆要求电子版论文的目录必须为黑色, 且某些教务要求打印版论文的文字部分为纯黑色而非灰度打印。

**\cite**、**\parencite** 和 **\supercite** 三个命令分别产生未格式化的、带方括号的和上标且带方括号的引用标记:**test-en**, [**test-zh**]、[**test-en**, **test-zh**]。

若要避免章末空白页, 请在调用 pkuthss 文档类时加入 **openany** 选项。

如果编译时不出参考文献, 请参考 **texdoc pkuthss**“问题及其解决”一章“其它可能存在的问题”一节中关于 **biber** 的说明。



## 致谢

### *pkuthss* 文档模版最常见问题：

在最终打印和提交论文之前,请将 `pkuthss` 文档类选项中的 `colorlinks` 替换为 `nocolorlinks`, 因为图书馆要求电子版论文的目录必须为黑色, 且某些教务要求打印版论文的文字部分为纯黑色而非灰度打印。

`\cite`、`\parencite` 和 `\supercite` 三个命令分别产生未格式化的、带方括号的和上标且带方括号的引用标记:`test-en`,`[test-zh]`、`[test-en, test-zh]`。

若要避免章末空白页, 请在调用 `pkuthss` 文档类时加入 `openany` 选项。

如果编译时不出参考文献, 请参考 `texdoc pkuthss`“问题及其解决”一章“其它可能存在的问题”一节中关于 `biber` 的说明。



# 北京大学学位论文原创性声明和使用授权说明

## 原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名： 日期： 年 月 日

## 学位论文使用授权说明

(必须装订在提交学校图书馆的印刷本)

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
- 因某种特殊原因需要延迟发布学位论文电子版，授权学校在一年/两年/三年以后在校园网上全文发布。

(保密论文在解密后遵守此规定)

论文作者签名:                   导师签名:                   日期:       年      月      日