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

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 10 日

版权声明

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，社会交通需求也日益增大。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，简单的交通控制技术已经不能满足需求。传统智能交通研究以路网中的单个空间位置点为研究对象，运用动力学，统计学，仿真学与机器学习相关理论对其交通流行为进行分析。随着交通系统的不断发展，交通系统逐渐呈现网络化态势，路网中各个节点之间紧密联系，单个空间位置点的研究已经不足以描述整个高速公路路网的宏观特性。随着交通事故逐渐成为交通系统的瓶颈所在，智能交通系统的研究出现了新的需求：如何找到交通系统中的关键节点，通过对这些关键节点进行处理，以达到减少交通路网瘫痪率、增加路网运行稳定性的目的。

在网络化的智能高速公路系统中，关键节点的研究集中在 1）对交通网络拓扑结构的研究；2）对交通网络信息进行统计研究；3）利用传播动力学，对微观站点进行相关研究。我们的目的是找到关键节点，在整体网络中提升运行效率，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）基于传统统计学，并在统计学的基础上，利用数据挖掘方法研究关键节点，路网中的车流流量随时间变化而变化，该方法只能静态分析路网关键节点，无法分析关键节点随着时间／路网流量的变化规律；方法 3）集中研究微观领域的交通特性，对整体路网的研究意义不大。在此我们提出一种基于宏观高速公路网络的目标模型，根据选取节点对模型的影响，选取关键节点。为了深入挖掘高速公路关键节点特性，我们提出了一个较复杂的概率求解模型。这个模型直接求解时间复杂度高，不符合智能交通系

统的实时应用需求。基于复杂网络的小世界特性，在此引入高速公路网络的分群算法，进行分治处理

因此，本文从智能交通的实际应用需求出发，针对现有关键节点挖掘研究方法的不足，深入进行两个方面的研究：提出一种描述语高速公路的关键节点与高速公路的通行状况之间联系的函数模型,针对现有复杂网络关键节点研究的局限性，从宏观层面提出目标函数，结合高速公路的实时路况，研究实时高速公路中关键节点的位置;提出一种基于高速公路路网的社群划分方法。针对现有社群划分方法中的分辨率限制与极端退化特性，结合高速公路路网的特点，建立适用于高速公路的社群划分模型，在一定程度上解决传统方法中的分辨率限制与极端退化特性。

本文的贡献主要有以下几点：

（1）提出一种高速公路路网关键节点挖掘模型，突破现有复杂网络的局限性

（2）提出一种基于高速公路路网的社群划分模型，降低高速公路关键节点挖掘复杂度，使得方法可以实际应用。

关键词：复杂网络，高速公路，关键节点，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

KEYWORDS: Complex network, Highway， Key node， Community

目录

插图

第一章　引言

1.1　研究背景

1.1.1　高速公路交通研究背景

交通系统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，对于交通系统管理人员来说，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%。毫无疑问，如何高效处理交通事故和预测事故发生点，一旦世故发生，最大限度地减少其影响是一个核心问题。第三，资源相对有限，尤其是中国高速公路正在逐渐走向免费，因此很难全面建立新的基础设施。同时，运输系统的有效性也越来越依赖于一个国家的处理紧急情况的能力（例如，大规模疏散和安全增强）。一个国家的技术竞争力，其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。

在过去的二十年中，智能交通系统（ITS）已成为一种提高交通系统的性能，提高行车安全有效的方式，并且为旅客提供更多的选择。上述一些问题可以通过智能交通系统，实施新的交通政策来解决。这

些政策是基于高速公路的特性提出的，高速公路的特性主要有网络性，小世界特性，社区结构特性、动态性以及周期性。

网络性

纵横交错的道路构成了复杂的交通路网，这使得交通系统具有了网络性质。网络中，不同的收费站构成了节点，相邻收费站之间的道路构成了网络中的边，节点之间通过车辆来交流。高速公路网络化，使得交通系统中的车流行为更加复杂，这对交通研究方法提出了更高的要求。网络化所引发的复杂性在于，路网中不同空间位置的交通流行为并非孤立产生，而是相互间存在着紧密关系，路网愈加庞大，关系愈为复杂。两个路网中不同空间位置的交通流之间存在着紧密关联，例如较多的车流从某些特定的入口进入路网，又从某些特定的出口流出路网，并且不同出口共享着某些车流来源。然而，传统的以单位置点为研究对象的交通流分析方法并不能有效利用车流之间的关联信息，因此它们已经无法再适用于网络化的交通系统。交通流之间的关联性促生了从路网视角进行全局交通流分析的需求，要求将路网中多节点的交通流行为同时进行学习。

小世界特性

小世界特性（Small world theory）又被称之为是六度空间理论或者是六度分割理论（Six degrees of separation）。小世界特性指出：社交网络中的任何一个成员和任何一个陌生人之间所间隔的人不会超过六个。在高速公路网络中，小世界特性的表现有所不同：网络中绝大部分车辆的跳数（车辆旅行途中经过的道路数量）小于 6 个.

无标度特性

现实世界的网络大部分都不是随机网络，少数的节点往往拥有大量的连接，而大部分节点却很少，节点的度数分布符合幂率分布，而

这就被称为是网络的无标度特性（Scale-free）。将度分布符合幂律分布的复杂网络称为无标度网络。在高速公路网络中，统计发现少量节点占有着大多数车辆。（上图）

社区结构特性

人以类聚，物以群分。复杂网络中的节点往往也呈现出集群特性。例如，社会网络中总是存在熟人圈或朋友圈，其中每个成员都认识其他成员。集群程度的意义是网络集团化的程度；这是一种网络的内聚倾向。连通集团概念反映的是一个大网络中各集聚的小网络分布和相互联系的状况。在高速公路网络中，这个特性体现在：高速公路的节点组成一个个社团，这些社团绝大部分车辆都驶向社团内部。

动态性以及周期性

交通路网是一种动态系统，随着时间的变化，其内部的交通流规律与运行模式都在不断变化。交通现象具有周期性，典型的例子是以日为周期的交通流交替运行模式。

1.1.2　关键节点挖掘研究背景

复杂网络的重要节点是指相比网络其他节点而言, 能够在更大程度上影响网络的结构与关键词功能的一些特殊节点. 近年来, 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, 更因为其广泛的实际应用价值. 由于应用领域极广, 且不同类型的网络中节点的重要性评价方法各有侧重, 学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网络. 其中, 节点代表系统的各种构成要素, 节点间的连边表示要素之间的联系. 最复杂的人类社会系统就可以用一个社会网络刻画, 节点是人, 人与人之间的各种关系构成社会网络中的链接. 本文最核心的研究问题就是如何识

别这些重要的节点. 所谓的重要节点是指相比网络其他节点而言能够在更大程度上影响网络的结构与功能的一些特殊节点. 所以传统的研究方向也主要是在这两个方向上进行。

在传统的研究中，网络结构是指整个网络的成网性，包括度空间分布、平均距离、连通性、聚类系数、度相关性等。网络功能涉及网络的抗毁性、传播、同步、控制等。

重要节点一般数量非常少, 但其影响却可以快速地波及到网络中大部分节点. 例如, 在对一个无标度网络的蓄意攻击中, 少量最重要节点被攻击就会导致整个网络瓦解; 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。这一转变, 是因为随着研究的深入, 人们发现宏观指标不能很好表现网络结构和功能上的特征, 真正精细可靠的解释, 哪怕是针对宏观现象, 也必须立足于微观上的深入认识. 类似地, 多年以前, 一批学者就提倡关注网络结构和功能的相互影响 [10], 但是早期的研究都集中在网络宏观或者中观上的一些特征与网络具体功能表现之间的关系, 所得到的一些结论, 譬如“热力学极限下无标度网络传染病 (SIS 模型) 没有阈值, 随机网络有阈值” [11] 等, 都只是一些统计上有意义, 大多数情况下正确, 定性上可以部分解释, 定量上无法开展预测的结果. 这是因为宏观指标以及基于宏观量的运算, 已经把很多个体的特征进行了“平均化”, 而一些非常关键个体的表现被这种“平均化”淹没了. 还是回到刚才的例子, 如果我们从个体出发, 比如仔细考虑一个节点在 SIS 传播动力学中可能的自维持特性, 就会得到颠覆性的结论: “热力学极限下随机网络上的 SIS 模型也没有阈值”[12]. 可见, 基于微观层面, 即节点个体的分析, 有望揭示网络功能上精细入微的特征. 总之, 随着网络科学研究从整体宏观

到个体微观的转变,重要节点的排序和挖掘已成为近年来的研究热点。

图 xxx 给出了复杂网络在传统方法中的样例。

上述复杂网络的关键节点方法经过不断的研究与发展，已经成功应用于众多研究领域中去。然而这些方法都有一定的局限性，或者是局限于微观节点，忽略了宏观节点之间的关系；或者是只关注路网的拓扑结构，没有关注路网的内在车流信息。

1.1.3　社群划分研究背景

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”[] 一; 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 []. 图 [] 中给出了上述统计特性的直观描述.

传统的研究方法主要分 1）基于划分的社区挖掘方法，即先找出社区间的所有链接，接着将它们全部删除，最后每个连通分支对应着一个社区；2）基于模块性优化的社区挖掘方法，即提出了一个用于刻画网络社区结构优劣的量化标准，被称之为模块函数 Q；3）基于标签传播的社区挖掘方法，即它没有特定的目标函数，而是通过一种直觉、富有启发的思想推断社区结构和设计算法. 标签传播类方法的启发式规则为“在具有社区结构的网络中，任一结点都应当与其大多数邻居在同一个社区内”.4）基于动力学的社群划分方法，比如说基于

Markov 随机游走理论的启发式求解策略；5）基于仿生计算的社区挖掘方法，主要包括蚁群算法和遗传算法。

上述研究或者基于没有实体的非物理复杂网络，如社交网络，或者具有相应的缺陷，如基于模块性优化方法中的分辨率限制与极端退化特性。

1.2 研究内容

综上所述，本文从交通实际问题的角度出发，针对现有复杂网络关键节点挖掘技术的不足，深入开展下述两项研究内容：

（1）提出一种度量高速公路节点重要性的研究模型，可以在宏观层面反映这些节点对高速公路网络的影响高，挖掘高速公路关键节点

（2）提出一种结合高速公路网络特性的社群划分算法，达到较强的收敛性与低误差。

1.3 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。第二章介绍了复杂网络关键节点研究的相关工作，结合交通问题的特点分析了现有方法的优势与不足。第三章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。从第四章开始的后续章节将论述本文的主要研究内容。第四章提出了一种复杂网络关键节点挖掘模型，给出了详尽的理论分析，并在多个数据集下进行了验证。第五章提出了一种基于高速公路交通网络的社群划分模型，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。第六章给出了混合模型在真实交通场景下的应用实例。第七章给出了全文的总结与未来工作展望。

第二章　复杂网络关键节点挖掘相关研究

复杂网络的重要节点是指相比网络其他节点而言, 能够在更大程度上影响网络的结构与关键词功能的一些特殊节点. 近年来, 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, 更因为其广泛的实际应用价值. 由于应用领域极广, 且不同类型的网络中节点的重要性评价方法各有侧重, 学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网络. 其中, 节点代表系统的各种构成要素, 节点间的连边表示要素之间的联系. 最复杂的人类社会系统就可以用一个社会网络刻画, 节点是人, 人与人之间的各种关系构成社会网络中的链接. 本文最核心的研究问题就是如何识别这些重要的节点. 所谓的重要节点是指相比网络其他节点而言能够在更大程度上影响网络的结构与功能的一些特殊节点. 所以传统的研究方向也主要是在这两个方向上进行。

在传统的研究中，网络结构是指整个网络的成网性，包括度空间分布、平均距离、连通性、聚类系数、度相关性等。网络功能涉及网络的抗毁性、传播、同步、控制等。

2.1　基于节点临近

该方法是最简单直观的方法, 度中心性考察节点的直接邻居数目, 半局部中心性考虑了节点 4 层邻居的信息. k-壳分解可以看作度中心性的一种扩展, 它根据节点在网络中的位置来定义其重要性, 认为越是在核心的节点越重要.

2.1.1 度中心性

社会网络分析中, 节点的重要性也称为“中心性”, 其主要观点是节点的重要性等价于该节点与其他节点的连接使其具有的显著性 []. 度中心性 (degree centrality)[] 认为一个节点的邻居数目越多, 影响力就越大, 这是网络中刻画节点重要性最简单的指标. 节点 vi 的度, 记为 ki, 是指与 vi 直接相连的节点的数目, 是节点最基本的静态特征. 在有向网络中, 根据连边的方向不同, 节点的度有入度和出度之分. 在含权网络中节点度又称为节点的强度 (strength), 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 [], 它认为一个节点的度越大, 能直接影响的邻居就越多, 也就越重要. 值得注意的是, 不同规模的网络中有相同度值的节点有不同的影响力, 为了进行比较, 定义节点 vi 的归一化度中心性指标为:

加入公式

其中, ki iaij , aij 即网络邻接矩阵 A 中第 i 行第 j 列元素, n 为网络的节点数目, 分母 n1 为节点可能的最大度值. 在有向网络中入度和出度有不同的意义 (如社交网络中入度代表受欢迎程度, 出度代表合群程度), 一般会分别计算入度和出度的中心性.

度中心性指标拥有简单、直观、计算复杂度低等特点. 在网络鲁棒性和脆弱性研究中, 针对无标度网络或指数网络, 如果攻击前一次性选择若干个攻击目标, 采用度中心性指标的攻击效果比介数中心性、接近中心性、特征向量中心性要好 (参见 6.1 节). 度中心性指标的缺点是仅考虑了节点的最局部的信息, 是对节点最直接影响力的描述, 没有对节点周围的环境 (例如节点所处的网络位置、更高阶邻居等) 进行更深入细致地探讨, 因而在很多情况下不够精确.

2.1.2　半局部中心性

度中心性指标计算方便简单, 但实际效果欠佳. 基于全局信息的方法, 如在下一节中介绍的介数中心性和接近中心性指标, 虽然具有较好的刻画节点重要性的能力, 但计算复杂度太高, 难以在大规模网络上使用. 为了权衡算法的效率和效果, Chen 等人 [] 提出了一种基于半局部信息的节点重要性排序方法, 简称半局部中心性 (semi-local centrality). 首先定义 N(w) 为节点 vw 的两层邻居度，其值等于从 vw 出发 2 步内可到达的邻居的数目, 然后定义

插入公式

其中 (j) 表示节点 vj 的一阶邻居节点的集合. 最终节点 vi 的局部中心性定义为

插入公式

可见, 半局部中心性涉及了节点的四阶邻居信息. 文献 [] 用 D-S 证据理论 (参见 5.6 节) 将本方法推广到了含权网络. 文献 [] 指出半局部中心性方法的计算复杂度随网络规模线性增长, 消耗非常少的计算时间, 就能够得到远好于度中心性和介数中心性的排序结果. 近期, Chen 等人 [] 还提出一种针对有向网络的半局部算法 (ClusterRank), 该算法不仅考虑了邻居节点的数量, 还考虑了聚类系数对信息传播的影响: 聚类系数越大越不利于信息的广泛传播. 两个数据集上的实验结果显示 Cluster- Rank 算法优于 PageRank 和 LeaderRank 算法, 并且计算复杂度最低.

2.1.3　k-壳分解法

度中心性仅考察节点最近邻居的数量, 认为度相同则重要性相同. 然而, 近期的一些研究表明在刻画节点重要性的时候节点在网络中的位置也是至关重要的因素. 在网络中, 如果一个节点处于网络的核心位置, 即使度较小, 往往也有较高影响力; 而处在边缘的大度节点

影响力往往有限. 基于此, Kitsak 等人 [] 提出用 k-壳分解法 (k-shell decomposition) 确定网络中节点的位置, 将外围的节点层层剥去, 处于内层的节点拥有较高的影响力. 这一方法可看成是一种基于节点度的粗粒化排序方法. 具体分解过程如下 []: 网络中如果存在度为 1 的节点, 从度中心性的角度看它们就是最不重要的节点. 如果把这些度为 1 的节点及其所连接的边都去掉, 剩下的网络中会新出现一些度为 1 的节点, 再将这些度为 1 的节点去掉, 循环操作, 直到所剩的网络中没有度为 1 的节点为止. 此时, 所有被去掉的节点组成一个层, 称为 1-壳 (记为 ks=1). 对一个节点来说, 剥掉一层之后在剩下的网络中节点的度就叫该节点的剩余度. 按上述方法继续剥壳, 去掉网络中剩余度为 2 的节点, 重复这些操作, 直到网络中没有节点为止. 更广泛地, 可定义初始度为 0 的孤立节点属于 0-壳, 即 ks=0. 网络中的每一个节点属于唯一的一层, 显然所有节点均满足 k ks . 图 1 给出一个 k-壳分解的示例. 其中 (a) 为原网络,(b), (c), (d) 分别表示 1-壳, 2-壳和 3-壳. 可见, 大度节点有可能因处于核心位置而拥有较大的 ks 值 (如图 1(d) 中的深色节点), 也可能因为处于边缘而具有较小的 ks 值 (如图 1(b) 中的深色节点). 在这个方法下, 大度节点不一定是重要节点.

插入图

k-壳分解法计算复杂度低, 在分析大规模网络的层级结构等方面有很多应用. 然而, 此方法也有一定局限性. 第一, k-壳分解法有很多不能发挥作用的场景. 比如在树形图, 规则网络和 BA 网络 [] 中, 所有 (或大部分) 节点都会被划分在同一层. 更极端的例子是星形图, 显然中心节点有最强的传播能力, 但是 k-壳分解的时候, 星形网络的所有节点会被划分在同一层 (ks=1). 第二, k-壳分解法的排序结果太过粗粒化, 使得节点的区分度不大. k-壳分解法划分的层级比度中心性方法划分的层级少很多, 很多节点处在同一层上, 它们之间的重要性难以比较. 第三, k-壳分解法在网络分解时仅考虑剩余度的影响, 这相当于认

为同一层的节点在外层都有相同的邻居数目, 显然不合理. Zeng 等人 [] 提出了在每一步剥去一部分外围节点之后, 同时考虑节点剩余的邻居数 kr 和节点 i 已经移除的邻居数 k e 的方法, 定义节点 v 的混合度 ii 为 kr ke , 根据新的混合度值对网络继续分层. ii 这种采用混合度值的 k-壳分解法能够很好地区分树形图以及 BA 网络中不同节点的传播能力, 并且分层的层数大大增加 (甚至可超过度中心性), 提高了节点传播能力的区分度. 另外, Liu 等人 [] 指出壳数相同的节点传播能力差距可能很大, 并提出了一种可以进一步区分具有相同壳数的节点的传播能力的排序方法, 从而较 Kitsak 等人 [] 的方法有所进步; Hu 等人 [] 将 k-壳分解法与社区结构相结合, 提出一种改良指标, 在 SIR 模型上的实验表明该方法较 Kitsak 等人的方法略佳.

2.2　基于路径临近

在交通、通信、社交等网络中存在着一些度很小但是很重要的节点, 这些节点是连接几个区域的“桥节点”, 它们在交通流和信息包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 []. 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 [].

2.2.1 离心中心性

在连通网络中, 定义 dij 为节点 vi 与 vj 之间的最短路径长度, 也称最短距离, 一个节点 vi 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 [], 即:

插入公式

网络直径定义为网络 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. 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.

2.2.2 接近中心性

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 vi 到网络中其他节点的平均最短距离:

插入公式

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

插入公式

上面定义的缺点是仅能用于连通的网络中, 文献 [] 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:

插入公式

如果节点 vi 和 vj 之间没有路径可达则定义 dij , 即 1dij 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.

2.2.3　Katz 中心性

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 []. 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 vi 相距有 p 步长的节点, 对 vi 的中心性的贡献为 sp(s(0,1) 为一个固定参数). 设 lp为从 ij 节点 vi 到 vj 经过长度为 p 的路径的数目. 显然 A2 l 2 a a , 其中元素 l 2即从节点 v 到 ij k ik kj ij ivj 经过的边数为 2 的路径的数目, 同理我们可以得到 A3, A4 · · · Ap · · · , 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:

插入公式

其中,I 为单位矩阵.K 矩阵中第 i 行 j 列对应的元素 kij 实际上就是我们所熟知的节点 vi 和 vj 的 Katz 相似性 []. 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 vj 的 Katz 中心性为矩阵 K 第 j 列元素的和:

插入公式

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 vi 与 vj 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 [] 来定义节点中心性.

2.2.4 信息指标

信息指标 (information indices)[] 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 (vi, vj) 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 qij. 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/qij 相当于以 2 个节点 vi 和 vj 为两端点的电阻值 (qij 相当于电导)[], 于是我们可以通过计算矩阵 R(r)(DAF)1 获得 q, 其中 ij ijD 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 (vi, vj) 间通过所有路径能够传播的信息总量为

插入公式

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

插入公式

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=(mij) 来表示网络中所有节点之间的关系, M 的每一个元素 mij 刻画了节点 vi 和 vj 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 vi 的重要性可表示为 Centrality(i)　j m . 由此可见, 只要我们能够给出 ij 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.

2.2.5　介数中心性

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 vi 的介数定义为

公式

其中, g 为从节点 v 到 v 的所有最短路径的数目, gstst st 为从节点 vs 到 vt 的 gst 条最短路径中经过 vi 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 n1, 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:

介数中心性可用于设计网络的通信协议、优化网络部署、检测网络瓶颈等. 王延庆 [] 将介数应用于负载网络, 提出用过载函数法研究网络的连接失效问题. 此外, Goh 等人 [] 提出的负载中心性 (traffic load centrality) 采用类似网络中信息包传递的机制: 每一对节点之间沿着最短路径传输一个单位的网络流, 如果最短路径不止一条, 则在几条最短路径的分叉处将网络流平均分配到这些最短路径上. 忽略时延, 网络中所有节点对之间都互不干扰地传输一个单位的信息流时, 一个

节点上传输过的网络流的数量称为该节点的负载. 一个节点的负载越大, 该节点就越重要. 介数中心性的计算时间复杂度较高, 使其在实际应用中受到限制, 相关讨论可参见文献 [].

2.2.6 流介数中心性

介数中心性仅考虑网络流通过最短路径传输. Yan 等人 [] 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 vs 到目标节点 vt 的最大流量是指 vs 与 vt 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)[] 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 vi 的流介数中心性为

插入公式

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.

2.2.7 随机游走介数中心性

从源节点 v 到目标节点 v 的随机游走的过程中当 i=s 或者 t 的时候,IsIt1. 该方法计算复杂度 st ststt 经过 vi 的次数可表征 vi 的重要性. 基于此, Newman[] 提出了基于随机游走的介数中心性算法 (random walk betweenness centrality). 在随机游走过程中短的路径计数次数较多, 相当于赋予其更高的权重. 在随机游走过程中, 如果网络流不断地从一个节点来回经过无疑会提高这个节点的介数中心性, 但是这样的

刻画实际上是毫无意义的. 为了避免这种偏差, 约定在一次随机游走中如果网络流两次分别从相反方向经过某一节点, 则它们对这个节点的介数中心性的贡献相互抵消. 于是, 节点 vi 的随机游走介数中心性可表示为

公式

2.2.8　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 (vs, vt), 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 vu 转发给节点 vv 一个从源节点 vs 到目标节点 vt 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 vu 指向 vv 的有向边. 用 s,t(u) 其中 I i 程中, 经过节点 vi 的次数. 事实上, 如果我们将网络看成一个电阻网络, 每条边的电阻值为 1, 从节点 vs 表示信息包从 v 到 v 的传递过程中, 经过节点的 v stst st u 概率, 显然 s,t(s)=s,t(t)=1, 用 Preds,t(v) 表示 R(s, t) 中节点 vv 的直接前驱的集合, 那么有向无环图 R(s, t) 中经过任意一个节点 vv 的概率可由下式得出:

gon shi

2.2.9　子图中心性

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)[]. 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 vi 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

公式

其中 at 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, a1 0;t=2 时, a2 为节点 v 的度值, 即 ii ii ia2 k , 此时, 子图中心性就等价于度中心性; t3 ii i 时, at 表示从点 v 开始, 经过 t 条边又回到 v 的路径 iii i 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

公式

其中,jj1,2,,n为邻接矩阵 A 的特征值,j 是 所对应的特征向量, i 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 []. 另外, 子图中心性的方法还能够应用于网络中模体的检测 [].

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.

2.3.1　特征向量中心性

特征向量中心性 (eigenvector centrality)[] 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 xi 为节点 v 的重要性度量值, 则:

公式

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, 是需要防范的关键节点 []. 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 [] 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), 有的人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=Ax+e, 其中 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 (Localiztion), 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 [] 对特征向量中

心性进行改进, 提出在计算节点 vi 的分值时, 求和中其邻居的分值不再考虑节点 vi 的影响.

2.3.2 累计提名

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) [] 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 pit 为节点 vi 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 0 1), 每个时间步每个节点从所 i 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 时刻的累积提名为

公式

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 vi 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 vi 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.

2.3.3 PageRank 算法

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 vi 在 t 时刻的 PR 值为

公式

jj 以 c 的概率均分给网络中所有节点, 以 1c 的概率均分给它指向的节点. 该过程实际上是考虑到了现实中网络用户除了通过超链接访问页面之外, 还可以通过直接输入网址的形式对网页进行访问的行为, 从而保证了即使是没有任何入度的网页也有机会被访问到. 其实质是将有向网络变成强连通的, 使邻接矩阵成为不可约矩阵, 保证了特征值 1 的存在. 由此可得含参数 c 的 PageRank 算法: 值都达到稳定时为止. 公式 (27) 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

公式

参数 c 的取值要视具体的情况而定. c 取值越大收敛越快, c=0 时回到公式 (27). c 取值越大算法的有效性越低, c=1 时所有节点都有相同的 PR 值. 针对万维网的网页排序, 以前的研究显示, c=0.15 是一个比较好的参数. PageRank 算法作为谷歌搜索引擎的核心算法, 它在商业应用上的极大成功激发了人们深入研究 PageRank 的热忱, 研究者们提出了一系列基于 PageRank 的改进算法. 例如 Kim 和 Lee[] 为了避免悬挂节点囤积 PR 值的问题, 将每一步到达悬挂节点的 PR 值平均分

给网络中的 n 个节点, 即将概率转移矩阵中悬挂节点所在的列的 n 个元素修改为 1/n; PageRank 中从一个网页上的链接中挑选下一个访问目标时是等概率的, Zhang 等人 [] 认为这 n 个目标网页出度越大的越有可能被点击, 并提出 N-step PageRank 算法用以描述这一思想. 2012 年 Brin 和 Page[] 以相同的题目重新出版了当年提出 PageRank 算法的博士学位论文, 在文中他们对这十几年的网页排序算法进行了回顾, 并就如何用 PageRank 实现大规模搜索进行了深入讨论. 另外, 作为有向网络节点排序最经典的算法, PageRank 及其改进算法广泛应用于其他领域, 如对期刊的排序 []、对社交网络上用户的排序 []、对风投公司 (VC) 的排序 []、对科学论文的排序 [] 73] 以及科学家影响力的排序 [] 77] 等.

2.3.4 LeaderRank 算法

PageRank 算法中, 每一个节点的随机跳转概率都是相同的, 即从任意网页出发, 采用输入网址来访问其他网页的概率相等. 然而在现实中人们在内容丰富的热门网页 (出度大的节点) 上浏览的时候选择使用地址栏跳转页面的概率要远小于浏览信息量少的枯燥网页 (出度小的节点). 另一方面, PageRank 算法中的参数 c 的选取往往需要实验获得, 并且在不同的应用背景下最优参数不具有普适性 []. LeaderRank 算法的出现很好地解决了以上两个问题. 在有向网络的随机游走过程中, 通过添加一个背景节点以及该节点与网络中所有节点的双向边来代替 PageRank 算法中的跳转概率 c, 从而得到一个无参数且形式上更加简单优美的算法. LeaderRank 算法在某一页面输入网址访问下一个页面的概率就相当于从这个页面访问背景节点的概率, 这个概率和一个网页上的链接数负相关, 链接数越多, 网页的内容越丰富, 越倾向于从本地的链接访问, 访问背景节点的概率就越低. 注意, 背景节点的存在同样保证了网络的强连通性. 初始时刻给定网络中除背景节点 vg

以外的其他节点单位资源, 即 LRi01， ig; LRg00. 经过以下的迭代过程直到稳态:

公式

LeaderRank 算法在衡量社会网络中节点的影响力等方面有非常优异的表现 [], 因此得名. 实验发现 LeaderRank 比 PageRank 在很多方面表现得更好:(1) 与 PageRank 相比收敛更快 []; (2) 能够更好地识别网络中有影响力的节点, 挖掘出的重要节点能够将网络流传播的更快更广; (3) 它在抵抗垃圾用户攻击和随机干扰方面相比 PageRank 有更强的鲁棒性. 这些优点使得 LeaderRank 算法广受关注. 标准 LeaderRank 算法中背景节点和所有节点的连接都一样, Li 等人 [] 对此提出改进, 认为从背景节点出发访问其他节点时, 入度大的节点应该有更高的概率被访问到. 如果一个节点 v 的入度为 kin , 则背景节点 ii 指向 v 的边权 w (kin), 网络其他节点之间的连接 i gi i 的权重都等于 1, 由此得到改进后的 LeaderRank 的迭代公式为

公式

这种改进更加重视网络中的大度节点, 在多个数据集上的实验发现新方法比标准的 LeaderRank 的性能在多个方面均有提升. 虽然这一方法的提出最初是为了提升 LeaderRank 算法在无权网络中的排序效果, 但是这种思路也可以应用到含权网络中, 关于 LeaderRank 算法在含权网络中的扩展参见 5.5 节

2.3.5　HITs 算法

一个网络中不同类型的节点功能不同, 每个节点的重要性往往不能由单独的一个指标给出, HITs 算法 [] 赋予每个节点两个度量值: 权威值 (Authorities) 和枢纽值 (Hubs). 权威值衡量节点对信息的原创性, 枢纽值反映了节点在信息传播中的作用. 枢纽页面是那些指向权威页面的、链接数较多的页面, 反映网页上链接的价值. 节点的权威值等

于所有指向该节点的网页的枢纽值之和, 节点的枢纽值等于该节点指向的所有节点的权威值之和. 因而, 节点若有高权威值则应被很多枢纽节点关注, 节点若有高枢纽值则应指向很多权威节点. 简单地说, 权威值受到枢纽值的影响, 枢纽值又受到权威值的影响, 最终通过迭代达到收敛. 在一个包含 n 个节点的网络中, 定义 ait 和 hit 分别为节点 vi 在时刻 t 的权威值和枢纽值, 于是在每一时间步的迭代中:

公式

HITs 首次用不同指标同时对网络中的节点进行排序, 具有开创意义. HITs 除了可以用于确定一个节点上多个相互关联的属性, 还可以处理更复杂的排序问题 [], 譬如在信誉评价系统中如何评价用户的信誉度以及产品的质量 []. 这类评价系统通常包含两类节点 (用户和产品), 信誉排序问题解决的是包含两类节点的各自的排序问题. 与 HITs 类似的是两类节点的分数值也是相互影响的, 最终通过迭代寻优获得两类节点的排序值. 例如文献 [] 利用这种思路提出一种可以有效抵抗恶意评分的排序方法, 该方法认为一个商品得到的打分反映了这个商品的质量, 自然地, 应该给可信度高的用户更大的权重; 反过来, 一个用户打分的可信度, 可以用他的打分和商品质量的接近程度来衡量.

需要指出的是, 特殊的网络结构会影响 HITs 算法、PageRank 算法这类应用邻居之间相互传递打分值进行排序的方法的表现. 例如万维网中广泛存在紧密连接社团 (tightly-knit community), 社团内节点间非常紧密的链接关系会使这些节点的权威值和枢纽值相互增强 (mutual reinforcement), 从而使网页的排序结果更倾向于将社团内部的页面排在前面而偏离搜索的主题, 出现主题漂移 (topic draft) 现象 [].

2.3.6 自动信息汇集算法

Kleinberg 与其合作者对 HITs 算法进行了改进, 提出了自动信息汇集 (automatic resource compilation,ARC) 算法 []. HITs 算法仅考虑网

页之间的链接关系 (即仅考虑网络结构), ARC 算法在此基础上, 还考虑了页面内容与搜索主题的相关性, 给每个链接赋予不同的权值, 提高页面排序的真实可靠性. 算法的具体过程如下: 取一个含有搜索主题 T 的网页的增广集, 这个集合中的网页抽象为节点, 它们之间的链接抽象为节点之间的连边. 每个节点 vi 都有权威值 ai 和枢纽值 hi, 所有节点的初始权威值设为 1. 假设某一个页面上有一个指向另一个页面的链接, 如果链接周围有较多关于搜索主题 T 的内容, 则认为链接的权值较大. 记 t 为链接前后 B 字节范围内关于主题 T 的内容出现的次数, 定义链接的权值 w =t+1, 在每 ij 步迭代之后进行归一化. 作者提出 ARC 算法时建议 B=50. 接下来, 通过下面的迭代过程使权威值和枢纽值达到稳定:

公式

与此类似, 文献 [] 也提出一种同时考虑页面之间的链接和页面内容的排序算法, 与 ARC 不同的是它对页面内容采用的是语义分析技术.

2.3.7　SALSA 算法

SALSA 算法 [], 即链接结构的随机分析法 (stoch- astic approach for link structure analysis), 是 HITs 算法的另一种改进. SALSA 算法不仅考虑了用户在浏览网页时顺着网页之间的链接方向访问网页, 还考虑了逆着链接方向访问原来的网页的情况. SALSA 算法用随机游走的方法, 通过访问网页的马尔科夫过程来确定网页的权威值和枢纽值的大小. 万维网用有向网络 G 表示, 所有入度不为零的节点构成权威集合 SA, 所有出度不为零的节点构成枢纽集合 SH, 两类节点之间的关系用无向边来表示: 图 G 中从节点 vi 指向 vj 的边表示为边 (iH, jA), 由此将原始网络 G 转换为无向二分网络 G , 图 2 给出一个示例. 用 G 中长度为 2 的路径模拟用户上网的随机游走过程, 则每一个随机游走的路

径都是从集合 SA 到集合 SH 再到集合 SA 或从集合 SH 到集合 SA 再到集合 SH, 其中每一个从集合 SH 到集合 SA 的路径都是沿着链接方向访问, 每一个从集合 SA 到集合 SH 的路径都表示逆着链接方向访问. 每一随机游走过后节点上的权值都会进行重新分配. 于是可以根据枢纽值和权威值定义两个随机游走过程. 对于计算枢纽值而言, 初始时刻赋予枢纽集合中的每个节点一单位初始权值, 用向量 h0 表示, 权值转换的过程可表示为, 其中权值转换矩阵 H 的元素 (H)ij h 其中 ai 为二分网络 G 的邻接矩阵元素, 如果节点 vi (SA) 与节点 v (SH) 相连接则 ai 1, 否则 ai0.kc 表示二分图中节点 vc 的度, 当 vcSA 时 kc 相当于节点 vc 在图 G 中的入度, 当 vcSH 时 kc 相当于节点 vc 在图 G 中的出度. 类似地, 权威值的转换过 程为 atAat1, 其中转移矩阵 A 的元素 (A) a 表示节点 v 将其权威值传给节点 v 的概率, 即: 多次迭代后每个节点上的值都达到稳定时停止迭代, 于是得到节点最终的权威值和枢纽值. 由于计算枢纽值和权威值的随机过程是相互独立的, 因此不会出现两者相互增强的情况, 相比 HITs 算法而言, SALSA 算法能够更好地避免主题漂移的问题. SALSA 算法实际上考虑的是一个基于二部分图的随机游走过程, 这一思路也被成功地应用在信息挖掘的另外两个领域中, 即基于网络结构的链路预测问题 [] 和个性化推荐算法 []. 实际上这里介绍的 SALSA 算法和推荐算法中的物质扩散算法如出一辙 [], 其区别在于以下几点: (1) 推荐系统中的物质扩散算法通常只考虑扩散两步的结果, 并不考虑稳态的结果; (2) 在个性化推荐中初始向量的设定根据目标用户不同而异, 而 SALSA 算法不会针对某一个节点设置不同的初始向量值; (3) 在推荐算法中通常只考虑用户没有选择过的产品的排序结果, 而 SALSA 考虑的是对所有节点的排序结果.

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 [] 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.

2.4.1 节点删除的最短距离法

破坏性反映重要性. 节点删除的最短距离法 [] 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”[]. 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 n1. 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 Nk (k=1, 2, ···, s) 表示一个节点 vi 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 s s N N , 记由于删除节点 v 造成 t1 rt1 t r i 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 vi 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:

公式

djk 为删除节点 vi 之前 vj 与 vk 间的最短距离. 注意, 当 j 或 k=i 的时候, 相当于直接损失; 当 jki 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.

2.4.2　节点删除的生成树法

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 [] 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 aii 替换为节点 vi 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 t0 为矩阵 L 的任意一个元素 lpq 的余子式 Mpq 的行列式, 即: t0 Mpq . 删除任意一个节点 vi, 网络的邻接矩阵变为 Ai, 然后用上面的方法计算网络的生成树个数为 ti. 由此可定义节点 vi 的中心性指标为

公式

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.

2.4.3　节点收缩法

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 []. 如果 vi 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

公式

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.

2.4.4 残余度的中心性

为了研究网络的抗毁性, Dangalchev[] 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 [] 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 [] 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 vi 之后, 定义其残余接近中心性为

公式

其中 djk(i) 为删除节点 vi 之后, 节点 vj 与 vk 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)1) 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.

2.5 权网络中的节点中心性

无权网络采用粗粒化的二分法来表示网络中节点间的联系 (有边为 1, 无边为 0), 不考虑联系的强弱信息. 然而边的权重信息能帮助我们更加细致地理解网络的结构与功能. 如在社交网络中, 边的权值可代表情感关系的强弱、交流与服务的频次、任务执行时间的长短等. 科学家合作网络中可用两个科学家合作论文的数量刻画两个科学家的联系紧密性. 航空运输网络中可以用两个机场之间所有班次上的座

位数表示这两个机场的通勤情况. 那么, 如何能够有效地利用网络的边权重信息进行重要节点的挖掘呢? 到目前为止, 大多数的研究思路都是将基于无权网络中心性指标在含权网络上进行扩展应用, 专门针对含权网络进行设计的方法鲜见.

2.6　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 []、女生用餐伙伴选择网络 []、空手道俱乐部网络 [] 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 vi 比 vj 更重要, 单独考察 vi 比 vj 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.

2.6.1　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 (i/n) 表示移除 i/n 比例的节点后, 网络中属于巨片 (giant component)[] 的节点数目的比例, 网络的鲁棒性 (robustness) 可用 R-指标刻画 []:

公式

显然, 不论对何种算法, 星形图中, R 取最小值 (1/n1/n2), 完全图中 R 取最大值 (11/n)/2, 当 n 比较大时 R0,1/2. 可定义 V12R 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 (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 [] 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 (图 4(a)) 和 6(图 4(b)), 移除节点时采用同时移除的方法.

2.6.2 用传播动力学模型评价排序算法

复杂网络上传播研究的对象极广 [], 比如通信网络中的病毒传播 []、社会网络中的信息传播 []、电力网络中的相继故障 []、经济网络中的危机扩散等 []. 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 [] 和 SIR 模型 []. 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 [], 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流传播地又快又广, 则说明该重要节点排序方法优于其他方法. 例如文献 [] 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 [] 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.

第三章　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 Girven 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 (1ink 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[]Í 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[] 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[] 朝将谱图理论引入模块性优化中。2008 年，Blondel 等人 u6】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。22 提出了著名的标签传播算法 (1abel 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 [] 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 [] 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 [] 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 130j 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 [] 提出了映射平衡算法 infomap．该方法基于最小描述长度 (MDI。) 原理Í川，通过信息传播扩散技术探测网络社区结构．2011 年，Morfirescu 等人 [] 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 [341 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 [] 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 Í71 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 [] 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)．RWAC() 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．

第四章　高速公路关键路段识别模型

对于交通运输 [1]、水利传输 [2]、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，甚至会达到负担不起预算的水平。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关键路段进行管理。即将资源投放到高速公路路段集合的一个关键子集，同时尽量增加高速公路的通行效率，以达到宏观层面增强路网稳定性的目的，实现事故前的预防，事故后的快速恢复。

这一章主要研究的是如何对高速公路关键路段挖掘问题进行建模，以及围绕着安徽、山西、北京的收费站车辆数据，进行建模之后的求解。

4.1　模型定义

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 表示边的集合，也就是高速公路中路段的集合。对于通过高速公路出行的车辆，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（$0<P_e<1$）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、布置资源而减小。定义管理者的决策向量 y=$\{y_1$，y_2，

$\cdots$，$y_n\}$，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，管理者会进行维护和投资管理，改善路段状况；0 表示非关键路段，暂时不关注。因为每一条路段都有一定的概率损毁，所以用 C_{e_i} 来表示第 i 个路段是否损毁，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 $\boldsymbol{c}=\{C_{e_1},C_{e_2},\cdots,C_{e_n}\}$，c 表示路网的某一种拓扑结构，$\boldsymbol{C}$ 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，所有路段在下一时刻都有一定的概率损毁，同时这个损毁不会影响到其他路段的损毁率。在传统研究网络可靠性的相关文献中 [3], 都是基于这个假设做的研究。

2）M>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，我们默认对于该车辆来说，他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一定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，作为本文的先验概率。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 $C^j = \{C_{e_1}^j, C_{e_2}^j, \cdots, C_{e_n}^j\}$ 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 $P^j = \prod_{i=1}^{n} (P_{e_i} C_{e_i}^j + (1 - P_{e_i})(1 - C_{e_i}^j))$ 表示高速公路网络拓扑变成这个拓扑结构的概率。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，进而路网拓扑结构概率矩阵 Z 也会发生变化。

在此提出关键路段挖掘模型：

$$L(\boldsymbol{y}) = -E(T(\boldsymbol{c}|\boldsymbol{y})) \tag{4.1}$$

其中，$T(\boldsymbol{c}|\boldsymbol{y})$：

$$T(\boldsymbol{y}) = P(K|\boldsymbol{c}) \sum_{k \in K} X_k \tag{4.2}$$

$\boldsymbol{y}$ 表示管理者想要投资维护的路段，$T(\boldsymbol{c}|\boldsymbol{y})$ 表示当关键路段集合为 $\boldsymbol{y}$，路网拓扑结构为 c 的时候，高速公路的整体通行时间。式4.1中对时间的期望取负，转化为通行效率。模型的研究目标是探究如何选取关键路段，对这些关键路段增加投入，使得在同样的预算情况下，整个路网的通行效率得到最大的提升。结合式4.1，式4.2，得到展开式：

$$Max(L(\boldsymbol{y})) = -Min_{\boldsymbol{y}} \sum_{\boldsymbol{c}\in C} P(\boldsymbol{c}|\boldsymbol{y})P(K|\boldsymbol{c}) \sum_{k\in K} X_k \tag{4.3}$$

式中 $\boldsymbol{y}$ 表示关键路段集合，假设高速公路网络的路段数量为 n，则 $\boldsymbol{y}$ 为 n 维向量，对于 $\boldsymbol{y}$ 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；$\boldsymbol{c}$ 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；$P(\boldsymbol{c}|\boldsymbol{y})$ 表示当关键路段集合为 $\boldsymbol{y}$ 时，高速路网的拓扑结构为 $\boldsymbol{c}$ 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；$P(K|\boldsymbol{c})$ 表示当路网拓扑结构为 $\boldsymbol{c}$ 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。

4.2 基于博弈的用户出行方式

当用户出行 $O-D$ 矩阵与路网拓扑结构 c 确定时，所有用户的总出行时间期望就已经确定了。

4.3 子模性分析

4.3.1 子模性定义

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 即满足子模性。

4.3.2　子模性证明

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

定义:

$$I = L(\boldsymbol{y} + \boldsymbol{\varepsilon}) - L(\boldsymbol{y}) - (L(\boldsymbol{Y} + \boldsymbol{\varepsilon}) - L(\boldsymbol{Y})) \tag{4.4}$$

不妨假设 $Y = y + \varepsilon_2$，公式 4.4 转化为: $I = L(\boldsymbol{y} + \boldsymbol{\varepsilon_1}) - L(\boldsymbol{y}) - (L(\boldsymbol{y} + \boldsymbol{\varepsilon_1} + \boldsymbol{\varepsilon_2}) - L(\boldsymbol{y} + \boldsymbol{\varepsilon_2}))$

令 $J = L(\boldsymbol{y} + \boldsymbol{\varepsilon_1}) - L(\boldsymbol{y})$，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: $\frac{dy}{dx} = \sum(\sum_{c_1|y+\varepsilon} P(c_1) - \sum_{c_2|y} P(c_2))X_k$。显然 $\sum_{c_1|y+\varepsilon} P(c_1) * X_k$ 具有单调非减性，导数恒大于等于 0。模型的子模性得到证明

由于对于具有子模性的模型，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。

4.4　贪心求解

贪心算法主要用于简化算法的复杂度，采用一步步获取局部最优解的方式，得到最终解集。在本小节中，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。不断迭代遍历直到关键路段的数量达到预算:

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，枚举 C_n^B 种情况，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)

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

function GREEDY($ODMatrix\ G = V, E\ B\ P_e$)
  $res \leftarrow 0$
  $Array \leftarrow []$
  $k \leftarrow 0$
  $l \leftarrow 0$
  while $len(Array) \leq B$ **do**
    for $i \in E - Array$ **do**
      if $L(Array + i) > k$ **then**
        $k = L(Array + i)$
        $l = i$
      end if
    end for
    $res \leftarrow k$
    $Array \leftarrow Array + l$
  end while
  return $Array$
end function

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

function ENUMERATION($ODMatrix\ G = V, E\ B\ P_e$)
  $res \leftarrow 0$
  $Array \leftarrow []$
  $k \leftarrow 0$
  for $l \in \Omega$ **and** $len(l) \leq B$ **do**
    if $L(l) > k$ **then**
      $k = L(l)$
      $Array = l$
    end if
  end for
  return $Array$
end function

是计算路段 i 的中心性函数，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。式中 f_i 表示路段 e 的流量。:

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: $Array \leftarrow []$
4: $k \leftarrow \{\}$
5: **for** $i \in E$ **do**
6: $k \leftarrow \{i, Z(i)\}$
7: **end for**
8: $SortbyValue(k)$
9: $Array \leftarrow k[0 : B]$
10: **return** $Array$
11: **end function**

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: $Array \leftarrow []$
4: $k \leftarrow \{\}$
5: **for** $i \in E$ **do**
6: $k \leftarrow \{i, f_i * P_i\}$
7: **end for**
8: $SortbyValue(k)$
9: $Array \leftarrow k[0 : B]$
10: **return** $Array$
11: **end function**

4.5 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。

4.5.1 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，其中主要的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。

还有高速公路路段损毁概率，这个通过统计历史的路段损毁次数获得，部分数据库已有，部分通过新闻抓取。路段的损毁包括交通事故损毁，重大自然灾害损毁，重大堵车事故等。

4.5.2 实验结果

图4.1，4.2给出了在不同时间区段下，几种方法的最终结果比较。图4.1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4.1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大小也在变化，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后效果的影响，流量越大，关键路段维护后造成的效益越大。图4.2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4.1一样，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，可以通过研究某个月的数据，获取关键路段，同时这些关键路段在宏观角度来看具有普适性。

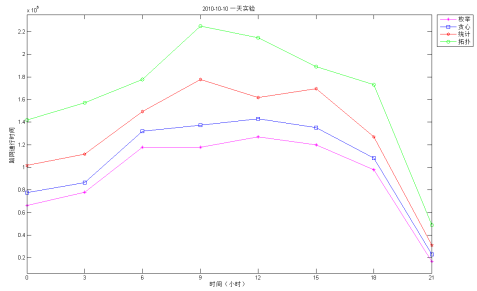


图 4.1　关键路段挖掘：以 1h 为区间

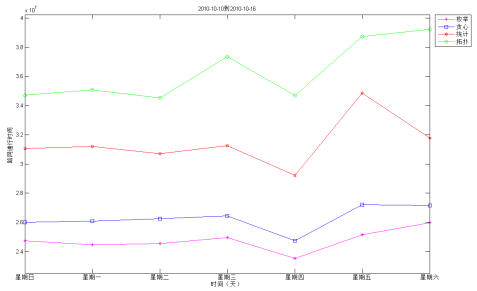


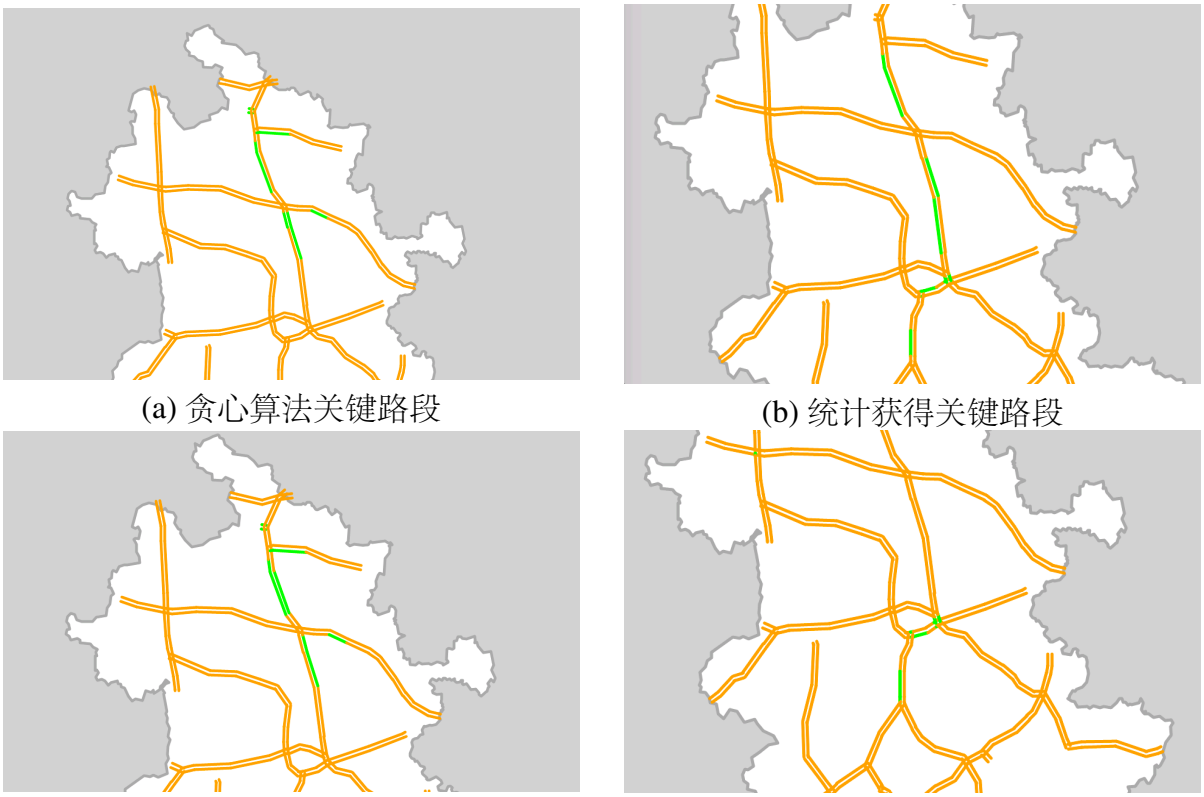
图 4.2　关键路段挖掘：以 1d 为区间

表4.1给出了不同方法求得的路网通行效率。表格中的数值是 $L(\boldsymbol{y})$ 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 4.1　算法结果集

图4.3给出了关键路段在路网中的分布图，图4.3(a) 是用贪心算法求得的关键路段集合，图4.3(b) 是高速公路统计方法获得的路段集合。图4.3(c) 是基于枚举所得的最优解集，图4.3(d) 是基于路网拓扑结构



(a) 贪心算法关键路段　(b) 统计获得关键路段

(c) 枚举获得最优解　(d) 基于拓扑结构的关键路段

图 4.3　不同方法求得的关键路段结果图

选取的关键路段集合。图4.3(a) 中颜色的变化和粗细的变化表示路段在贪心求解过程中，路段被选择的顺序；图4.3(b) 中颜色的变化表示路段的重要程度。对比两图可以发现，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节点，需要经过一些计算才能求出；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节点和枚举方法差距较大。

4.5.3　时间复杂度分析

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：$O(n * \log(n))$

其中，后两个方法可以用大根堆将时间复杂度优化到 O(n)。实验时间如表5.1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 4.2　不同方法的运行时间

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节点识别问题，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节点挖掘方法虽然在时间上运行极快，但是在精度上打不到要求。

4.6　本章小结

本章提出了一种面向高速公路网络的关键节点挖掘模型，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，本章分析了模型的子模性，单调性。特别的，本文的实验中使用枚举方法，获取高速公路中的最优解。结果表明该模型的贪心算法解集可以很好地逼近数据中真实解，并且在时间复杂度上有了规模性的优化，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起到不错的作用，但是这是基于目前研究的场景是静态关键路段挖掘模型，以及高速公路只有部分路段产生过阻断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)，当管理者需要更加快速得到实时反馈的时候，上述方法收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 O-D 特点，提出相应的解决手段。

第五章　高速公路社群划分方法

上一章介绍了面向高速公路的关键节点挖掘模型，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。虽然说高速公路上只有部分路段出现过损毁情况，n 的规模比较小，勉强可以求解，但是当高速公路网络扩大时，指数级别的复杂度不可接受。

5.1　模型分析

模型需要从输入的代表关键节点的离散 0-1 向量 $\boldsymbol{y}$，求得高速公路网络通行效率的期望。对于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，本质上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 $\boldsymbol{y}$ 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，现有的研究并没有一些比较合适的优化方法。

5.2 高速公路社群划分模型

5.2.1 模型定义

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 $c = \{v_1, v_2, ...v_m\}$，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 $C = \{c_1, c_2, ...c_u\}$；其中 $v_i \in V$，$c_i \bigcap c_j = \emptyset$，$\sum_{i=1}^{u} \sum_{v \in c_i} v = V$。

基于高速公路社群划分的关键节点挖掘算法主要采用分治思想，将一个难以直接解决的大问题，分割成一些规模较小的相同问题，以便各个击破，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节点，之后再用一定的方法合并。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系统中，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就是 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[] 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：

$$Q = \frac{1}{2m}\sum_{ij}[A_{ij} - \frac{k_i k_j}{2m}]\delta(c_i, c_j) \tag{5.1}$$

式5.1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 []，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 []。在此提出面向高速公路的社群划分模块化函数 Q:

$$\vartriangle Q = [\frac{\sum_{in} C + 2k_{i,in}}{2m} - (\frac{\sum_{tot} C + k_i}{2m})^2] - [\frac{\sum_{in} C}{2m} - (\frac{\sum_{tot} C}{2m})^2 - (\frac{k_i}{2m})^2] - L(i) \tag{5.2}$$

公式5.2用于在遍历过程中，判断节点应该属于哪一个社群。式中，$\sum_{in} C$ 表示社群 C 内部的所有边的权重和；$\sum_{tot} C$ 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。

$L(i)$:

$$L(i) = \frac{k_{i,c_1} - k_{i,c_2}}{k_{c_1,c_2}} \tag{5.3}$$

式5.3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。

在本节模型中，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。$L_{ij} = \sum_{e \in E_{ij}} e$，式中 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 $W_{ij} = \frac{f_{ij}}{L_{ij} * T}$。由于社群划分算法具有极端退化特性，

5.2.2 模型分析

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二维地理空间，且有明确的位置；高速公路网络中的边是一种实体联接，具有明确意义，并不是抽象空间中所定义的关系，能够明确表示线路之间的相互关系，线路在整个网络中的重要程度以及网络的局部和全局效率；高速公路交通网络中节点的长程联接需要一定成本，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；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度分布。

在以前的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 5.1，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。如图5.2，可以较直观的看出高速公路的社群特性。

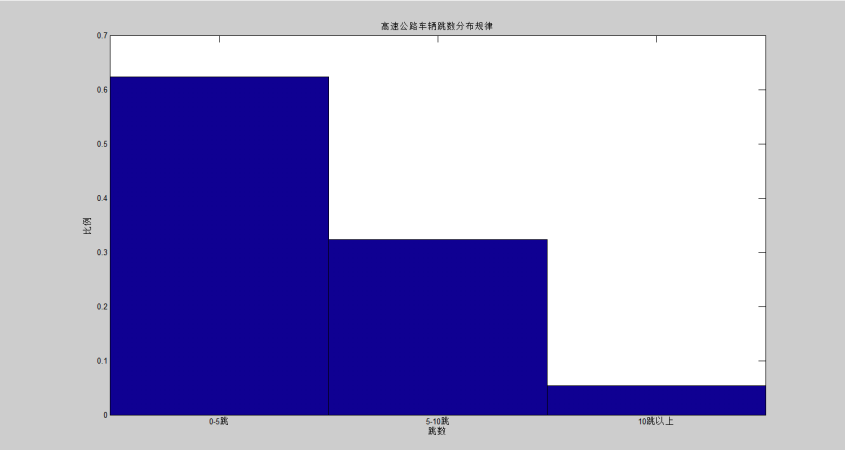


图 5.1　fig1

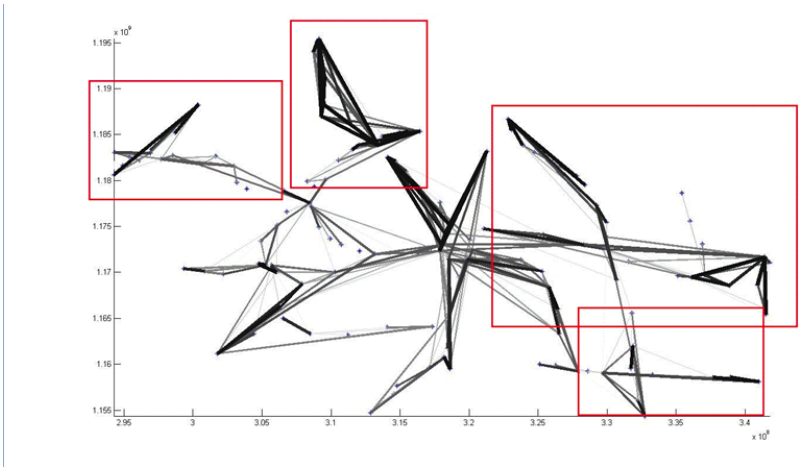


图 5.2　fig1

5.2.3　模型实现

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节点，分治计算，最后进行合并，达到优化时间复杂度的目的。在此引入基于模块性优化的社区挖掘方法。

2004 年，Newman 和 Girvan[] 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：

$$Q = \frac{1}{2m}\sum_{ij}[A_{ij} - \frac{k_i k_j}{2m}]\delta(c_i, c_j) \tag{5.4}$$

式5.1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；$m = \frac{1}{2}\sum_{ij} A_{ij}$。

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。[xxx] 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5.3），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。

在此，结合高速公路路网特性以及本文的研究目标，作出以下几项改进：

1）在边权中引入路段物理长度：高速公路网络是具有实体的物理网络。为了找到符合本文要求的社区，我们认为只要两个收费站有 O-D 交流，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距离就是 1 跳或者其他权值，但是对于物理网络，它具有实体距离。在此定义两个收费站 i，j 之间的距离，令 $e = e_1, e_2 ... e_n$ 表示 i 和 j 之间的最短路径中的路段集合，本文认为 i 和 j 的距离 $D = \sum len(e_k)$。这个方法可以解决一部分孤立点和交叉社区

2）采用模拟退火思想，不断变化边权——增加距离的权重，直

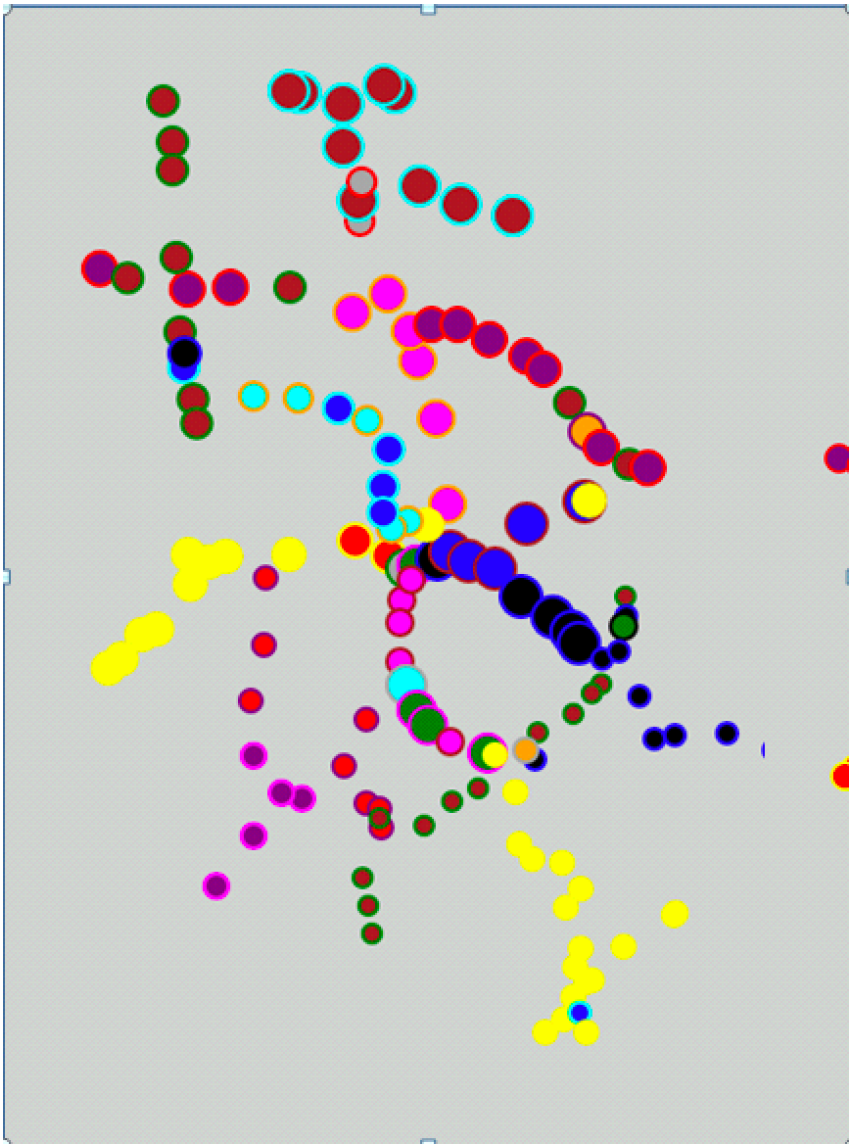


图 5.3　fig1

到解集收敛，或者新的解集的规模变化大于退火温度：这个方法主要用于使解集收敛，解决极端退化特性。

社群划分基准函数：

$$\Delta Q = [\frac{\sum_{in} C + 2k_{i,in}}{2m} - (\frac{\sum_{tot} C + k_i}{2m})^2] - [\frac{\sum_{in} C}{2m} - (\frac{\sum_{tot} C}{2m})^2 - (\frac{k_i}{2m})^2] - L(i) \tag{5.5}$$

公式5.2用于在遍历过程中，判断节点应该属于哪一个社群。式中，$\sum_{in} C$ 表示社群 C 内部的所有边的权重和；$\sum_{tot} C$ 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。

$L(i)$:

$$L(i) = \frac{k_{i,c_1} - k_{i,c_2}}{k_{c_1,c_2}} \tag{5.6}$$

式5.3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。下面说明算法步骤。

初始权重设为 f/l，f 表示两个节点之间的流量，$l = \sum len(e)$ 表示两个节点之间的物理距离。e 表示两个节点之间的最短路径。初始情况下，每个节点都构成一个社群，对节点进行遍历，运用公式5.2，判断节点应该属于哪一个社群。循环到节点社群无变化或者社群分类进入循环的状态，停止本次迭代，改变权值，进行下一轮迭代，直到收敛。

伪代码如下:

采取自下而上的策略，可以有效解决低分辨率问题；而基于高速公路的物理网络特性，对边权值进行处理，可以增加算法结果的收敛程度；最后用模拟退火思想，逐渐加强距离的权重，达到消除孤立点，建立物理层面社群的目的。

5.3 基于社群划分的复杂网络关键节点挖掘

上一章节介绍了如何对高速公路进行社群划分，本节介绍如何结合高速公路社群划分算法进行关键节点挖掘。

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

function COMMUNITY($ODMatrix\ G = V, E\ B$)
  $res \leftarrow [[\{0\ 0\}\ \{1,1\}\ \cdots\ \{n,n\}]]$
  $tmp \leftarrow [\{\}]$
  $pre \leftarrow [[\{0\ 0\}\ \{1,1\}\ \cdots\ \{n,n\}]]$
  $k \leftarrow 0$
  $l \leftarrow 0$
  $T \leftarrow 100$
  while $|len(res) - len(pre)| \leqslant T$ **do**
    $res = res[-1]$
    $pre = res$
    while $res[-1] \notin res[0:-1]$ **do**
      $tmp \leftarrow res[-1]$
      for $i \in E$ **do**
        for (**do**$C \in tmp \& |C| \leq B$)
          if $\triangle\ Q > k$ **then**
            $l \leftarrow C$
            $k \leftarrow \triangle\ Q$
          end if
        end for
        $tmp[l] \leftarrow i$
      end for
      $res\ add\ tmp$
    end while
    T–
  end while
  return res
end function

5.3.1　合并策略

假设社群已划分完毕，针对每一个高速公路社群，利用贪心算法获取关键节点集合，最后对所有子社群的结果进行合并，得到最终结果。贪心算法和高速公路社群划分的思路见前两节。由于贪心算法可以计算出每一个关键节点选出后，高速公路运行效率的增量，所以问题转化为投资问题：假设一共有 k 个社群，将这 k 个社群看作 k 种货物。总共预算为 B，当第 x 种货物投入 i 预算时，收益为 $f(x_i)$。投资问题可以利用动态规划，伪多项式时间内求解。

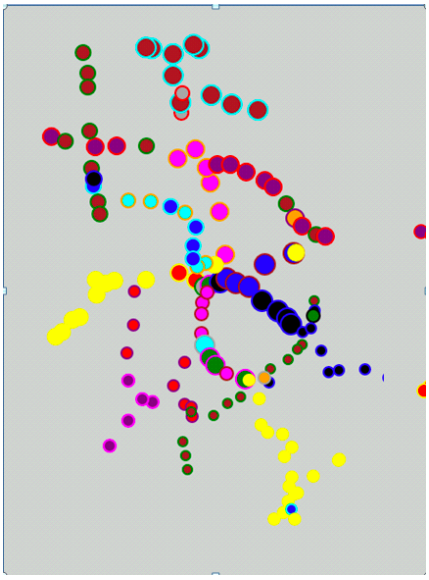


图 5.4　fig1

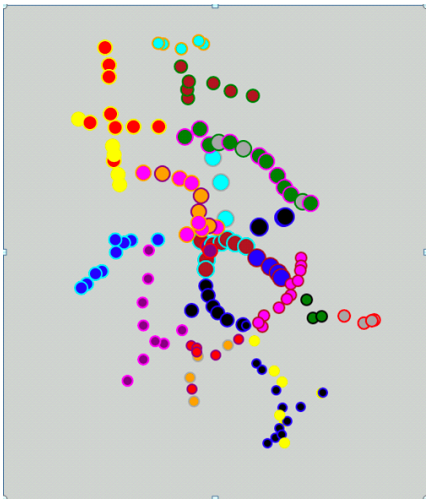


图 5.5　fig2

5.3.2　投资问题

5.4　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。

图5.4是基于带权的模块性函数的简单分群结果，该方法根据基本模块化函数 Q 进行分群，将路段之间的流量直接作为权值。实验中，最终的分群结果收敛于几百个解构成的解集合，图5.4是从中随机挑选的一个。可以看出图中有很多孤立点，而且很多社群相互交叉，不符合高速公路的物理网络特性，不符合高速公路社群划分的意义。图5.5是基于目标函数5.1的社群划分结果，该试验最终收敛于两个解构成的解集合，我们发现该方法最终可以消除孤立点，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。图5.6直接将具有交叉节点的社群合并的结果。

分群后，采用分治方法计算关键路段集合。最后将求的的关键路段带入到第一章的模型中，计算路网通行效率的提升率。图5.7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5.8给

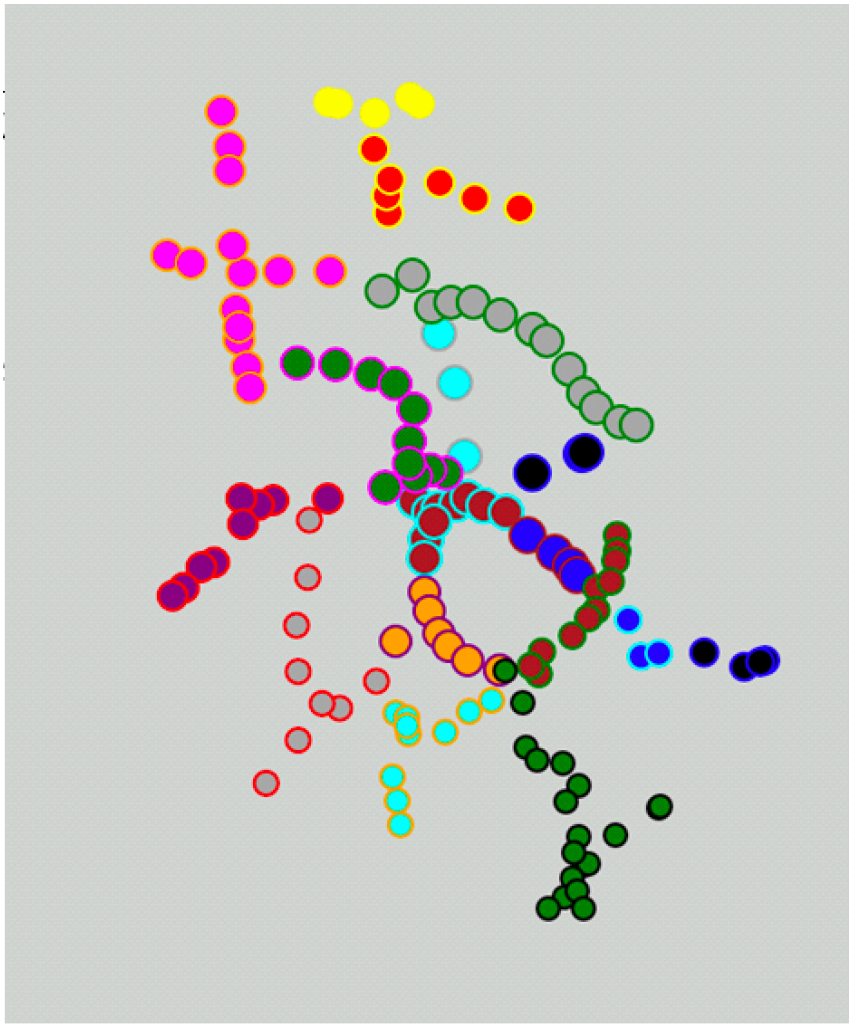


图 5.6　fig1

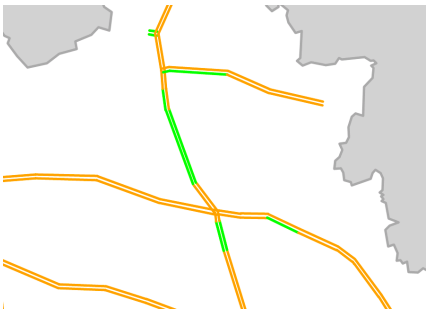


图 5.7　图片还得再画

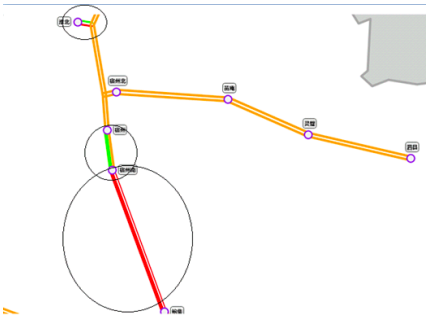


图 5.8　图片还得再画

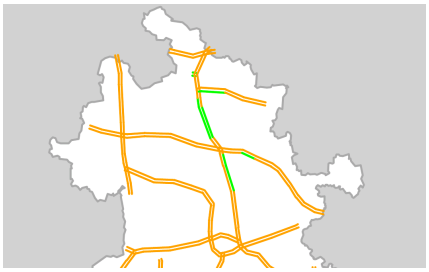


图 5.9　fig1

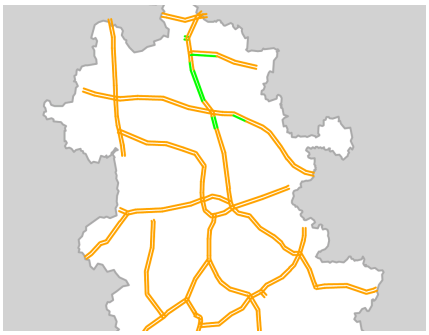


图 5.10　fig2

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。

下图给出不同方法选出的关键路段集合，图5.9给出了枚举方法选出的关键路段集合，图5.10给出了简单贪心算法给出的关键路段集合，图5.11给出了结合社群划分的关键节点识别算法的结果，图5.12给出了基于统计学的关键节点集合。观察图5.9和图5.10，发现两者选取的关键节点具有很强的相似性。

误差分析：

表5.1描述了枚举方法和直接贪心方法之间的误差，直接贪心和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。

关键节点选取误差分析：

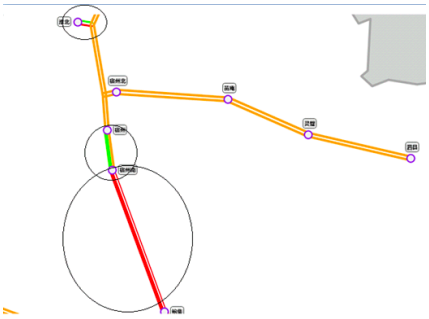


图 5.11　fig1

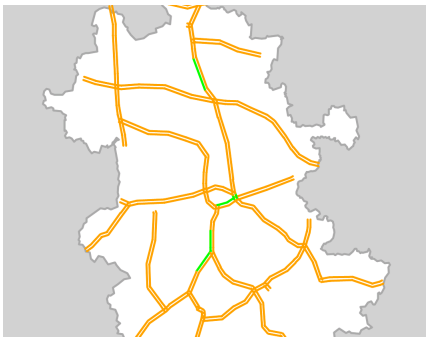


图 5.12　fig2

	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 5.1　example of table

表5.2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节点准确率逐步上升。

运行效率分析：

由表5.3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5.1来看，精度误差处于可接受范围$(1/e)$。

5.5　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。

	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 5.2　example of table

	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 5.3　example of table

结论

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。

参考文献

[1] 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.

[2] Fortunato and Santo. "*Community detection in graphs*". *Physics Reports*, **2010**, *486*(3–5): 75–174.

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