

PaperPass专业版检测报告

简明打印版

比对结果(相似度)：

总 体：15% (总体相似度是指本地库、互联网的综合比对结果)

本地库：15% (本地库相似度是指论文与学术期刊、学位论文、会议论文数据库的比对结果)

期刊库：8% (期刊库相似度是指论文与学术期刊库的比对结果)

学位库：13% (学位库相似度是指论文与学位论文库的比对结果)

会议库： 1% (会议库相似度是指论文与会议论文库的比对结果)

互联网：5% (互联网相似度是指论文与互联网资源的比对结果)

编 号：58BEFB0B61C3EDWZR

版 本：专业版

标 题：链路预测及交通拥堵恢复序列的预测

作 者：李星

长 度：55031 字符(不计空格)

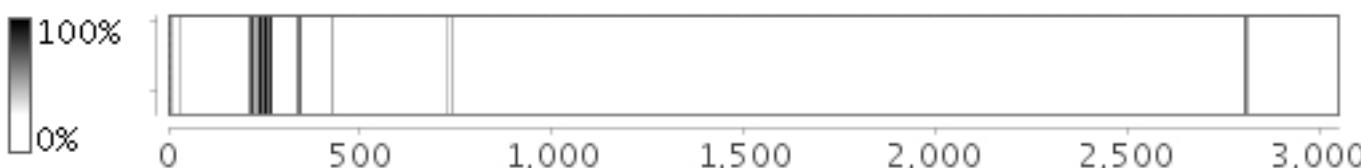
句 子 数：3047 句

时 间：2017-3-8 2:25:15

比对库：学术期刊、学位论文（硕博库）、会议论文、互联网资源

查真伪：<http://www.paperpass.com/check>

句子相似度分布图：



本地库相似资源列表（学术期刊、学位论文、会议论文）：

1. 相似度：4% 篇名:《复杂网络链路预测》

来源：学术期刊 《电子科技大学学报》 2010年5期 作者:吕琳媛

2. 相似度：4% 篇名:《复杂网络的链路预测:基于结构相似性的算法研究》

来源：学位论文 湘潭大学 2011 作者:白萌

3. 相似度：3% 篇名:《复杂交通网络拥堵特性及控制方法研究》

来源：学位论文 西南交通大学 2009 作者:赵月

4. 相似度：2% 篇名:《移动通话网络中链路预测问题的研究》

来源：学位论文 西安理工大学 2014 作者:潘剑

5. 相似度：2% 篇名:《复杂网络上链路预测的研究》

来源：学位论文 华北电力大学 2012 作者:邢登华

6. 相似度 : 2% 篇名:《二分网络链路预测方法研究》
来源: 学位论文 合肥工业大学 2013 作者: 马吴迪
7. 相似度 : 1% 篇名:《具有时间感知的加权网络链路预测研究》
来源: 学位论文 中南大学 2014 作者: 涂一娜
8. 相似度 : 1% 篇名:《静态功能脑网络数学建模方法的研究》
来源: 学位论文 太原理工大学 2015 作者: 牛力敏
9. 相似度 : 1% 篇名:《复杂网络中链路预测问题的研究与实证》
来源: 学位论文 西安理工大学 2011 作者: 商超
10. 相似度 : 1% 篇名:《基于标签传播的链路预测算法研究与应用》
来源: 学位论文 北京交通大学 2014 作者: 辛霆麟
11. 相似度 : 1% 篇名:《复杂网络观察》
来源: 学术期刊 《复杂系统与复杂性科学》 2010年2期 作者: 吕琳媛 陆君安 张子柯 闫小勇 吴晔 史定华 周海平 方锦清 周涛
12. 相似度 : 1% 篇名:《基于复杂网络的链路预测研究》
来源: 学位论文 河北大学 2014 作者: 赵延乐
13. 相似度 : 1% 篇名:《动态多维社会网络中链路预测方法研究》
来源: 学位论文 山东师范大学 2012 作者: 滕兆明
14. 相似度 : 1% 篇名:《电子邮件通信实体关系挖掘与分析研究》
来源: 学位论文 电子科技大学 2014 作者: 吴祖峰
15. 相似度 : 1% 篇名:《基于博弈的社会网络个性化好友推荐算法研究》
来源: 学术期刊 《计算机科学》 2015年9期 作者: 杨阿桃 汤庸 王江斌 李建国
16. 相似度 : 1% 篇名:《基于层次随机图模型的复杂脑网络链路预测研究》
来源: 学位论文 太原理工大学 2015 作者: 田甜
17. 相似度 : 1% 篇名:《多维社会网络中链路预测方法研究》
来源: 学术期刊 《计算机应用与软件》 2011年9期 作者: 滕兆明 王红 张华青 马晓慧
18. 相似度 : 1% 篇名:《复杂网络中链路的可预测性》
来源: 学术期刊 《复杂系统与复杂性科学》 2014年1期 作者: 许小可 许爽 朱郁筱 张千明
19. 相似度 : 1% 篇名:《基于相似性的链接预测方法研究》
来源: 学位论文 哈尔滨工程大学 2012 作者: 李淑玲
20. 相似度 : 1% 篇名:《基于社团特性的链路预测算法的研究》
来源: 学术期刊 《广东技术师范学院学报(社会科学版)》 2015年2期 作者: 雷方元 蔡君
21. 相似度 : 1% 篇名:《社交网络链路预测研究综述》
来源: 学术期刊 《信息与电脑》 2015年21期 作者: 杨星 鲁天琦 陈美灵 刘超 张树华
22. 相似度 : 1% 篇名:《科学知识网络中的链路预测研究述评》
来源: 学术期刊 《中国图书馆学报》 2015年3期 作者: 张斌 马费成
23. 相似度 : 1% 篇名:《有向与加权网络的链路预测》
来源: 学位论文 湘潭大学 2011 作者: 张扬夫
24. 相似度 : 1% 篇名:《基于信任和相似标签的链接预测算法》
来源: 学位论文 云南财经大学 2014 作者: 范思理
25. 相似度 : 1% 篇名:《一种基于节点相似性的链接预测算法》
来源: 学术期刊 《中国科技论文》 2013年7期 作者: 张健沛 姜延良
26. 相似度 : 1% 篇名:《基于预测误差修正的时序链路预测方法》
来源: 学术期刊 《电子与信息学报》 2014年2期 作者: 邓志宏 老松杨 白亮
27. 相似度 : 1% 篇名:《在线社交网络中的链路预测问题研究》

来源：学位论文 中国科学技术大学 2014 作者：靳婷

28. 相似度：1% 篇名：《基于链路预测的我国航线网络演化机制研究》

来源：学位论文 南京航空航天大学 2012 作者：程凯

29. 相似度：1% 篇名：《基于拓扑相似性的动态网络链路预测方法》

来源：学位论文 西安电子科技大学 2013 作者：邓琨

30. 相似度：1% 篇名：《MapReduce环境下的并行复杂网络链路预测》

来源：学术期刊 《软件学报》 2012年12期 作者：饶君 吴斌 东昱晓

31. 相似度：1% 篇名：《复杂网络中连边预测方法研究》

来源：学位论文 武汉理工大学 2012 作者：郑汉彬

32. 相似度：1% 篇名：《复杂网络中集聚系数对链路预测算法的影响》

来源：学术期刊 《科技视界》 2014年12期 作者：黄子轩 马超 徐瑾辉 黄江楠

互联网相似资源列表：

1. 相似度：4% 标题：《复杂网络的路预测：基于结构相似性的算法研究.pdf文档全文免费阅读、...》

<http://max.book118.com/html/2015/1026/27946833.shtml>

2. 相似度：3% 标题：《复杂交通网络拥堵特性及控制方法研究_图文_百度文库》

<http://wenku.baidu.com/view/d3b642d076eeaeaad1f3304b.html>

3. 相似度：2% 标题：《复杂交通网络拥堵特性及其控制方法的研究.pdf文档全文免费阅读、在线看》

<http://max.book118.com/html/2015/1001/26509748.shtml>

4. 相似度：1% 标题：《复杂网络链路预测的理论、算法和应用研究 - MBA智库文档》

<http://doc.mbalib.com/view/2417f46c115269ffaf22600ee3c09fb2.html>

全文简明报告：

摘要

在自然世界以及人类的生活当中，我们所能接触的许多系统都可以被抽象为复杂网络。 { 49%：这样的复杂网络包括因特网、世界贸易网络、社交网络、城市交通网络、蛋白质相互作用网络等。 } { 57%：实际的复杂网络当中往往存在着信息缺失的情形，复杂网络中的链路预测是指如何通过已知的 } { 95%：网络节点以及网络结构等信息预测网络中尚未产生连边的两个节点之间产生连接的可能性。 } { 46%：链路预测是对于网络分析和数据挖掘的一项关键技术。 } { 55%：近年来，随着复杂网络科学的快速发展，链路预测也受到了越来越多的关注。 } 链路预测算法具有广泛地实用与研究价值，在生物信息、电子商务、交通运输等多个领域都取得了不少成果。 { 40%：链路预测当中基于局部相似性的链路预测指标只需要利用节点的局部信息（如节点的度和其邻居节点）即可得到节点间的相似性指标。 } { 45%：基于局部相似性的链路预测算法由于其较小的时间复杂度和较高的预测准确率，基于局部相似性的链路预测算法在应用上得到了大家的青睐。 } { 41%：本文想在保证局部相似性算法低复杂度的前提下，提升链路预测算法的预测准确率。 } 交通网络属于复杂网络的一种，交通网络中的拥堵现象在一定程度上可以理解为网络中边的缺失， 基于链路预测的启发，我们将利用链路预测的思路来解决交通拥堵问题。 当交通运输网络当中出现多条线路的拥堵情形时，我们将利用链路预测的思路对拥堵边进行重要性的评价。 我们提出了基于链路预测思路的反相关性算法，根据其对交通拥堵恢复序列的预测，我们将依次对拥堵边进行恢复。

本文的主要工作及成果包括：

1) 分析了链路预测中9种经典的基于局部信息的相似性指标，根据链路预测指标，我们会对网络当中的节点对进行评分，进而确定这些边的重要性。 {71%：这9种指标分别是共同邻居指标、Salton指标、Jaccard指标、Sorenson指标、大度节点有利指标、大度节点不利指标、} {47%：Adamic-Adar指标、资源分配指标和LHN指标（这是9种单独的评价指标）。} 由于基于局部相似性的链路预测指标只需要根据节点的度、最近邻居等节点的局部信息来对边的重要性进行评分， {47%：所以这类指标的计算复杂度较低并适用于较大的网络。} 但由于其所利用到的信息有限，其预测准确度也并非很高。

{47%：2) 以上9种指标，如CN算法将所有的共同邻居等同看待，仅利用共同邻居的数量作为节点对间相似度的评分函数，} {82%：没有区分出不同的邻居节点对链接预测的影响；} {97%：AA算法和RA算法虽然区分了每个不同的共同邻居对链接预测的不同的影响力，但是它们都只关注于共同邻居本身，而忽略了这些共同邻居之间的相互影响。} {48%：由于聚类系数表示复杂网络中节点间联系的紧密程度，高的聚类系数表示邻居节点之间有着更高的连接可能性。} 我们将聚类系数这一复杂网络特征添加到现有的9种经典局部相似性算法之中，用以区分共同邻居节点对于链路形成的不同贡献程度。 在5种真实数据集上的实验结果表明了基于聚类系数对现有9种经典链路预测指标优化的可行性。 此外，链路预测的准确性研究多基于丢失边的预测，对于错误连边的准确性预测少有涉及，本文的实验部分同时涉及了丢失边的预测以及错误连边的预测。 传统的链路预测比较实验中，对于测试集多选取10%这一固定值，为了说明算法的健壮性，我们针对不同大小的测试集进行了比较实验。

3) 关于交通拥堵方面的研究，先前多基于提升网络的运载能力。 然而，再优良的网络设计也不能完全杜绝交通堵塞问题。 基于本文链路预测研究的启发，我们提出了基于链路预测的交通拥堵恢复序列预测算法，我们将交通网络中的拥堵边看作链路预测中的缺失边， 根据链路预测算法对其重要性进行评分，根据网络中拥堵边的重要性对其进行依次恢复，从而在最短的时间内恢复网络的运载能力。 本文提出的基于链路预测的反相关性恢复算法分别是反优先连接指标、反相加优先连接指标、反三阶相似性指标、反局部路径指标， {41%：这些算法都是基于节点的反相关性设计的，该类指标认为度较小的节点更有利保证网络的连通性，} {54%：这在之前的链路预测研究当中少有提及。} 我们在一种模拟网络（BA网络）及两种真实交通网络之上进行了大量的实验， {44%：实验比较了链路预测算法、基于链路预测的反相关性算法和随机恢复方法的恢复效果，} 实验表明基于链路预测的反相关性算法在交通拥堵恢复序列预测方面优异的表现。

关键词： 链路预测算法； 聚类系数； 丢失边预测； 错误边预测； 交通拥堵恢复序列的预图表目录5

第1章 绪论6

1.1 研究背景与意义6

1.1.1 链路预测6

1.1.2 交通拥堵问题8

1.2 论文研究问题9

1.2.1 链路预测算法9

1.2.2 交通拥堵恢复序列预测10

1.3 国内外研究现状11

1.3.1 链路预测11

1.3.2 交通拥堵及预测13

1.4 本文主要研究工作及贡献14

1.4.1 基于聚类系数的链路预测算法研究15

1.4.2 利用链路预测思维进行交通拥堵恢复序列的研究。 15

1.5 本文的章节安排16

第2章链路预测方法分析18

2.1 链路预测问题描述18

2.1.1 丢失边的预测18

2.1.2 错误边的预测18

2.2 链路预测算法流程19

2.3 数据集划分21

2.3.1 随机抽样21

2.3.2 逐项遍历21

2.3.4 k折叠交叉检验21

2.3.5 滚雪球抽样21

2.3.6 其他方法21

2.4 经典的链路预测算法22

2.4.1 共同邻居指标22

2.4.2 Salton 指标23

2.4.3 Jaccard 指标23

2.4.4 Sorensen 指标24

2.4.5 大度节点有利指标24

2.4.6 大度节点不利指标24

2.4.7 LHN 指标24

2.4.8 Adamic-Adar指标25

2.4.9 资源分配指标25

2.5 评价指标26

2.5.1 接受者操作特征曲线下面积26

2.5.2 准确度27

2.5.3 排序分27

2.6 本章小结27

第3章 基于聚类系数的链路预测算法29

3.1 引言29

3.2 链路预测算法伪代码29

3.3 聚类系数的定义31

3.4 基于聚类系数的链路预测算法改进32

3.4.1 基于聚类系数的共同邻居指标32

3.4.2 基于聚类系数的Salton 指标32

3.4.3 基于聚类系数的Jaccard 指标32

3.4.4 基于聚类系数的指标33

3.4.5 基于聚类系数的HPI 指标33

3.4.6 基于聚类系数的HDI 指标33

3.4.7 基于聚类系数的LHN 指标33

3.4.8 基于聚类系数的AA 指标33

3.4.9 基于聚类系数的资源分配指标34

3.5 数据集介绍34

3.5.1 数据集描述35

3.5.2 数据集的网络特征的统计35

3.6 实验结果分析36

3.6.1 实验环境36

3.6.2 参数设定36

3.6.3 实验结果以及分析36

3.7 本章小结42

第4章交通拥堵恢复序列预测问题分析43

4.1 问题描述43

4.2 解决思路43

4.3 流量模型44

4.3.1 吞吐量定义44

4.3.2 衡量指标44

4.4 实验流程46

4.5 评价方式48

4.6 实验伪代码49

4.7 本章小结50

第5章交通拥堵恢复序列算法51

5.1 算法描述51

5.1.1 偏好连接算法51

5.1.2 本地路径算法51

5.1.3 三阶相似性算法51

5.1.4 反偏好连接算法52

5.1.5 反相加偏好连接算法52

5.1.6 反本地路径算法52

5.1.7 三阶相似性算法52

5.2 数据集描述52

5.2.1 数据集描述52

5.2.2 网络特征统计55

5.3 实验结果分析56

5.3.1 加州路网上的结果分析56

5.3.2 不同网络损毁比例对于实验的影响58

5.3.3 不同网络尺寸对于实验的影响59

5.3.4 不同大小的流量对于实验的影响60

5.4 本章小结60

第6章总结与展望62

6.1 总结62

6.2 进一步工作62

致谢64

参考文献66

图表目录

图 1.1 蛋白质相互作用网络

图 1.2 恐怖分子犯罪网络

图1.3 北京上班早高峰

图1.4 链路预测图示

图1.5 交通拥堵恢复序列

图 2.1 复杂网络丢失边预测

图2.2 复杂网络错误边预测

{ 61 % : 图 2.3 链路预测算法流程图 }

图 2.4 CN算法图示

图2.5 计算RA指标

图3.1(a) 链路预测算法伪代码

图3.1(b) 链路预测算法伪代码

图3.2 聚类系数计算示例

图3.3 数据原始格式

图 3.4 (a) Elegans网络缺失边发现

图 3.4 (b) Elegans网络错误边发现

图 3.5 (a) Football网络缺失边发现

图 3.5 (b) Football网络错误边发现

图 3.6 (a) Jazz网络缺失边发现

图 3.6 (b) Jazz网络错误边发现

图 3.7 (a) USAir网络缺失边发现

图 3.7 (b) USAir网络错误边发现

图 3.8 (a) Karate 网络缺失边发现

图 3.8 (b) Karate 网络错误边发现

图 4.1 交通拥堵示例

图 4.2 值变化 ($R=100$, $S=1.8$, $N=20$)

图 4.3 实验整体思路

{ 54 % : 图 4.4 每个时间片内对数据包的处理流程 }

图 4.5 实验程序流程图

图 4.6 不同的交通恢复序列算法下的流量模拟伪代码

图 5.1 BA 网络样例

图 5.2 加州路网及美国航空网络原始数据

图 5.3 加州路网上的实验结果

图 5.4 BA 网络上不同拥堵率对实验结果的影响

图 5.5 不同尺寸的 BA 网络对实验结果的影响

图 5.6 美国航空网络上不同大小的流量对于实验结果的影响

表 1.1 链路预测算法比较

表 3.1 数据集的网络统计特征

表 5.1 网络特征统计

第 1 章 绪论

1.1 研究背景与意义

{ 71 % : 在现实世界中，从道路交通到 Internet，从神经系统到生态系统，从蚁群到人类社会，无不属于复杂系统的范畴。 } { 96 % : 在这些系统内，往往相互作用的个体众多且作用种类复杂。 } 如何通过现有复杂网络信息来挖掘暂未被发现的信息以及预测复杂网络未来一段时间内可能出现的新型网络结构已经成为了一个极其重要的研究点。 { 42 % : 链路预测便是对于复杂网络当中缺失信息的还原与预测的一项关键性技术。 } { 41 % : 交通运输网络做为复杂网络的一种，当交通网络当中的多条线路发生拥堵之时， } 我们可将多条线路的恢复顺序问题抽象

为链路预测问题，依据链路预测算法对不同的拥堵边进行评分，并按照其重要程度依次对其进行恢复。

链路预测

关于网络结构的研究已有数百年的历史，网络学科的理论发展经过了规则网络、随机网络和复杂网络三个阶段。 { 42 % : 网络上的首项研究可以追溯到1736年欧拉关于哥尼斯堡七桥问题的提出与解决，这属于规则网络时代的研究。 } { 43 % : 之后的很长一段时间，网络学科都没有很大的发展。 } { 44 % : 1959年，匈牙利的著名数学家 Erds 和 Rnyi 建立了著名的随机图(ER)理论，这才将网络研究带入到了第二个阶段[1]。 } { 52 % : 复杂网络的开启主要归功于 Watts 和 Strogatz 在1998出的小世界模型以及 Barabsi 在1999提出的无标度网络[2]。 } { 48 % : 近年来，随着复杂网络学科的快速发展，作为其分支之一的链路预测研究也得到了越来越广泛的关注。 } { 78 % : 网络中的链路预测是指根据网络现有的节点属性以及网络拓扑等信息来预测网络中尚未产生连边的两个节点之间产生链接的可能性[3]。 } 链路预测包含了对于未知边的发掘以及网络演化中即将产生的边的预测。 链路预测很好的将网络与信息科学进行了连接，可以很好地处理信息科学中缺失信息的还原与预测。

{ 100 % : 复杂网络作为复杂系统的一种拓扑近似，在构建过程中，由于时间和空间或者实验条件的限制，有不少链接并未探测到，而导致许多链接是未知的。 } { 100 % : 并且复杂系统之所以复杂，它是随时间动态演化的，与之对应的网络结构就会有节点和边的添加或去除。 } { 100 % : 这样在采样过程中难免有错误或冗余的链接出现。 }

图 1.1 蛋白质相互作用网络

{ 91 % : 此探讨根据已知的网络信息对缺失的边以及未来的边进行预测，不仅能进一步加深对网络结构和演化机制的理解，还能对相关实验工作提供一定的指导。 }

{ 100 % : 以(Yeast)蛋白质相互作用网络为例，人们对其所知有限，约有80%的相互作用是未知的。 } { 100 % : 为了探测可能存在的相互作用，用室内实验的方法一一检测可能存在的链接，尤其当网络节点众多而又稀疏时，通常花费巨大，耗时耗力。 } { 95 % : 与其盲目的检测所有可能存在的链接，不如根据已知的网络结构及其它信息来预测可能存在的链接。 } { 91 % : 如果能找到合适的链路预测方法且预测的精度足够高的话，那么我们便可以有选择有目标的进行实验，在很大程度上减少了实验所需的花费和人力。 } { 65 % : 图1.1即为一种蛋白质相互作用网络。 }

在其他领域上，链路预测可以识别犯罪网络的结构，如图1.2所示[4]，使利用协同过滤的推荐系统克服数据稀疏的问题，检测并控制通过email传播的电脑病毒等等。 { 43 % : 链路预测算法在实际生产生活中的可应用性是它受到广泛关注的重要原因。 } { 100 % : 链路预测的方法也可直接用于信息的推荐。 } { 100 % : 比如 Facebook 等在线社交网站的成功证明，基于共同的爱好推荐尚不存在但可能性很大的链路作为潜在的朋友关系，能够帮助用户找到新的朋友，进而提高对网站的信任度。 }

图 1.2 恐怖分子犯罪网络

另外，它还可以作为评价网络演化机制的一种方式。 { 86 % : 为了解释网络演化过程中节点连接成边的内在机制，各种各样的驱动因素被提了出来，如随机重连(random connection) [5]、优先连接(preference attachment)[6]、优先去除(preference depletion)[7]、稳定性限制(stability constraints) [8]、基于欧几里德距离的节点相似度(node-similarity based on Euclidean distance)[9]等等。 } { 100 % : 它们从不同的角度唯象的解释了复杂网络的结构及演化特征，缺乏统一的标准判别孰优孰劣。 } { 100 % : 如果把每一种模型当作链路预测的一种算法，就可

以利用链路预测中的精度判别标准，基于一些已知的演化数据，对不同的演化模型做一个统一的评价。}

{ 84 % : 其实作为数据挖掘的一个子问题，链路预测在信息科学领域早已展开了研究，尽管如此，这个问题仍然得到了越来越多的关注，开始在复杂网络领域内蓬勃发展。} 原因在于：

{ 94 % : (1) 借助于各种数学和物理思想方法及先进的计算机手段，人们研究的网络体系越来越大，结构越来越复杂，} { 100 % : 原来的链路预测方法远不能满足人们对预测精度和效率的需要，因此有必要发展新的链路预测算法。}

{ 95 % : (2) 长期以来，网络的结构与功能是复杂网络研究的基本问题之一，网络的结构决定了网络的功能，而网络的功能又影响着网络的结构。} { 100 % : 以基因网络和电子电路网络为例，每一种子结构代表网络的某一部分功能。} { 100 % : 链路预测既然是对网络结构的一种预测，} { 100 % : 更深入地研究它与链路预测的内在关系，} { 71 % : 比如细致地分析网络的一些普适子结构(比如模体[10]、社团等)及研究某些网络的特殊结构(比如等级结构[11])，} { 100 % : 不仅对发展新的算法有借鉴意义，还能丰富人们对网络结构和功能的认识。}

1.1.2 交通拥堵问题

{ 100 % : 复杂网络是对复杂系统非常一般的抽象和描述方式，它突出强调了系统结构的拓扑特征。} { 100 % : 原则上说，任何包含大量组成单元(或子系统)的复杂系统，当我们把构成单元抽象成节点，单元之间的相互作用抽象为边时，都可以当作复杂网络来研究。} { 88 % : 交通系统是一个典型的复杂系统，大量的实证研究表明，交通运输网络和其他网络一样，具有复杂网络的结构特性[12-15]，} { 54 % : 所以在对交通网络拓扑特性进行系统分析的基础上，结合复杂网络当中的一些技术与方法，进而提出有效地缓解交通拥堵的道路规划及控制策略，} 是具有重大科学意义的课题。

{ 100 % : 城市是人口居住、社会活动、商务往来、物流集散和文化休闲高度集中之地，而交通则是城市的血液和命脉，} { 100 % : 作为一项支撑国民经济和社会发展的关键性基础条件，城市道路交通系统直接关系到国计民生。} { 91 % : 令人遗憾的是，随着交通供需矛盾的加剧，交通拥挤问题已经成为了世界各大城市的常见病之一。}

{ 100 % : 所谓交通拥挤是指交通需求(一定时间内想要通过道路的车辆数)超过某条道路的交通容量时，形成的交通滞留在道路上的现象。} { 95 % : 一般来说，城市交通拥挤的产生有三种方式：} { 100 % : 暂时的路障、永久能力瓶颈和网络某路段需求的随机波动。} { 98 % : 在实际的交通出行中，某一路段的拥堵在城市交通网络中会逐渐向相邻的路段和节点进行传播，进而会造成局部交通的瘫痪。}

//

图1.3 北京上班早高峰

{ 100 % : 交通拥挤的直接危害是使交通延误增大，行车速度降低，带来时间损失，低速行驶还导致耗油量增加及出行成本增加。} { 98 % : 美国德克萨斯交通研究所通过对美国85个城市区域的道路和高速公路系统的机动性的评估，估算2013年，} { 78 % : 美国城市区域的交通拥挤总计造成了37亿 h的出行延误和105亿 L的燃油耗费，分别比2012年增加了7%，} { 75 % : 900万 h和3亿 L，折算的经济损失总额超过630亿美元。}

{ 100 % : 交通拥挤使得汽车尾气排污量增加，导致环境恶化，人民生活质量下降。} { 91 % : 目前各国城市

大气污染物主要是一氧化碳CO、碳氢化合物HC、氮氧化合物、浮沉等，而汽车是这些污染物的主要排放源(达35%左右)。} {95%：研究发现，CO和HC的排放受行驶工况的影响很大，机动车排放CO和HC的数量随速度的增加呈负指数下降[16]。} {96%：美国环境保护总署的一项研究表明，当干道车速由10 m·l e/h增加到20 m·l e/h时，} {87%：碳氢化合物的排放大约减少40%，氮氧化合物的排放大约减少20%[17]。} {100%：此外，交通拥挤还使交通事故增多，而交通事故的发生又使交通拥挤加剧，从而形成恶性循环。} {89%：由此可见，无论哪种方式产生的交通拥挤或拥挤的加剧，不仅会造成巨额的经济损失，而且如果发展严重甚至还会导致城市功能的瘫痪及次生灾害的蔓延[18-19]。}

1.2论文研究问题

1.2.1链路预测

{59%：链路预测是指根据网络通过已知的网络结构等信息预测出网络中未知链接或者是预测出不存在的链接。} {51%：下面给出链路预测两种情形的定义。}

定义1 复杂网络 (Complex Network)： {56%：复杂网络本质上是一个图形结构 $G = (V, E)$ ，其中 V 表示复杂网络当中的节点的集合，} {47%： E 表示网络当中的边的集合，若网络当中的两个节点存在交互关系，则他们之间存在一条边 e } 复杂网络 (V, E) 中节点的数目为 $|V|$ ，边的数目为 $|E|$ ，若该网络为全联通网络，全集 U 的数目为 $|V|^*(|V|-1)$ 。

定义2 链路预测之丢失边预测 (Link Prediction of Missing Links)： 给定复杂网络结构 $G(V, E)$ 以及缺失边的比例 p ，缺失边预测问题的目标是预测出边的集合 E' ，并将 E' 添加到复杂网络 $G(V, E)$ 之中。 E' 满足 $E' = \{e_1, e_2, e_3, \dots, e_n\} (U - E)$ 且 $|E'| = p^* |V|$ 。

定义3 链路预测之丢错误边预测 (Link Prediction of Spurious Links)： 给定复杂网络结构 $G(V, E)$ 以及错误边的比例 p ，错误边预测问题的目标是预测出边的集合 E' ，并将 E' 从复杂网络 $G(V, E)$ 之中删除。 E' 满足 $E' = \{e_1, e_2, e_3, \dots, e_n\} E$ 且 $|E'| = p^* |E|$ 。

{42%：现我们使用某个链路预测算法，根据网络现有的连边信息，算法会对所有未连接的边进行评分。} 若我们选取边 e ，该边的两个端点分别为 x, y ，根据链路预测算法，我们计算得到 e 的分值为 S_{xy} 。 {44%：将评分从高到低进行排序，分值越高表示边出现的概率越大。} 以图1.4为例，左侧是完整的网络，包含8个节点和11条边。 右侧的网络由于某些原因丢失了两条边{ag、fd}。 现我们希望利用链路预测的方式来恢复这两条边，由于网络节点数为8，所以共有节点对28对。 我们利用{ab、ac、ae、gh、gf、ch、ef、dh、dc}这样9条边的已知信息， 来对28-9=19条边进行评分，好的链路预测算法对于{ag、fd}这两条未知边的评分将高于其他17条不存在的边， 此时我们优先恢复评分较高的边，即{ag、fd}。

图1.4 链路预测图示

1.2.2交通拥堵恢复序列预测

{52%：交通运输网络的运行效率影响着每个人的日常生活。} 现有的关于交通拥堵问题的研究主要集中在提升网络的负载能力之上，其中主要的研究包括节点资源分配、交通路由策略的研究和交通网络的设计。 虽然这些设计及控制方式有效的提升了网络的运载能力，但是随着日益增加的出行需求，交通拥堵问题日益增多。 本文需要解决的交通拥堵问题便是，当交通网络当中出现多条线路（边）的拥堵之时， 以怎样的一个顺序对交通进行

疏导恢复可以在短时间内最为有效的提升网络的运载能力。受链路预测的启发，我们可以将拥堵边近似看作不存在的边，利用链路预测的思维对这些不存在的边进行评分，根据评分依次恢复各条边（运输线路）。本问题的关键点是利用怎样的链路预测算法可以最为近似的刻画各条线路的重要性，进而有效恢复运输网络。

定义4交通网络（Traffic Network）：{40%：交通网络属于复杂网络的一种，本质上也是一个图形结构 $G=(V, E)$ } {47%：其中 V 表示交通网络当中运输节点的结合， E 表示交通网络当中运输线路的集合，线路连接着交通运输节点。}

定义5交通拥堵恢复序列预测（Traffic Congestion Recovery Sequence Prediction）：给定交通网络结构 $G(V, E)$ 以及拥堵边集合 $E=\{e_1, e_2, \dots, e_n\}$ ，交通拥堵恢复序列预测问题的目标是预测出边的有序集合 $E=\{eh_1, eh_2, eh_3, \dots, eh_n\}$ ，并按序恢复 E 。 E 和 E 满足 $EE=E$ ，且 $|E|=|E|$ 。

{51%：如图1.5所示，现有的交通网络包含节点集合{1, 2, 3, 4, 5}，边集合{[1, 5], [1, 3], [2, 5], [2, 4], [2, 3], [3, 5], [4, 5]}，现在网络当中{[1, 3], [4, 5]}两条边发生了拥堵。在现有疏通资源有限的前提下（如道路上只有一辆清障车），我们面对两种线路的恢复选择。显然[1, 3]和[4, 5]的恢复顺序对于网络整体流量的改变是不同的，现在我们想找出这样的一个恢复顺序，该恢复顺序最有利于网络流量的提升。}

对于该问题进行拓展，若三条边发生了损毁，可供我们选择的恢复顺序便有了 $3! = 6$ 种。

图1.5 交通拥堵恢复序列

1.3 国内外研究现状

1.3.1 链路预测

{69%：早先的链路预测算法主要是基于马尔科夫链和机器学习的方法，文献[20]应用马尔科夫链进行网络的链路预测和路径分析。} {70%：文献[21]将基于马尔科夫链的预测方法扩展到自适应性网站的预测中。} {93%：此外，文献[22]提出一个回归模型在文献引用网络中预测科学文献的引用关系，方法不仅用到了引文网络的信息，还有作者信息、期刊信息以及文章内容等外部信息。} {95%：应用节点属性的预测方法还有很多，如文献[23]利用网络的拓扑结构信息以及节点的属性，建立了一个局部的条件概率模型进行预测。} {95%：文献[24]基于节点的属性定义了节点间的相似性，可以直接用于进行链路预测。} 然而该类方法需要节点的属性等外部信息，然而很多的情况下这些信息是难以获取的。比如在社交网络之中，我们可以轻易得知个体的连边关系（如果为好友则为连边），但是对于个体的属性信息却是难以获得的。我们难以获得个体的出生日期或者兴趣爱好等个体属性，即使获取了节点的属性值也难以判别该属性是否真实值。此外，该类方法的另一个弊端便是计算的复杂度极高，因此该类方法在实践之中使用的并不是很多。

{53%：近些年，基于网络的拓扑结构进行链路预测的方法逐渐成为主流。} {53%：基于网络的拓扑结构进行的链路预测方法又分为基于相似性的链路预测和基于似然分析[的链路预测。}

{72%：利用节点间的相似性进行链路预测的一个重要前提假设是，两个节点之间相似性越大，他们之间链接的可能性越大。} {69%：基于相似性的链路预测算法又分为基于局部信息算法的相似性指标、基于路径的相似性指标、基于随机游走的相似性指标和其他的类似性算法。} {63%：基于共同邻居的相似性指标包含众多算法：} {48%：大度节点有利指标[25]、大度节点不利指标[26]、AA指标[27]、RA指标[26]等；} {71%：基于路径的相似性指标包含局部路径指标[28]、Katz指标[29]等；} {58%：基于随机游走的指标又包含了基于随机游走的

余弦相似性[30]，有重启的随机游走指标[31]。} 除此之外，还包含[32]认为结构一致性越强的网络越容易被准确预测丢失的链路。 { 42 % : 我用结构一致性，提出了一种新的名为结构微扰法的新的链路预测方法。}

{ 100 % : 链路预测的另一类方法是基于最大似然估计的。} { 85 % : 文献[33]认为，很多网络的连接可以看作某种内在的层次结构的反映，基于此，} { 94 % : 文献提出了一种最大似然估计算法进行链路预测，该方法在处理具有明显层次组织的网络(如恐怖袭击网络和草原食物链)时具有较好的精确度。} { 96 % : 但是，由于每次预测要生成很多个样本网络，因此其计算复杂度非常高，只能处理规模不太大的网络。} { 73 % : 文献[34]假设观察到的网络是一个随机分块模型的一次实现，在该模型中节点被分为若干集合，两个节点间连接的概率只与相应的集合有关。} { 91 % : 文献所提出的基于随机分块模型的链路预测方法，可以得到更好的结果。} { 49 % : 相比于基于相似性的链路预测方法，基于似然分析的链路预测可以获得更高的准确性，但是计算复杂度极高，} 以至于在处理几千个节点之时便感到十分费力，因此更不用说是去处理社交网络等大型网络了。 { 48 % : 因此，近些年关于链路预测的研究主要还是聚焦于基于相似性的方法。}

{ 44 % : 除了对于不含权重的无向网络的链路预测，链路预测还包含着加权网络的链路预测[35]、有向网络的链路预测[36]、二部分网络的链路预测[37]。} 链路预测的应用方向除了预测蛋白质之间的相互作用、社交网络的好友推荐之外还包括网络重构、网络模型评价方法、节点的标签分类等。 关于三类算法的时间复杂度、准确率、信息获取难度以及应用场景见表1.1。

表1.1 链路预测算法比较

链路预测算法

时间复杂度

准确率

信息获取

应用能力

机器学习

最高

小规模网络

基于似然分析

容易

小规模网络

基于相似性

不低

容易

大规模网络

1.3.2 交通拥堵及恢复序列的预测

现代社会在很大程度上依赖于一些基础设施的有效运行，这些基础设施包括因特网、航空网络、电力网络、航空网络等[38]。 { 48 % : 这些系统既不随机又不规则，因此可以使用复杂网络来对其进行抽象。 } 在复杂网络当中，节点表示网络当中的基本单元，边表示这些基本元素之间的交互[39]，例如航空网络当中，节点表示机场边表示机场之间的航线。

在交通网络的研究当中，提升网络的运载能力是其主要的目标之一。 在本文当中的一个主要研究课题便是交通网络当中的拥堵问题。 在过去的几十年当中，如何避免交通的拥堵一直是一大热点并被广泛研究着。

网络的运载能力通常是由网络在非拥堵情况下所能承受的最大流量所衡量的[40]。 关于交通拥堵问题的研究主要分为网络结构的设计、节点的资源分配以及路由策略这三大类。

在网络结构对于交通拥堵影响方面：

设计一个有效的网络结构来提升网络的运载能力是一项充满挑战的任务。 Boccaketti等人意识到在运输网络当中，网络的整体运载能力很大程度上依赖于每条边上的运载需求。 { 44 % : 这也就是说，机场当中的出行乘客或者地铁当中的工作出行人员。 } 作者通过删除掉网络当中的一小部分节点来模拟交通崩溃或者是因特网当中的路由损坏的情况。 他们发现在同质网络当中，不管是随机的损毁或是刻意的攻击，网络都不会出现大规模的拥塞现象。 然而在异质性的网络当中，对于Hub节点的攻击会导致严重的网络瘫痪。 因此，交通运输网络设计为同质性的网络会更有效的应对突发性的攻击[41]。 Toroczkai 等人通过通过随机和无标度的方式生成了梯次化互穿聚合物网络，他们在这两类网络之上比较了交通的拥堵情况。 他们发现通过无标度的方式生成的网络更加不易于发生拥堵现象[42, 43]。 Roger 等人的研究再次证实了交通网络只有处理的流量在一定范围内才不会发生拥堵，当网络中的流量超过一个阈值便会导致网络的整体崩溃。 他们同时发现网络在发生拥堵之时可分为三个阶段。 首先，随着网络流量的加大，虽然节点的处理能力降低了，但是交通系统暂时不会崩溃。 然后，随着流量的继续加大，节点处理能力不会继续下降。 最后，当处于临界值之时，继续增加少量流量便会导致交通网络的整体崩溃 [54]。

在节点的资源分配方面的研究：

在现实的场景当中，一旦交通系统设计成型便很难对其结构进行改变，所以上述研究的有效性主要存在于交通网络的设计阶段。 上述关于交通网络拓扑结构的研究思路是建立在网络当中所有的节点或者边的处理能力相同前提下。 这一简化的场景让研究人员更好的将精力集中于网络的拓扑结构对于网络运载能力的影响，然而在真实的交通系统当中资源在网络当中的分配极少有完全平均的情况。 { 42 % : 在交通网络的资源分配对于网络影响的研究方面，Zhang等人提出了一个非常实用而且高效的方法来进行资源分配。 } 他们的团队发现在资源总量一定的前提下，通过给处理流量较多的节点更多的资源，处理流量较少的节点较少的资源可有效的提升网络的运载能力。 通过在异质性网络上的实验，他们发现该分配方案可有效的提升网络的运载能力达到两个数量级[45]。 Zhao 等人测试了两种资源分配方案： (1) 节点的处理能力与该节点的度正相关； (2) 节点的处理能力与该节点的介

数正相关。 { 46 % : 他们发现 , 根据方案1 , 随机网络和无标度网络比规则网络更不易发生拥堵。 } 根据方案2 , 针对不同的拓扑结构以及网络尺寸 , 网络的运载能力相差不大 , 所以该方案更适合于在路由网络当中的使用 [46]。

在分析不同的路由策略对于交通影响方面 :

Daniele 等人在无标度网络上测试了一个路由策略 , 该策略平衡了交通流量和网络拓扑 , 他们发现不同的路由策略对异质性网络的影响要大于同质性的网络 [47] 。 Gang 等人提出了利用有效路径替代最短路径的路由方案 , 他们发现在交通网络当中度较大的节点更容易发生拥堵 , { 42 % : 他们利用有效路径的算法将通过中心节点的流量分散到非中心节点之上。 } 通过模拟实验发现 , 有效路径这一路由策略相比较最短路径算法 , 可提升网络的运载能力达 10 倍数之多 [48] 。 { 46 % : Xiang 等人通过队列记录节点的流量信息的方式提出了一种全局的动态路由策略。 } { 47 % : 通过该策略 , 网络的运载能力得到了进一步的提升。 } 该方案虽然提升了网络的最大运载能力 , 但是会导致出行的整体时间的增加 [49] 。 谭飞提出了一种将网络的静态拓扑与流量的静态变化相结合的路由策略 , 该策略有效的缓解了 Hub 节点的处理压力 [50] 。 Danila 等人提出了对于介数较大的节点绕行的方案 , 通过维护一个路由表 , 该路由方案相比于最短路径的路由方式可显著的提升网络的运载能力 [51] 。

[52] 文章从节点控制角度出发 , 通过对节点分流率这一因素入手 , 建立城市交通拥堵疏导的节点控制策略模型 , 在动态系统最优控制模型的基础上加入分流率作为控制变量 , 探索拥堵环境中通过人工干预对交通路网中交通流的运行规律的作用。

1.4 本文主要研究工作及贡献

针对链路预测以及交通拥堵恢复序列问题这两个问题。 { 43 % : 本文提出了将聚类系数加入现有 9 种基于局部相似性的链路预测指标之中 , 对其进行优化。 } 其次 , 应对交通拥堵问题 , 我们将其近似看作链路预测问题 , 通过链路预测算法得到交通拥堵恢复序列。

{ 61 % : 1) 基于聚类系数的链路预测算法研究 }

{ 50 % : 由于基于局部信息的链路预测算法具有较高的准确度以及较低的时间复杂度 , } { 41 % : 其应用范围相较于其他两类链路预测算法也有着更为广泛的应用场景。 } { 45 % : 所以 , 我们的主要研究点是对于基于局部信息的链路预测算法的优化。 } { 76 % : 由于共同邻居指标、 Salton 指标、 Jaccard 指标、 Sorensen 指标、大度节点有利指标、大度节点不利指标、 } { 50 % : Adamic- Adar 指标、资源分配指标和 LHN 指标这 9 种指标多基于共同邻居信息 , } { 41 % : 没有区分共同邻居以及共同邻居之间的链接对于指标的影响 , 所以其预测准确率并非很高。 } 经过对网络的形成机制的研究之后 , 我们发现网络节点的聚类系数这一网络特征与网络的预测准确性之间存在一定的关联性。 通过对现有 9 种经典的链路预测指标进行分析之后 , 我们将网络节点的聚类系数这一特征加入到这 9 种算法之中 , { 44 % : 对其进行优化 , 之后的实验结果证明了改进后的算法相较于原先的 9 种算法有着更高的预测准确率。 }

在实验部分 , 除了一般链路预测文献的常规实验。 我们同时将算法应用到了缺失边的预测以及错误边的预测之上 , 其中错误边的预测在先前的研究中很少涉及。 通过对美国航空网络、线虫神经网络、国际足球联赛网络、爵士乐演奏家合作网络、跆拳道社会网络这 5 类显示数据上的实验 , 表明了改进后的算法在缺失边的预测以及错误边的预测之上均有超过原算法的准确率。 除此之外 , 过往研究将测试集的比例固定在 10% 这一数值之上 , 我们的实验对此进行了扩展 , 我们将测试集的比例伸展到了 4% 到 20% 范围之间 , 实验在不同信息缺失率之上均表明了算法优化后的准确率得到了提升 , 这一参数范围的变化也验证了我们算法的健壮性。

{ 41 % : 2) 利用链路预测思路进行交通拥堵恢复序列预测的研究 }

通过文献的阅读，我们发现关于交通拥堵问题的研究主要分为3大类。相关规模下的不同拓扑结构的网络其交通运载能力会有很大的不同，第1类方法希望通过网络结构的设计来减少交通的拥堵，这主要发生在交通网络的设计阶段。相同结构的交通网络，若其交通节点的流量处理能力不同，其相应的运载能力也会有所差别。第2类方法考虑的是对于节点资源（处理能力）的合理分配。该类方法希望找到更优的资源分配方案，进而提升网络的运载能力，这发生在设计阶段之后。第3类方案是对于路由策略的研究。{ 47 % : 交通出行当中迪杰斯特拉算法是最为广泛使用的出行方案。} 然而，在交通网络当中，若大家均使用最短路由策略进行游走，便会导致部分介数较大的边发生拥堵，该类方案通过改进路由方案来提升网络的运载能力，这主要发生在交通的控制阶段。

上述3种关于交通问题的研究目的都是提升网络的运载能力，这主要发生在交通拥堵发生之前。但是，这些方案无法完全避免交通的拥堵。一但发生拥堵之后，我们需要一个紧急的恢复方案。由于链路预测可以对网络中边的重要性进行区分，本文尝试将链路预测的方法运用到交通恢复序列的预测之上。{ 41 % : 基于对路网节点重要性的分析，本文提出了基于链路预测方法的一系列反相关性算法（RAPA、RSA、RLP）。} 当交通网络当中多条线路发生瘫痪之时，通过该类反相关性算法，我们对于这些拥堵的线路进行重要性评分，优先疏通评分较高的线路。通过在模拟网络（BA）网络以及美国航空网络、加州路网这3类交通网络上的模拟，我们得出了反相关性算法在交通拥堵序列的预测上的优良表现。也就是说，在交通发生拥堵之后，利用链路预测反相关性算法能够最有效地恢复交通网络运载能力。我们在得出初步结论后，今后将对课题进行拓展，将研究内容向更加实际的方向上推进。利用链路预测的思路来解决交通恢复序列这样一个问题尚属于探索型课题，希望本课题的成果能对以后相关领域的研究提供一点指引或启发。

1.5 本文的章节安排

本文共分为六章，每章具体内容安排如下：

第1章，绪论部分。{ 41 % : 介绍了链路预测的相关背景，以及在现实研究当中的应用价值，提出了利用节点的聚类系数这一信息对现有的链路预测当中的基于相似性的局部算法进行改进。} { 40 % : 此外，我们还提出了基于链路预测思维进行交通拥堵恢复序列的预测。}

{ 66 % : 第2章，链路预测方法分析。} 我们将链路预测分为了丢失边的预测以及错误连边的预测这两种情形。我们分析了9种基于局部信息的链路预测算法，该类算法多基于节点本身以及其邻居节点的信息来进行链路预测。{ 45 % : 此外我们介绍了链路预测当中对于数据集的划分和链路预测准确性的评价指标。}

{ 50 % : 第3章，基于聚类系数的链路预测。} 通过分析讨论，我们发现聚类系数这一指标可进一步区分节点间不同的共同邻居，因此我们将聚类系数加入现有9种链路经典的链路预测算法之中，得到了改进后的算法。{ 43 % : 5类数据集之的对比实验表明了聚类系数可提高链路预测算法的准确率。}

第4章，交通拥堵问题的研究。本章想研究的问题是当出现多条交通线路同时发生拥堵之后，在资源有限的前提下，以怎样的一个顺序对线路进行修复可以最大程度上恢复整个网络的流量。本章提出了利用链路预测的思维进行交通拥堵恢复序列的预测，并介绍了交通的流量模拟方法。

第5章，交通拥堵恢复序列算法。{ 42 % : 基于链路预测，我们提出了交通恢复序列算法，算法分为两类

, 一类是直接的链路预测算法 , } 另外一类是结合交通网络实际情况而提出的基于链路预测的负相关性算法。真实以及模拟数据集上的实验结果表明了基于链路预测的负相关性算法可很好的预测出恢复边的顺序。

第6章，总结和展望。 对本文的研究内容做一总结，并对还存在的问题以及以后的研究计划进行分析和总结。

第2章链路预测方法分析

{ 59 % : 本章主要介绍了链路预测的基础概念及相关技术。 } { 43 % : 链路预测问题主要包含了对缺失边的恢复以及对错误链接的纠正这两类。 } 对于链路预测算法准确率的计算，我们需要经过数据集的划分、利用链路预测的相似性指标对边进行评分、 { 52 % : 利用某种评价指标计算链路预测算法的准确率这3个步骤。 } { 59 % : 其中对于相似性指标的分析是本章的讨论重点。 }

2.1链路预测问题描述

链路预测中主要有两类的需求，这两类需求分别是丢失边的预测以及错误连边的预测。丢失边的预测主要是指网络中存在但是尚未被发现的边，如人类的蛋白质网络中，大量的蛋白质相互作用仍未被人类发现。错误连边主要是指实际网络中不存在，但是被错误加入其中的边，如QQ有时会向我们进行好友推荐，而实际生活中我们并不认识该人。

以下将对链路预测中的丢失边预测以及错误连边预测分别进行讨论。

2.1.1 丢失边的预测

在链路预测中，为了测试链路预测（链路预测的评分指标）的准确性，我们通常将E分为两个部分，训练集ET和验证集EP。训练集和测试集的划分原则为 $ET \cup EP = E$ ，且 $ET \cap EP = \emptyset$ 。{ 44 % : 我们会将属于U但是不属于E的边称为不存在的边，在预测算法的准确性时，我们是基于训练集中的信息，验证集仅用作数据的验证。 }

{ 42 % : 如图2.1所示，其中图左侧是完整的网络，该网络拥有13个节点，18条边，全集包含 $(13 \times 12) / 2 = 78$ 条边，其中不存在的边有60条。 } 我们在原网络的19条边中选出4条作为验证集，如右侧图中虚线所表示的{ci, dm, ak, mg}，剩下的14条边就是训练集。之后我们就需要根据我们选用的算法将64条未知边（其中验证集4条，不存在的边60条）赋予一个分数值，并按降序排序。假如在全部的分数值中，我们的验证边的分数相比较不存在的边排名越靠前，那么就代表我们的算法对于此网络的精确地越高。极端的结果是，根据某种链路预测的评分，4条验证边的得分均高于60条不存在的边，这时的预测准确率便为100% [3]。

2.1.2 错误边的预测

与丢失边的预测略有不同，对于错误边的预测主要是找出网络中不存在的边，也就是常说的噪音干扰。{ 49 % : 为了测试链路预测指标的准确性，我们需要模拟错误连边的情况。 } 在图2.2中左侧的网络是正确的信息，我们将网络中添加的边用作验证集EP以模拟噪声情况，此时的训练集 $ET = E + EP$ 。右侧网络中的虚线表示验证集合，即

图 2.1 复杂网络丢失边预测

$E=\{ab, ai, ak, ac, kj, km, kl, lc, jm, cm, mh, mg, cd, md, me, mf, ij, hg\}$ 共18条， $EP=\{bc, de, ef, jh\}$ 共4条。同丢失边的预测，链路预测是基于训练集中的信息来进行预测，此处的 $ET=E+EP$ ，同丢失边的情况略有不同。

图2.2 复杂网络错误边预测

2.2链路预测算法流程

{ 59 % : 链路预测算法的计算流程如图2.3所示。 } { 40 % : 该流程的主要目的是计算链路预测指标的准确性。 } 预测指标的准确性分为对缺失边的预测和错误连边的预测这样两种情况。该算法流程的核心点在于链路预测指标的选取，对于该部分我们在图2.3当中对其进行了圈定。 { 42 % : 在链路预测研究领域，通常将链路预测指标（打分函数）称作链路预测算法。 } { 43 % : 所以，在下文当中若提及链路预测算法，等价于某种特定的链路预测指标。 } 链路预测当中的评分指标会根据训练集的信息，对验证集以及不存在的边进行评分，{ 50 % : 评分越高的边，我们在链路预测当中认为其存在的可能性越大。 } { 46 % : 优先恢复链路预测指标评分较高的未知边。 }

{ 61 % : 图 2.3 链路预测算法流程图 }

2.3经典的链路预测方法

{ 45 % : 本章所选取的算法是链路预测中最为经典的基于节点的局部相似性算法。 } { 60 % : 该类算法属于基于节点相似性的链路预测算法，该类算法的一个前提假设是两个节点之间相似性越大，两点间存在链接的可能性也越大。 } 在链路预测流程当中，这9种指标对应于图2.3当中画虚线的那一步骤（链路预测评分指标）。

{ 48 % : 度量节点的相似性有很多种方式，比如利用节点的属性就是一种简单而又直接的方式。 } 举例来说，若两个人具有相同的年龄，所在城市相同，职业相同，兴趣相同时，我们认为这两个人具有很高的相似性，这两个人认识的可能性也大于属性差异较大的两个人。

{ 44 % : 基于节点属性的相似性虽然可以获得较高的链路预测准确率，但是获取节点的属性往往是十分困难的。 } 比如我们常用的微博，我们的很多类信息是对他人不可见的（身份证号，手机号等），即使获得了用户的一些信息，该类信息也未必真实可靠，比如说年龄、身高、性别、所在地等等。由于节点属性的难以获取，所以利用节点属性进行链路预测便遇到了很大的障碍。

{ 52 % : 与节点属性相比较，网络的拓扑结构更易于获取同时也更为可靠。 } { 48 % : 基于拓扑相似性的链路预测算法的准确率高低很大程度上取决于该算法是否符合网络的形成机制。 }

相比于计算全局网络结构的算法，本文所选取的算法有计算复杂度较低的特点，这几种方法也常见于对改进算法的比较之中。 { 51 % : 链路预测指标的评分方式如图2.4所示。 } 针对丢失边的预测和错误边的预测，评分函数所计算的节点对也有所不同。有图2.4可以看出来，针对丢失边的预测，链路预测的评价指标将计算所有的未知边，从未知边当中恢复评分较高的边。然而，针对错误边预测，链路预测指标所计算的边是已经观察到的边，根据对这些边的计算结果，我们从中删除得分较低的一部分边。

2.4 链路预测指标的评分方式

{ 72 % : 2.3.1 基于共同邻居指标的链路预测方法 }

Common Neighbors (CN) : { 57 % : 对于网络中的节点 v_x , (x) 表示该节点的所有邻居, 同样的(y) 表示节点 v_y 的所有邻居集合。 } { 42 % : 下式表示的便是节点 v_x 和节点 v_y 的所有共同邻居的数目, 该算法认为两个节点间共同邻居的数目越多的话, 该两个节点认识的可能性就越大。 }

$$S_{xy} = |(x)(y)| \quad (2.1)$$

在现实的生活中, 如果两个陌生人有很多的共同好友, 那么这两个人在未来认识的可能性也就越大。再例如, 在科学家合作网络中, 某两位科学家有着越多的共同合作者, 那么这两个人在未来合作的可能性就越大。

如图2.5所示, 对于点 A 和点 D 来说, 根据定义 $(A)=\{B, C\}$,

$(D)=\{B, C\}$, $(A)(D)=\{B, C\}$ 。因此按照上述公式来计算, $S_{AD}=2$ 。

而对于点 A 和点 E 来说, $(A)=\{B, C\}$, $(E)=\{B\}$, $(A)(E)=\{B\}$, 得 $S_{AE}=1$ 。

图 2.5 CN算法图示

{ 41 % : 在计算两个节点 (node1, node2) 的共同邻居时, 只需遍历网络中的节点 V, 若某个节点 node_x 与 node1、node2 均相连, } 那么我们将共同邻居数目 +1, 直至遍历完所有的节点。故该指标的计算复杂度为 $O(|V|)$ 。

2.3.2 基于Salton指标的链路预测方法

{ 48 % : Salton指标又称作余弦相似性指标, k_x 和 k_y 分别表示节点 v_x 和节点 v_y 的度的大小。 } 与 CN 指标相比较, 该式多出的分母部分表示在两个节点共同邻居数目一定的时候, 本身节点度越小, 两个节点连接的可能性越大。比如在微博当中, 大 V 之间会有着更为广阔的社交空间, 所以他们会有更多的共同邻居, 所以反而本身好友较少的前提下有着相同的共同邻居的两个节点在未来认识的可能性较大。

$$S_{xy} =$$

$$|(x)(y)|$$

$$(2.2)$$

{ 47 % : 根据网络拓扑结构, 我们可方便的得到每个节点的度, 故分母值求解的时间复杂度为 $O(1)$, 故该指标的计算复杂度为 $O(|V|)$ 。 }

2.3.3 基于Jaccard指标的链路预测方法

Jaccard 指标提出于一百多年前, 相比于 CN 指标, 该指标认为在共同邻居数目一定的前提下, 两节点邻居数目越多, 相连的可能性越小。 { 41 % : 或者说, 两个人的总邻居数目一定, 共同邻居的数目越多, 这两个人相识的可能性越大。 }

$$S_{xy} =$$

$$|(x)(y)|$$

$$|(x)(y)|$$

(2.3)

类似于CN指标，该指标只需要在求得节点x和节点y的邻居信息的前提下，求出两个节点的并集，该指标的计算复杂度为 $O(|V|)$ 。

2.3.4 基于Sorensen指标的链路预测方法

该指标常用于生态系统网络的预测，该公式可看做是Salton指标的一个简单的变形。

$$S_{xy} =$$

$$2|(x)(y)|$$

(2.4)

类似于Salton指标，分母部分两个节点的度运算改为相加，该指标的计算复杂度为 $O(|V|)$ 。

2.3.5 基于大度节点有利指标的链路预测方法

hub promoted index(HPI): { 45%：该指标主要被用来刻画新陈代谢网络中每对反应物的拓扑相似性，该公式认为在共同邻居的数目一定的前提下，度较小的节点更有利与网络的连。 }

$$S_{xy} =$$

$$|(x)(y)|$$

$$\min\{$$

$$y\}$$

(2.5)

该指标的计算量在于求解两个节点的共同邻居，分母部分的两个节点度运算为求其中度较小的数值，该指标的计算复杂度为

$$O(|V|)$$

2.3.6 基于大度节点不利指标的链路预测方法

hub depressed index(HDI): 该定义与HPI类似，只是分母部分略有不同，该公式认为在共同邻居的数目一定的前提下，度较大的节点更有利与网络的连接。

$$S_{xy} =$$

$$|(x)(y)|$$

$$\max\{$$

$$y\}$$

(2.6)

与大度节点有利指标类似，分母部分的两个节点度运算为求其中度较大的数值，将该指标的计算复杂度为 $O(|V|)$

2.3.7 基于LHN指标的链路预测方法

该公式的形式与Sorensen类似，仅将分母部分的两个元素由相加改为了相乘。

$$S_{xy} =$$

$$|(x)(y)|$$

(2.7)

{ 41%：在CN的基础上，增加了分母的度计算，时间复杂度不变，依然为 $O(|V|)$ 。 }

2.3.8 基于Adamic-Adar指标的链路预测方法

Adamic-Adar (AA) : { 48%：这是对于CN指标的一种改进，该指标不再继续认为共同邻居的贡献作用相同，该指标的主要思想是度较小的共同邻居节点的贡献度大于度较大的共同邻居节点。 } 符号

表示计算所有元素的和。

$$S_{xy} =$$

$$z|(x)(y)|$$

$$\log$$

(2.8)

举例来说，在微博当中受关注较多的人大多是领域专家或者名人，如姚晨。因此，关注者之间不一定存在很大的共同点。相反的，若两人共同关注的人的粉丝数目很少，此时，这两人之间具有重叠的社交圈的概率大大增大，因此两人之间相互认识的可能性也大大增大，即两者之间存在连接边。

该指标在遍历所有节点的同时，将共同邻居节点的度记录下来，时间复杂度不变，仍旧为O(|V|)。

2.3.9 基于资源分配指标的链路预测方法

Resource Allocation (RA) : { 57 % : 指标在形式上类似于AA指标，该算法是周涛、吕林媛等人受到网络中资源分配模型的启发而提出的。 } { 51 % : 网络中不直接相连的两个节点vx和vy，vx需要借助共同邻居向vy节点发送资源。 } { 50 % : 假设每个节点都有一个单位的资源并平均发送给他的共同邻居，显然vy从vy处接受的资源总量为下式中的Sxy。 }

$$S_{xy} =$$

$$\frac{z(x)(y)}{|N(y)|}$$

(2.9)

图2.6计算RA指标

如图2.6所示，节点x和节点y的共同邻居为c1，c2。{ 49 % : 判断x节点和y节点之间存在连边的可能性，我们使用RA指标对节点对Sxy进行评分，那么节点x和节点y的相似性分值为 }

$$S_{xy} =$$

$$= 0.42$$

{ 43 % : RA指标和 AA指标最大的不同处是赋予共同邻居权值大小的区别，当网络度的平均值较小时两个指标差别不大， } { 44 % : 当网络度的平均值较大时，差异便变得明显。 }

同AA指标，时间复杂度为O(|V|)。

{ 50 % : 上述9种链路预测指标中以 CN指标为基础，考虑两个端点度的影响，从不同的角度和方式产生了 Salton指标， } { 77 % : Jaccard指标， Sorensen指标，大度节点有利指标，大度节点不利指标、LHN指标。 } { 48 % : 在共同邻居指标的基础上，若考虑两个端点的度的大小，根据不同的生成方式我们便得到了AA指标和RA指标。 } { 58 % : 上述9种指标均为考虑共同邻居节点之间的相互关系。 }

2.4 数据集划分

由上述关于链路预测的定义，我们发现在链路预测中我们会用到两类边，该两类边称作训练集和验证集，不同的验证集和训练集的划分会在一定的程度上影响结果的准确性。该两类边的选取主要是由验证集的选取决定的，因为一旦获得验证集 EP之后，在丢失边的预测中训练集 $ET = E - EP$ ，在错误边的预测中 $ET = E + EP$ 。{ 41 % :

常见的验证集的选取主要有随机抽样、逐项遍历、k折叠交叉检验，滚雪球抽样。} 现对该几类方法做简要的介绍，为了便于介绍，我们将以未知边的预测为例，关于错误连边中验证集的选择稍加拓展即可。

2.4.1 随机抽样

在网络 $G(E, V)$ 中，共有节点 N 个，边 M 条。 { 50% : 现在我们需要划分其中比例为 p ($0 \leq p \leq 1$) 的边作为验证集，随机抽样即从 M 条边中随机抽取 $p \cdot M$ (若非整数，向上取整) 条边组成验证集。 }

2.4.2 逐项遍历

{ 48% : 针对某些较小的网络，逐项遍历法是一种更为精准的数据集划分方式。 } { 64% : 在逐项遍历法中，我们每次从网络中选取一条边作为验证集，其余的边则作为训练集使用。 } 对该条边进行预测时我们会得到一个准确率，依次遍历网络中的 M 条边，我们会得到 M 次的准确率，{ 48% : 求 M 次准确率的平均值，该平均值即为我们算法的准确率。 } { 40% : 该方法相较于随机抽样法中每条边都有被选中的机会。 }

2.4.3 折叠交叉检验

{ 48% : 该方法是逐项遍历法的一种改进方法，我们将边随机分成 k 份，选取其中的一份作为验证集，剩余部分作为训练集，此时我们会得到一个预测准确率。 } { 44% : 遍历 k 份边，共得到 k 次的预测准确率，求该 k 次准确率的平均值。 } 10 折叠交叉验证是最常见的方法，此时相当于选取 10% 的边作为验证集，即 $p=0.1$ 。 { 47% : 当 k 等于 M 时，该方法退化为逐项交叉检验法。 }

2.4.4 滚雪球抽样

该方法先随机访问一些被调查者，然后再邀请这些调查者推荐调查对象。 { 63% : 在实际应用中类似于广度优先搜索。 } 在初始化时，我们随机选择一些节点，此后依次访问该类节点的邻居节点，并将这些点加入样本中，该过程直至满足样本的数量要求为止。

2.4.5 其他方法

{ 62% : 其他常用的抽样方法还包括熟识者抽样、随机游走抽样、基于路径抽样。 }

在链路预测的相关研究中，数据集的划分多采用随机抽取的方式，为了便于比较，本文也将采取随机抽取的方式来获取验证集。 { 42% : 此外，先前的研究当中，验证集的选取大多为固定值 10%，为了验证本文提出改进算法的健壮性，} 我们将 p 值设置为 [0.04, 0.08, 0.12, 0.16, 0.20] 这五个不同的值。

{ 74% : 2.5 链路预测准确率评价指标 }

{ 55% : 衡量链路预测算法准确性的指标主要有 AUC、准确度以及排序分三种。 } { 49% : 他们三种指标对于准确度的衡量侧重点有所不同。 } 现分别就三种指标做简单的介绍。

2.5.1 接受者操作特征曲线下面积

Area under the receiver operating characteristic curve (AUC)：在丢失边的预测之中，每次从测试集和不存在的边

当中分别取出一条边，链路预测算法根据训练集信息分别对两条边进行评分， { 52 %：如果测试集中边的评分大于不存在的边，那么加1，如果测试集中边的评分等于不存在的边，那么加0.5， } { 69 %：如果测试集中边的评分大于不存在的边，那么加0。 } { 47 %：如此往复比较 n 次，若测试集中的边分值大于不存在的边的次数为 n'，测试集中的边分值等于不存在的边的次数为 n''。 }

在错误边的预测中，每次从 E 和 EP 中分别取出一条边，如果 E 中边的评分大于 EP 中的评分，那么加 1，如果相等加 0.5，否则加 0。 { 43 %：如此往复比较 n 次，若 E 中的边分值大于 EP 边的次数为 n'，等于的次数为 n''。 }

根据先前的划分，网络中的未知边分为不存在的边和测试集当中的边。 虽然测试集中的边被划分在了未知边中，但好的算法是应该可以区分出不存在的边和测试集当中的边的，即根据算法测试集中边出现的概率要高于不存在的边。 那么 AUC 值的计算公式为：

$$AUC =$$

(2.10)

显然，如果链路预测的算法不具备预测能力， $AUC=0.5$ 。 { 94 %：因此 AUC 大于 0.5 的程度衡量了算法在多大程度上比随机预测的方法精确。 }

2.5.2 准确度

准确度 (Precision)：相比于auc指标，有时我们只关心排在前面的几条结果是否正确。 { 63 %：准确度定义为在前 L 个预测中，预测准确的比例。 } 假设现在的验证集大小为 L，根据出现连边的可能性对于未出现的边进行排序，验证集评分在前 L 中的个数为 m，那么准确率为：

$$Precision =$$

(2.11)

L 大小的选择一般会对结果产生影响。

2.5.3 排序分

{ 60 %：排序分 (Ranking Score) 考虑了验证集在最终排序结果中的位置。 } { 54 %：链路预测算法会对所有未知边进行排序，令 H 表示未知边的集合，re 表示测试边 eE 在排序中的排名。 } 那么这条边的排序分为：

$$RS =$$

|H|

遍历所有的验证集，将各条测试边的排序分相加便得到总的排序分，显然排序分越低预测越准确。

{ 65 %：Precision 考虑的是排在前 K 位的预测是否准确，Ranking Score 侧重于所预测边的排序， } AUC 是三种

当中最为常用的一种指标，他可以很好的从整体上衡量算法的准确度， { 47 % : 本文中关于链路预测的准确度便是以 AUC来进行衡量的。 }

2.6 本章小结

{ 44 % : 本章将链路预测问题分为了丢失边预测以及错误边预测这两种情况。 } { 43 % : 我们分析并介绍了9种经典的链路预测算法，这9种算法基于节点间的局部相似性， } { 42 % : 由于只需利用到节点以及邻居节点的信息，所以这9种算法保持了较低的时间复杂度， } 这9种算法的时间复杂度均为 $O(|V|)$ 。 { 45 % : 由于该9种链路预测算法将节点间的共同邻居均等看待或是不考虑共同邻居节点之间的联系，所以其准确率较低。 } { 46 % : 此外，我们介绍了链路预测方法中数据集的划分以及链路预测准确性 }

{ 60 % : 第3章基于聚类系数的链路预测算法 }

在对9种经典的链路预测算法分析的基础上，本章提出了利用聚类系数对这9种算法进行优化的思路。实验部分，在5类真实的数据集，我们对优化前后共18种算法进行了预测准确率的比较。

3.1 引言

在第2章当中，我们分析并介绍了9种经典的链路预测指标，以上9种指标虽然准确率较高，但依然有着一定局限性。 { 55 % : 例如，CN 算法将所有的共同邻居等同看待，仅利用共同邻居的数量作为节点对间相似度的评分函数，没有区分出不同的邻居对链接预测的影响是不一样的； } { 97 % : AA 算法和 RA 算法虽然区分了每个不同的共同邻居对链接预测的不同的影响力，但是它们都只关注于共同邻居本身，而忽略了这些共同邻居之间的相互影响。 }

{ 45 % : Watts和Strogatz在1998年提出网络中聚类系数的概念，该指标为网络的一个局部性指标。 } { 47 % : 针对网络中的任意一个节点 v_i ，其聚类系数定义为该点所有邻居节点之间的连接数占邻居间最大连边可能性的比例，其数学表达式为： }

$$C_i =$$

1)

(3.1)

{ 45 % : 式子中 $|N(v_i)|$ 表示节点 v_i 的邻居节点之间连边的数目， k_i 表示节点 v_i 度的大小， C_i 即为节点 v_i 的聚类系数大小。 } { 45 % : 根据式子我们可知，在一个全网络图中，每个节点的聚类系数均为 1，在一个树状结构中，每个节点的聚类系数均为 0。 } { 45 % : 在图3.1种，节点 i 的度为 7，节点 i 的邻居间相邻的数目为 5，所以 $C_i =$ }

21

图3.1聚类系数计算示例

{ 47 % : 聚类系数表示了节点间联系程度，高的聚类系数表示节点的邻居之间相识的可能性较大。 } { 100 % : 如果节点 v_1 连接于节点 v_2 ，节点 v_2 连接于节点 v_3 ，那么节点 v_3 很可能与 v_1 相连接。 } { 100 % : 这种现象体现了

部分节点间存在的密集连接性质。} 在生活中，若将人类的生活抽象为社交网络，若某人的聚类系数较大，那么该人的朋友之间相互认识的可能性越大，那么我们认为该人的好友之间产生连边的概率较大。

为了克服第二章所提到9种算法的缺点，本文提出了基于聚类系数的链路预测方法。{ 64 %：该算法所基于的想法是，节点的聚类系数越大，则尚未相连的邻接点间存在连边的概率越大。} 基于这个思想，本文提出一种基于聚类系数的新算法，该算法同时考虑了两端节点以及共同邻居的本身特征，能反应出节点间的相互作用关系。

{ 48 %：3.2基于聚类系数的链路预测指标准确率计算 }

{ 42 %：链路预测算法的伪代码如图3.2所示，对应于图2.3中的链路预测流程。} 我们输入某种链路预测指标，将经过该流程的处理，我们将得到该指标在某种网络当中的预测准确率，也就是我们链路预测当中常说的该链路预测算法（由于链路预测指标是算法的核心部分，{ 55 %：故也指代为链路预测算法）的准确率。}

图3.2代码中 if (kind==0) 这个条件判断语句是判断链路预测算法是在计算丢失边的准确率还是错误连接边的准确率，如果 kind等于0则表示对于丢失边的预测。9-14行的 score () 函数，便是我们的链路预测算法，该算法的功能是根据现有网络信息，对某个节点对进行评分，分数较高的节点对表明该节点对之间存在边的可能性较大。在1000轮的循环当中的某次循环中，若未知集（验证集和不存在边）中的验证集根据 score () 函数的评分高于不存在的边，我们将 sum加1，相等的话加0.5，小于的话加0。经过1000轮次循环后，我们将sum除以1000，所得的值即为该链路预测指标的预测准确率。相应的17-30行表示对于错误边的预测。

{ 53 %：图3.2链路预测指标准确率计算 }

{ 47 %：图3.2链路预测指标准确率计算（序） }

{ 58 %：3.3 基于聚类系数的链路预测算法改进 }

对于第2章中所分析的9种经典链路预测算法，我们引入聚类系数加以区分共同邻居之间联系的紧密程度，得到的改进后的相应指标为：

{ 55 %：3.3.1基于聚类系数化共同邻居指标的链路预测方法 }

$S_{xy} =$

$z|x(y)|$

(3.2)

Clustering Common Neighbors (CCN)：CCN指标与CN指标的区别在于，CN指标根据两点间的共同邻居的数目进行打分，也就是说对应于公式3.2中 C_z 恒等于1。而在CCN指标看来，共同邻居节点的贡献度是不同的，根据对聚类系数的分析，聚类系数较高的共同邻居节点的贡献度较大。如在社交网络当中我们共同关注粉丝较多的人相较于共同关注粉丝较小的人，我们认识的可能性较小。符号

表示计算所有元素的和。{ 41 %：由于 C_z 计算的时间复杂度为 $O(1)$ ，所以相比于CN指标，时间复杂度不变，依旧为 $O(|V|)$ 。}

{ 47 % : 3.3.2 基于聚类系数化 Salton 指标的链路预测方法 }

$S_{xy} =$

$|(x)(y)|$

(3.2)

Clustering Salton (CSalton) : 该指标与 Salton 指标区别在于分子部分 , 类似于 CCN 指标 , CSalton 指标将共同邻居节点的贡献度区别看待 , { 54 % : 同样认为较大的聚类系数的共同邻居节点的贡献度较大。 } { 42 % : 分子部分表示将共同邻居的聚类系数值进行累加。 } { 46 % : 由于分母部分度的计算不影响时间复杂度 , 所以总的时间复杂度为 $O(|V|)$ 。 }

{ 49 % : 3.3.3 基于聚类系数化 Jaccard 指标的链路预测方法 }

$S_{xy} =$

$z|(x)(y)|$

$|(x)(y)|$

(3.4)

Clustering Jaccard (CJaccard) : 改进类似于 CCN , 我们将共同邻居区分看待 , 分子部分进行共同邻居聚类系数值的累加 , 分母同 Jaccard 指标 , 计算时间复杂度为 $O(|V|)$ 。

{ 46 % : 3.3.4 基于聚类系数化 Sorensen 指标的链路预测方法 }

$S_{xy} =$

$z|(x)(y)|$

(3.5)

Clustering Sorensen(CSorensen) : 改进类似于 CCN , 分子部分进行共同邻居的聚类系数值的累加 , 分母部分将两个端点的度值相加。 由于分母部分的计算不影响整体的计算法咋读 , 所以整体的计算时间复杂度为 $O(|V|)$ 。

{ 50 % : 3.3.5 基于聚类系数化 HPI 指标的链路预测方法 }

$S_{xy} =$

$z|(x)(y)|$

$\min\{$ $y\}$

(3.6)

Clustering hub promoted index(CHPI): 改进原理类似于CCN，分子部分进行共同邻居的聚类系数值的累加，分母部分求两个端点度较小的值，
{ 44 % : 分母部分的计算不影响整体复杂度的计算，计算时间复杂度为 $O(|V|)$ 。 }

{ 50 % : 3.3.6 基于聚类系数化HDI指标的链路预测方法 }

 $S_{xy} =$ $|(x)(y)|$ $\max\{$ $y\}$

(3.7)

Clustering hub depressed index(CHDI): 类似于HPI，只是将分母部分求最大值计算改为求两个端点度较小的那个值，计算时间复杂度为 $O(|V|)$ 。

{ 50 % : 3.3.7 基于聚类系数化LHN指标的链路预测方法 }

 $S_{xy} =$ $|(x)(y)|$

(3.8)

Clustering LHN(CLHN): { 43 % : 类似于CSorenson，将对两个端点度的相加改为了相乘，计算时间复杂度为 $O(|V|)$ 。 }

{ 49 % : 3.3.8 基于聚类系数化AA指标的链路预测方法 }

 $S_{xy} =$ $|(x)(y)|$

(3.9)

Clustering Adamic-Adar(CAA): { 46 % : 改进原理类似于 CCN , CAA指标同时看重共同邻居节点度大小和聚类系数的区别 , } { 43 % : 算法认为较大的聚类系数较小的度大小的共同邻居节点更利于促进连边的形成。 } 计算时间复杂度为 $O(|V|)$ 。

{ 46 % : 3.3.9 基于聚类系数化资源分配指标的链路预测方法 }

$S_{xy} =$

$|(x)(y)|$

(3.10)

Clustering Resource Allocation (CRA) : 改进原理类似于 CAA , 只是分母部分赋予权值的方式有所不同 , 当网络度的平均值较小时两个指标差别不大 , { 44 % : 当网络度的平均值较大时 , 差异便变得明显。 } 计算时间复杂度为 $O(|V|)$ 。

3.5 数据集介绍

{ 46 % : 本章实验部分共用到五类数据集 , 这些数据集合是在pajek和stanford两个数据集当中收集而来。 }

Pajek数据集发布在网站<http://vlado.fmf.uni-lj.si/pub/networks/data/>上 , stanford数据集发布在网站<http://snap.stanford.edu/data/>之上。 { 40 % : 这两个网站之上搜集了大量的复杂网络相关数据供复杂网络相关的研究者自由使用。 }

////

(a) Elegans (b) Football (c) Jazz (d) USAir (e) Karate

图3.3数据原始格式

如图3.3所示 , (a)、(b)、(c)、(d)、(e) 分别为Elegans , Football , Jazz , USAir和Karate。 其原始数据均为txt文本格式。 其中Elegans , Football , Jazz , USAir数据格式相同 , 每行表示网络当中的一条边 , 该边由两个节点表示 , 为无向网络。 Karate数据略有不同 , 每条边标注了源节点和目标节点 , 依旧为无向网络。

3.5.1 数据集描述

本章共使用五类数据集 , 他们分别是USAir、Elegans、Football、Jazz、Karate , 现对这些数据做简要的介绍。

USAir表示的是美国的航线信息 , 节点表示机场 , 若两个机场之间有直达航线 , 那么该两点之间存在一条连接边。 Elegans表示的是线虫的神经网络 , 节点表示线虫的神经元 , 突触或间隙表示为连边。 Football数据中的节点代表国家队 , 这些国家中的球员频繁参加国外联赛 , 若国家队中有成员在另外一个国家的联赛中踢球 , 那么这两个国家队之间存在连边信息。 Jazz表示jazz演奏家之间的合作网络。 { 49 % : Karate是社会网络分析领域中的经典数据集 , 该网络构造了美国一所大学空手道俱乐部中34名成员的社会关系 , } 若两人在现实之中是好友关系 , 那么该两人之间存在一条连边。

3.5.2 数据集的网络统计特征

表3.1数据集的网络统计特征

网络

节点

|V|

|E|

密度

聚类系数C

平均度[k]

密度

[d]

异质性H

Elegans

297

2148

0.0488

0.2923

14.4646

2.4553

1.8008

Football

118

0.1983

0.3389

6.7428

2.1226

1.4880

Jazz

198

2742

0.1405

0.6174

27.6969

2.2350

1.3951

USAir

332

2126

0.0386

0.6252

12.8072

2.7381

3.4638

Karate

34

78

0.1390

0.5706

4.5882

2.4081

1.6932

{ 50 % : 表3.1是实验所使用到的5类数据集的相关指标统计。 } |V|代表网络中节点的数目。 |E|代表网络中边的数目。 D表示网络的密度，计算公式为

$2|E|$

$|V|(|V|-1)$

{ 55 % : C 表示节点的聚类系数，上文已经给出了计算公式。 } { 64 % : [k] 表示网络中节点度的平均值。 } { 44 % : [d] 表示网络的平均最短距离，该值需要计算所有节点对之间的最短距离再求平均。 } { 41 % : H 表示网络节点度之间的异质性，若节点的度分布越不平均，该值则越大，计算公式为 }

[]

，分母部分先求所有节点度的平均值再平方，分子部分是先求各个节点度的平方再求平均值。

3.6 实验结果分析

3.6.1 实验环境

实验运行于配置为Intel(R) Core i7-4790K CPU @ 4.00GHz 4GB. 16GB RAM的电脑，采用Python环境。 第五章设计实验环境相同。

3.6.2 参数设定

为了保证试验的准确性，当我们确定某一个 p 值之后，我们会对测试集进行 100 次划分，划分之间彼此不关联，这主要是防止测试集的随机性对试验结果造成的误差。测试集划分好之后，我们将内层循环次数确定为 1000 次，也就是说总循环比较的次数为 100000 次，该次数对应于上述 AUC 指标中的 n。

3.6.3实验结果以及分析

(1)Elegans 网络中缺失边的发现与错误边的纠正，图3.4 (a) 为缺失边的发现，图3.4 (b) 为错误边的纠正。下图中共 18 种算法进行比较，其中原算法 9 种，加入聚类系数改进后的算法 9 种。为了便于比较，我们将某一算法和改进后的算法一一对应，这主要体现在绘图时线段的颜色，如下图所示 CN 和加入聚类系数后的 CCN 算法均由黑色线段表示，RA 和 CRA 均由红色表示。此外，原算法用虚线表示，改进后的算法用实线表示。{ 54 % : 结果图的横坐标表示所选取的训练集的比例 p ，纵坐标表示算法的预测准确度。 }

从 Elegans 的实验结果图可以看出改进后的算法相比较原算法在准确率上有所上升，该效果的上升不仅在于发现丢失边的情况同时适用于错误连边的纠正。{ 48 % : 随着测试集比例的上升，改进后的方法同样优于原算法。 }

详细数值见表3.2。

图 3.4 (a) Elegans 网络缺失边发现

图 3.4 (b) Elegans 网络错误边发现

表3.2 (a) Elegans网络缺失边发现

FRACTION

Algorithm

0.04

0.08

0.12

0.16

0.20

CN

0.84665

0.8477

0.8354

0.82515

0.81625

CCN

0.86415

0.85955

0.85345

0.8423

0.8333

RA

0.86395

0.85945

0.8536

0.8499

0.8323

CRA

0.8646

0.864

0.86

0.8483

0.83975

0.86 , 0.8483 , 0.83975

AA

0.86085

0.86335

0.8457

0.84675

0.8355

CAA

0.8677

0.8607

0.85185

0.8421

0.8405

HPI

0.8006

0.79595

0.7926

0.788

0.7827

CHPI

0.851

0.84405

0.83705

0.82295

0.82715

HDI

0.78065

0.7737

0.7648

0.7654

0.75685

CHDI

0.83135

0.83085

0.82

0.8083

0.8054

Salton

0.8018

0.80065

0.77645

0.7843

0.7789

CSalton

0.8497

0.8402

0.8303

0.8286

0.8143

Sorensen

0.79325

0.79015

0.7737

0.7708

0.77205

CSorensen

0.8465

0.836

0.82875

0.8188

0.8114

Jaccard

0.79205

0.78305

0.78595

0.7706

0.77365

CJaccard

0.83795

0.83055

0.8245

0.81645

0.80725

LHN

0.72015

0.7208

0.72165

0.71505

0.7215

CLHN

0.7831

0.78325

0.77545

0.77275

0.7716

表3.2 (b) Elegans网络错误边边发现

FRACTION

Algorithm

0.04

0.08

0.12

0.16

0.20

CN

0.8642

0.85485

0.8562

0.86345

0.8526

CCN

0.8832

0.8766

0.86885

0.87535

0.86745

RA

0.88305

0.8758

0.8731

0.8733

0.86095

CRA

0.88705

0.87785

0.8743

0.87055

0.86925

AA

0.88275

0.86655

0.87015

0.87185

0.86705

CAA

0.8868

0.87425

0.8708

0.87395

0.86925

HPI

0.83185

0.83655

0.83635

0.84105

0.839

CHPI

0.87435

0.8681

0.8625

0.86545

0.85135

HDI

0.7931

0.78955

0.7954

0.7937

0.7875

CHDI

0.8395

0.83935

0.82705

0.8268

0.82535

Salton

0.8262

0.82655

0.82905

0.82685

0.8277

CSalton

0.8704

0.8601

0.8553

0.85895

0.86185

Sorensen

0.81155

0.8061

0.8179

0.8138

0.81325

CSorensen

0.8544

0.847

0.84125

0.8433

0.8421

Jaccard

0.8053

0.81345

0.81055

0.81635

0.80845

CJaccard

0.85495

0.8477

0.8462

0.8502

0.8463

LHN

0.73555

0.7278

0.7269

0.7363

0.72795

CLHN

0.80005

0.80085

0.79845

0.7905

0.79565

(2)Football 网络的实验结果如图3.5所示。 实验结果表明改进后的算法在整体上明显优于原先的算法，此外在链路预测的过程中我们发现当 $p=0.08$ 时，预测的准确率达到最高值。

详细数值见表3.3。

图 3.5 (a) Football网络缺失边发现

图 3.5 (b) Football网络错误边发现

表3.3 (a) Football网络缺失边发现

FRACTION

Algorithm

0.04

0.08

0.12

0.16

0.20

CN

0.63255

0.65485

0.6433

0.6421

0.6498

CCN

0.65085

0.6914

0.6648

0.6528

0.6624

RA

0.63065

0.657

0.6389

0.6402

0.64245

CRA

0.6579

0.68145

0.66025

0.65685

0.6581

AA

0.6408

0.66025

0.64475

0.6347

0.65235

CAA

0.65125

0.6872

0.66515

0.65475

0.66975

HPI

0.58345

0.59665

0.6

0.5972

0.5921

CHPI

0.6295

0.63265

0.6341

0.63035

0.6279

HDI

0.5705

0.5839

0.5879

0.57245

0.58155

CHDI

0.6027

0.6342

0.61185

0.59915

0.61625

Salton

0.5862

0.5935

0.58865

0.5783

0.5912

CSalton

0.6186

0.6466

0.62555

0.6249

0.6274

Sorensen

0.5827

0.597

0.5829

0.58365

0.6023

C Sorensen

0.63015

0.6425

0.62735

0.6127

0.626

Jaccard

0.591

0.59725

0.5919

0.5806

0.584

C Jaccard

0.6137

0.6381

0.6205

0.6105

0.63105

LHN

0.4938

0.5079

0.5086

0.51855

0.5315

CLHN

0.53695

0.55415

0.54475

0.55145

0.56315

表3.3 (b) Football网络错误边边发现

FRACTION

Algorithm

0.04

0.08

0.12

0.16

0.20

CN

0.6883

0.67635

0.6805

0.6868

0.67995

CCN

0.7206

0.70935

0.7074

0.71075

0.7058

RA

0.6881

0.67255

0.6716

0.6879

0.67925

CRA

0.7223

0.70715

0.69455

0.69975

0.70505

AA

0.69

0.6846

0.68045

0.68865

0.68935

CAA

0.71935

0.69895

0.711

0.7109

0.6985

HPI

0.64355

0.625

0.6423

0.6438

0.64495

CHPI

0.70435

0.6753

0.69355

0.69195

0.6803

HDI

0.60725

0.6022

0.5971

0.6066

0.6039

CHDI

0.6552

0.64705

0.63065

0.6406

0.64795

Salton

0.62605

0.6221

0.62685

0.62905

0.62905

CSalton

0.6897

0.66865

0.6673

0.67195

0.67035

Sorensen

0.62865

0.6197

0.62135

0.618

0.6277

CSorensen

0.68075

0.66095

0.66875

0.65685

0.6657

Jaccard

0.61925

0.61965

0.6176

0.61415

0.61805

CJaccard

0.67515

0.65925

0.65715

0.65265

0.66

LHN

0.4914

0.48465

0.51305

0.5101

0.5045

CLHN

0.57395

0.5585

0.5712

0.55845

0.57815

(3)Jazz 网络的实验结果如图3.6所示。 改进后的算法只在少数几个点之上不能超越原算法，整体表现仍然十分优异。 整体上 CCN 和 CRA 表现最好，LHN 表现最差，由于在 Jazz 网络当中的整体的预测准确率较高， 所以除了 LHN 算法之外，其他算法的预测准确率差别并不大，但是依据聚类系数改进后的算法整体上超过了改进前的方法。

详细数值见表3.4。

图 3.6 (a) Jazz 网络缺失边发现

图 3.6 (b) Jazz 网络错误边发现

表3.4 (a) Jazz网络缺失边发现

FRACTION

Algorithm

0.04

0.08

0.12

0.16

0.20

CN

0.92655

0.9161

0.9137

0.91405

0.91585

CCN

0.9268

0.91955

0.9186

0.92165

0.91615

RA

0.9398

0.93175

0.9324

0.9326

0.92625

CRA

0.94235

0.9326

0.93305

0.9397

0.92955

AA

0.9299

0.92135

0.92835

0.9267

0.9192

CAA

0.9291

0.92835

0.9292

0.92495

0.9205

HPI

0.92025

0.9196

0.9172

0.9171

0.90945

CHPI

0.92475

0.9246

0.9226

0.9223

0.91725

HDI

0.9227

0.9213

0.9209

0.9168

0.912

CHDI

0.923

0.9194

0.9266

0.92025

0.92015

Salton

0.93345

0.93325

0.93115

0.93085

0.9295

CSalton

0.9401

0.9386

0.93205

0.93665

0.9345

Sorensen

0.9307

0.9298

0.92185

0.9248

0.92255

CSorensen

0.93515

0.9315

0.9298

0.91985

0.92325

Jaccard

0.92615

0.92845

0.9275

0.9238

0.9233

CJaccard

0.93295

0.9298

0.9277

0.92635

0.92685

LHN

0.87425

0.872

0.8727

0.86855

0.86955

CLHN

0.8861

0.8892

0.8949

0.8849

0.88065

表3.4 (b) Jazz网络错误边边发现

FRACTION

Algorithm

0.04

0.08

0.12

0.16

0.20

CN

0.9539

0.9483

0.95025

0.95355

0.947

CCN

0.95645

0.95215

0.94945

0.9504

0.9491

RA

0.9733

0.97115

0.96865

0.96825

0.96305

CRA

0.97935

0.9788

0.97515

0.97595

0.969

AA

0.962

0.9565

0.9554

0.95875

0.95605

CAA

0.9654

0.9642

0.96055

0.95785

0.9527

HPI

0.9619

0.9596

0.9619

0.9587

0.9625

CHPI

0.96185

0.96235

0.96095

0.95525

0.95835

HDI

0.9536

0.9503

0.95185

0.95205

0.95095

CHDI

0.96

0.95315

0.9532

0.9504

0.94835

Salton

0.9665

0.96775

0.9656

0.96365

0.96395

CSalton

0.9678

0.96615

0.96475

0.9636

0.9667

Sorensen

0.9676

0.96105

0.96345

0.96385

0.96065

CSorenson

0.9636

0.96025

0.9598

0.95635

0.9613

Jaccard

0.96575

0.9622

0.96385

0.95935

0.9634

CJaccard

0.9625

0.96255

0.95705

0.9614

0.9589

LHN

0.91715

0.91705

0.9152

0.92075

0.91955

CLHN

0.934

0.92445

0.9318

0.9243

0.9283

(4) USAir网络的实验结果如图3.7所示。该实验结果分层现象十分明显，改进后的算法曲线明显在改进前算法的曲线之上，这表明加入聚类系数这一节点的局部信息之后，{51%：该9种算法在美国航空网络的预测上效果明显提升。} 与上述数据集合类似，LHN指标表现较差。

详细数值见表3.5。

图3.7 (a) USAir网络缺失边发现

图3.7 (b) USAir网络错误边发现

表3.5 (a) USAir网络缺失边发现

FRACTION

Algorithm

0.04

0.08

0.12

0.16

0.20

CN

0.92645

0.9237

0.9277

0.92335

0.9175

CCN

0.93235

0.9336

0.93205

0.9251

0.921

RA

0.947

0.94685

0.9443

0.93615

0.9338

CRA

0.9489

0.95075

0.94825

0.9447

0.93785

AA

0.93895

0.9416

0.93145

0.9328

0.9318

CAA

0.93885

0.938

0.9343

0.9315

0.9259

HPI

0.869

0.87485

0.8645

0.86515

0.86175

CHPI

0.919

0.92205

0.9122

0.9067

0.90295

HDI

0.8852

0.8829

0.8826

0.87505

0.88075

CHDI

0.91245

0.90755

0.91195

0.9005

0.89765

Salton

0.90645

0.9071

0.90065

0.89395

0.89245

CSalton

0.92825

0.9298

0.9239

0.91895

0.90915

Sorensen

0.8876

0.89525

0.88925

0.8876

0.88395

CSorensen

0.91545

0.9147

0.9145

0.9059

0.90295

Jaccard

0.8957

0.89465

0.8878

0.885

0.8826

CJaccard

0.9147

0.9137

0.90935

0.9068

0.90305

LHN

0.7537

0.76525

0.7653

0.77205

0.7754

CLHN

0.8144

0.81955

0.82055

0.8256

0.8148

表3.5 (b) USAir网络错误边边发现

FRACTION

Algorithm

0.04

0.08

0.12

0.16

0.20

CN

0.941

0.94425

0.9428

0.9424

0.945

CCN

0.95285

0.95135

0.946

0.9454

0.9478

RA

0.95965

0.95935

0.9587

0.959

0.95805

CRA

0.9603

0.9623

0.9599

0.962

0.95115

AA

0.95365

0.955

0.95015

0.95105

0.955

CAA

0.95475

0.9539

0.95495

0.953

0.9507

HPI

0.9416

0.9403

0.9441

0.9487

0.9431

CHPI

0.9532

0.95205

0.9498

0.9503

0.95305

HDI

0.89885

0.8973

0.9011

0.89895

0.9012

CHDI

0.9267

0.92525

0.9238

0.91615

0.91595

Salton

0.9327

0.9354

0.92745

0.93125

0.93505

CSalton

0.9463

0.9466

0.9431

0.94595

0.9434

Sorensen

0.91415

0.91425

0.91705

0.9188

0.91545

CSorensen

0.93645

0.93735

0.92925

0.9291

0.9274

Jaccard

0.9139

0.91945

0.91775

0.9165

0.9171

CJaccard

0.936

0.9338

0.92685

0.9302

0.92815

LHN

0.78355

0.78535

0.77395

0.78865

0.78475

CLHN

0.8629

0.85675

0.84985

0.8539

0.8578

(5) Karate 网络的实验结果如图3.8所示。与上述四种网络结果共同验证了算法改进的有效性，此外我们发现这些链路预测算法对错误连边的发现要略优于预测不存在的边。

详细数值见表3.6。

图 3.8 (a) Karate 网络缺失边发现

图 3.8 (b) Karate 网络错误边发现

表3.6 (a) Karate 网络缺失边发现

FRACTION

Algorithm

0.04

0.08

0.12

0.16

0.20

CN

0.7026

0.67925

0.6643

0.6525

0.6536

CCN

0.72965

0.68785

0.6853

0.65845

0.65075

RA

0.76045

0.7356

0.72095

0.68895

0.6854

CRA

0.76655

0.7354

0.7372

0.69495

0.688

AA

0.7422

0.72325

0.70855

0.6782

0.6832

CAA

0.74625

0.7232

0.70785

0.69075

0.6928

HPI

0.7326

0.7037

0.68675

0.66715

0.65635

CHPI

0.7437

0.7144

0.696

0.6715

0.65295

HDI

0.60155

0.588

0.58605

0.5825

0.58915

CHDI

0.6364

0.60105

0.61245

0.6049

0.60355

Salton

0.6458

0.62455

0.6229

0.60625

0.60175

CSalton

0.70075

0.6542

0.6603

0.63435

0.61925

Sorensen

0.60995

0.60435

0.6081

0.59625

0.59875

CSorensen

0.66055

0.6248

0.6323

0.6177

0.60295

Jaccard

0.6238

0.59725

0.5899

0.5876

0.59195

CJaccard

0.65395

0.6222

0.61995

0.6041

0.59

LHN

0.60125

0.5982

0.60145

0.5923

0.5816

CLHN

0.637

0.6195

0.61825

0.6103

0.5846

表3.6 (b) Karate网络错误边边发现

FRACTION

Algorithm

0.04

0.08

0.12

0.16

0.20

CN

0.72165

0.68985

0.6956

0.70695

0.71195

CCN

0.7702

0.7311

0.73105

0.7416

0.72985

RA

0.7788

0.75095

0.7394

0.75405

0.7558

CRA

0.7662

0.75125

0.7425

0.75545

0.75795

AA

0.7608

0.72685

0.7337

0.73385

0.7385

CAA

0.77325

0.74775

0.73195

0.75065

0.7403

HPI

0.76085

0.69945

0.71835

0.73385

0.7388

CHPI

0.7879

0.7474

0.75615

0.7631

0.7567

HDI

0.6026

0.5615

0.5628

0.58425

0.5979

CHDI

0.68625

0.6504

0.6539

0.6615

0.6618

Salton

0.67045

0.6141

0.631

0.6499

0.64855

CSalton

0.7414

0.6957

0.70355

0.7045

0.7114

Sorensen

0.6291

0.5887

0.60405

0.6141

0.62915

CSorensen

0.71455

0.676

0.6808

0.6901

0.6852

Jaccard

0.6324

0.58715

0.594

0.6171

0.62565

CJaccard

0.7041

0.6571

0.6744

0.6824

0.6839

LHN

0.5987

0.5418

0.5527

0.57655

0.5798

CLHN

0.6829

0.6483

0.65485

0.66435

0.6623

本章小结

{ 58 % : 第二章所介绍的9种基于局部相似性的链路预测算法多基于共同邻居指标改进而成， } { 47 % : 这9种指标没有很好地区分不同的共同邻居节点之间的差异。 } 本章当中，我们引入了聚类系数对该9种经典的算法进行优化，加入聚类系数后的算法将让我们更好地区分不同的共同邻居之间的差异。 本章在5类复杂网络数据集之上进行了广泛的实验，实验比较了基于聚类系数改进前后两类算法，改进前后算法均有9种。 相比于常规的链路预测算法准确率的比较，我们比较了（1）不同错误信息比例下实验的效果； { 41 % : （2）同时比较了链路预测算法在缺失边以及错误连边的预测上的准确率。 } 实验均表明了基于聚类系数的改进思路对于这9种数据集的有效性。 实验同时说明，链路预测在较小的验证集比例下准确率较高，较大的网络规模也可提升算法的预测准确性。

{ 46 % : 第4章交通拥堵恢复序列预测问题分析 }

本章分析了交通拥堵恢复序列这一问题，提出了利用链路预测的方法对拥堵线路进行按序恢复。 为了验证交通拥堵恢复序列预测的有效性，我们介绍了一种广泛使用的流量模拟方法。

4.1 问题描述

{ 57 % : 定义 $G(V, E)$ 为一个交通运输网络，其中 V 表示网络中节点的集合， E 表示网络中边的集合。 } E 表示网络中发生拥堵的线路。 如图4.1所示，现有的交通网络包含节点集合 $\{1, 2, 3, 4, 5\}$ ，边集合 $\{[1, 5], [1, 3], [2, 5], [2, 4], [2, 3], [3, 5], [4, 5]\}$ ，现在网络当中 $\{[1, 3], [4, 5]\}$ 两条边发生了拥堵。 在现有疏通资源有限的前提下（如道路上只有一辆清障车），我们面对两种线路的恢复选择。 显然 $[1, 3]$ 和 $[4, 5]$ 的恢复顺序对于网络整体流量的改变是不同的，现在我们想找出这样的一个恢复顺序，该恢复顺序最有利于网络流量的提升。

对于该问题进行拓展，若三条边发生了损毁，可供我们选择的恢复顺序便有了 $3! = 6$ 种。

图 4.1 问题描述

4.2 解决思路

分析上述交通恢复序列的预测问题，我们发现与链路预测的定义有一定的相似性。 当交通线路发生拥堵之时，我们可将其简化为链路预测当中边丢失的情况。 根据链路预测的算法，我们可以对丢失边进行评分，根据评分结果我们将待预测的边进行排序， 分数较高的排序也靠前，链路预测方式将依次对这些边进行恢复。 所以，在交通拥堵恢复序列的预测中，我们可近似将拥堵边看作是缺失边，根据链路预测算法，我们对这些边进行评分，分数较高的我们将优先恢复。

虽然交通恢复序列的预测我们可将其抽象为链路预测问题，但是与链路预测问题有所不同的是， 如图4.1所示，我们虽然会将 $\{[1, 3], [4, 5]\}$ 这样的两条边看作是缺失边， 但由于信息掌握的差异，我们确信 $\{[1, 3], [4, 5]\}$ 这样的两条边是存在的， 因此对于未知边的评分，我们不会对不存在的边进行评分。 也就是说，我们不会对象 $\{[1, 2], [4, 3]\}$ 这样不存在的边进行预测，我们根据算法所计算的边仅是像 $\{[1, 3], [4, 5]\}$ 这样的

拥堵边。 另外一点的不同之处是，由于交通网络当中弱连接效应的存在，未必是连接着度较大节点的边对于提升网络运载能力发挥着更大的作用， 因此在之后的算法部分我们设计出了基于链路预测思路的反相关性算法。

在实验部分我们将比较反相关性恢复算法，随机恢复算法和链路预测恢复算法这三类方法的恢复效果。

4.3 流量模型

我们需要一种定量的方式来衡量网络的处理交通流量的能力，从而在恢复网络中一些损毁边之后能够比较不同恢复方案的优劣。 我们就用吞吐量来代表这个能力。

4.3.1 吞吐量定义

实验所利用的运载能力测量方法是参考文献[53]的思路，并结合实际测试数据后设计成的。 具体计算方法如下：

在交通网络当中我们以数据包来模拟交通的运载。 每个时间节点，网络中每个节点产生 R/N 个数据包，这里 N 是网络的规模， R 是流量的产生速率，即 R 是每一个时间点网络中总共产生的流量数。 这些数据包被随机分配终点，同时，我们给每一个节点设置一个传输容量 C_i ，这表明在每个时间节点内根据路径通过点 i 到达下一个节点的数据包数量最多为 C_i 。
{ 50 % : 当节点不能通过时会排在队列中，按照先进先出的原则处理。 } { 43 % : 因此，数据包的路径选择在交通过程中也是很重要的。 } { 41 % : 我们知道，在互联网中域内的路由选择使用的是最短路由方法。 } 因此，我们这里在计算吞吐量时也让数据包沿着最短路路径前进。 { 49 % : 当一个数据包到达了终点，那么就将它从系统中删除。 }

为了分析完全自由的状态到出现阻塞这一过程，我们使用一个参数(R)来表示上面这段描述。

4.3.2 衡量指标

参数(R)的定义如下公式所示，

$$() =$$

$$\lim$$

$$[]$$

(4.1)

在这里 $W(t)$ 代表在时间点 t 时网络中的数据包数量， $W=W(t+t)-W(t)$ ， $[...]$ 代表长度为 t 的时间窗口内的平均值。 换句话说，参数代表现存的交通流与经过足够长时间累积的注入的交通流的比值，或者说网络中残存的数据包比例。 很明显，在通常的交通网络中，即网络中没有阻塞，系统是可以处理新产生的车辆的，因此参数接近 0。 相反，在阻塞的情况下，产生的数据包太多导致无法运输，那么存在与网络中的数据包也会很多，使得接近 1。

图4.2是我们从某次实验中记录下来的模拟过程中的变化情况。 我们看到一开始值非常的高，这是因为最开始生成的数据包都没有到达终点，所有的车都存在于网络之中， 而当时间片走到40左右，值迅速降低并不断向

0.07靠近，这说明当改网络稳定下来的时候，该网络的值大约为0.07左右，即网络中会残存7%的数据无法处理。

图4.2值变化($R=100$, $S=1.8$, $N=20$)

随着 R 值得增长，交通网络的压力会越来越大，值也会从一直是趋近于 0 的状态变为非 0 并随着 R 的增长而增长。因此，我们用值从 0 变到非 0 的临界 R 值来代表网络的吞吐量，记做 R_c 。也就是说， R_c 代表该网络能完全地、不留残余的处理的最大的数据包的数量， R_c 越大代表交通网络的效率越高。

这里存在一个问题， R 值与值得关系看似的单调的，也就是直觉上我们认为 R 越大，那么值也就越大，但实际上临近 R_c 附近的 R 值所对应的值一直在 0 附近波动，再加上我们的实验有一定随机性，也就是说加入 R_1 [R_2 ， R_1 所对应的 1 并不一定小于 R_2 对应的 2]。这就造成了我们在寻找 R_c 时，只能从 0 一点一点增加，而不能使用二分查找，这就大大增加了实验需要的时间。

4.4 交通拥堵恢复序列预测分析模型

根据不同的道路拥堵序列的恢复算法，相同的拥堵状况将得到不同的恢复序列。本节将介绍流量模拟实验，根据模拟结果我们将能够判断不同预测算法的优劣。本章实验部分的开发语言为 python，并主要基于复杂网络相关的包 networkx 进行实验。

{ 54 % : 如图4.3所示，实验部分主要分为5个大的步骤。 }

图4.3实验整体思路

第一步，根据不同的数据集，我们将生成相应的复杂网络结构，该网络结构也包含节点以及边信息。对于美国航空网络以及加州路网数据，我们所获取的原始数据均为 txt 文本格式，其中每行均包括两列数据，数据内容为 (node1 ID , node2 ID)，该数据表示的是一条边信息，该边连接了 node1 和 node2。依次遍历每行数据，并将该条边添加到现有网络之上，直至文件最后一行。这样根据不同的实验需求，我们就可以构建相应的美国航空网络以及加州路网。关于 BA 网络，networkx 包当中已经内置了生成接口，只需要传递相应参数即可。本次实验生成的 3 类网络均是无权无向网络。

第二步，便是模拟交通拥堵的情况。本次研究过程通过移除网络当中相应边表示对应线路的拥堵情况，也就是说认为该线路拥堵的已经近似于不通了。{ 40 % : 对于边的移除方法，我们采取随机抽取的方式，也就是说认为每条线路发生拥堵是等概率事件。 }

第三步，根据 8 中不同的恢复方案对拥堵边进行恢复序列的预测。也就是说，如果网络当中有 5 条边发生拥堵，根据恢复算法，我们将对这 5 条边进行打分，分数高的优先进行恢复。

第四步，进行流量模拟实验，该步骤也是最为繁琐最为耗时的。该步骤是由多个时间片所构成的。单个时间片内处理过程如图 4.4 所示。对于网络当中的每个节点，我们对其维护一个队列，初始为空。网络当中的每个节点在同一个时间片内主要做两件事情，首先按概率产生一定数量的数据包并随机分配目的地节点，{ 65 % : 并按照迪杰斯特拉算法生成最短路径； } 其次运输数据包，若本时间片内经过该节点的数据包未处理完成则堆积在该节点的队列之中（本次设置每个节点单位处理能力为 1）。单位时间内，数据包运行距离为 1，即只能由一个节点运往其邻居节点，当数据包到达目的地节点时，丢弃数据包。每当流量模拟过程经过一定的时间片（本次实验

设置为1000），向网络当中增加一条边，此处边的增加顺序是由第三步所得到的，此步骤主要模拟每条道路需要的修复时间。 { 53 %：重复上述过程，直至添加完成所有的待修复的边。 }

{ 54 %：图4.4每个时间片内对数据包的处理流程 }

第五步，根据第四章第三节提到的方法，统计网络的拥塞率。当最后一条边修复完成之时，统计网络的拥塞率，据此比较各算法在交通拥堵序列预测方面的效果。

{ 66 %：流量模拟实验的大致流程如图4.5所示。 }

4.5 流量模拟实验

{ 48 %：不同的交通恢复序列算法下的流量模拟伪代码如图4.6所示。 } 我们首先创建网络 $G(V, E)$ ，若该网络为加州路网，由于网络包含了数百万条边，流量模拟实验耗时巨大，我们需要对该网络进行部分的截取，选取网络中部分区域信息。为了比较{ random , PA , RPA , RAPA , LP , RLP , SA , RSA}这8种交通拥堵预测算法的效果，我们将对实验进行1000次的循环，在某次循环中若某种恢复算法表现最优，我们对其进行加1处理。最后，我们将统计各个算法表现最优的次数。

4.6 评价方式

加上随机的序列恢复方式，我们共有8中交通恢复序列的预测方式，不同的恢复方案对于网络流量的改变有着不同的影响，其中一些方案会优于其他方案。首先我们会初始化一个大小为8的数组

`bestPerform=[0, 0, 0, 0, 0, 0, 0, 0]`，数组中的每个元素对应一种算法的表现效果，数组中元素所对应的算法分别为[Random , PA , RPA , RAPA , LP , RLP , SA , RSA]。在某一轮模拟当中，我们会模拟道路的损毁情况，并根据不同的恢复算法得到8中不同的恢复序列，根据不同的算法按序恢复交通线路，当我们完成道路的恢复之后，分别计算8种恢复算法下网络的交通拥堵率[R]。如果某个算法得到的拥堵率最低，那么久表示该算法在交通恢复方面的表现最为优异，我们便将在`bestPerform`中算法对应的元素+1。经过n轮之后，若某算法在`bestPerform`中对应值最大，那么久表示该算法在交通恢复序列的预测方面表现最好。

图4.5实验程序流程图

/图4.6 不同的交通恢复序列预测算法下的流量模拟

4.7 本章小结

本章描述了交通网络拥堵恢复序列预测问题。为了比较交通拥堵恢复算法表现的优劣，我们给出了一种经典的流量模拟方式，该方式以网络中的拥塞率来给出一个量化的比较标准。

第5章交通拥堵恢复序列预测算法

针对交通拥堵恢复序列的预测问题，我们提出了三类恢复性算法，这三类算法分别是链路预测算法、基于链路预测的反相关性算法，随机恢复算法。通过在美国航空网络、加州路网以及BA网络上的流量模拟实验，我们得出了基于链路预测的负相关性算法在恢复拥堵交通上的高效性。

5.1交通拥堵恢复序列预测算法

如图5.1交通拥堵恢复序列预测算法的流程图所示，对于交通网络当中的拥堵边（模拟实验通过随机产生），我们根据交通拥堵序列的预测性算法对这些边的重要性进行评分，根据得分的高低，我们对这些拥堵的边依次进行恢复。不同的交通拥堵恢复序列的预测指标将得到不同的拥堵边恢复顺序。本文将对基于链路预测指标的恢复方法、随机恢复方法、基于链路预测指标的反相关性恢复方法进行比较。其中随机恢复方案指的是对于 $\{e_1, e_2, e_3, \dots, e_n\}$ 这样的拥堵序列进行随机的重新排列，并按照排列后新的顺序进行恢复。若交通拥堵恢复序列预测算法有效，至少要保证恢复效果优于随机的方法。

图5.1交通拥堵恢复序列预测算法

5.2 链路预测指标

链路预测指标可以根据网络的拓扑信息对网络当中的缺失边进行判断，并查找出相应的缺失边。以下将给出3种链路预测指标，其中两种是经典的预测指标一种是本文提出的预测指标。

1. 基于偏好连接指标的链路预测方法

Preferential Attachment Index (PA): { 43 % : 对于网络中暂未连接的两个节点，这两个节点连接的可能性正比于该两个节点度的乘积。 } 偏好连接机制也适用于无标度网络的产生机制，在无标度网络的产生机制中，新增加的边的一个端点正比于该节点的度的大小。该机制也同样适用于非增长的网络，在这样的网络生成当中，我们每次先删除掉一条边，然后再添加一条边， { 58 % : 新增加的边连接节点 v_x 和 v_y 的可能性正比于 v_x 和 v_y 度的乘积。 }

$$S_{xy} = k_x * k_y \quad (5.1)$$

2. 基于本地路径指标的链路预测方法

Local Path Index (LP): { 59 % : 局部路径算法是在共同邻居的基础上考虑三阶邻居对于链路形成的影响。 } 定义如下式所示，其中参数用于调控三阶邻居影响的大小。 { 62 % : 显然，当 = 0 时，LP 算法就退化为了 CN 指标，即只考虑二阶邻居的影响。 } 下式中 A 表示为网络的邻接矩阵。 { 66 % : $(A^n)_{xy}$ 表示节点 v_x 和 v_y 之间路径长度为 n 的线路数目， } { 46 % : A^2 矩阵中第 x 行第 y 列的元素表示的是节点 v_x 到节点 v_y 之间长度为 2 的路径数目， } 即 v_x 和 v_y 的共同邻居数目。在 CN 算法中由于只简单的考虑两个节点的共同邻居数目，所以会导致相似性的分数过度集中（绝大多数的分数集中在 0, 1, 2, 3），这也就降低了算法的预测准确性。 { 43 % : LP 算法便解决了 CN 算法的聚集现象，提升了链路预测的准确性[28]。 }

$$S = A^2 + A^3 \quad (5.2)$$

{ 66 % : 3. 基于三阶相似性指标的链路预测方法 }

Three order Similarity Index (SA): 相比于前文所提到的链路算法，该算法是在本文当中首次提出的。 { 49 % : $3(x)$ 表示的是节点 x 的一阶邻居、二阶邻居、三阶邻居的合集， } { 41 % : S_{xy} 表示节点 x 和节点 y 三阶以内邻居的交集，该算法进一步的拓展了 CN 算法的差异性， } { 55 % : 力求进一步极高链路预测的准确性。 }

$$S_{xy} = |3(x) 3(y)| \quad (5.3)$$

5.3 基于链路预测的反相关性指标

以下的四个算法是在上文的三个算法的基础上进行的转换，我们称作反相关性恢复算法。与链路预测问题有所不同，交通拥堵恢复序列的预测是对于已知的拥堵边进行排序，这些边在网络当中是肯定存在的，然而在链路预测中是对于那些可能存在的边进行预测。同时，当交通网络当中连接较大度的边发生拥堵之时，流量可分流至邻接边。然而，若是连接较小度的边发生拥堵，网络的连通度必然降低，周边流量也难以得到分流，所以这样的拥堵边长期存在的话会更大程度的提高网络的拥塞率，所以交通网络当中的边的重要性评价方式与链路预测也有所不同。由于本质上负相关性算法优先恢复连接较小度的边，所以恢复过程中可更大概率的降低网络的拥塞率，这也可能是网络当中另外一种形式的弱连接效应。

$$\{ 58 \% : \text{基于链路预测的反偏好连接方法} \}$$

Reverse Preferential Attachment Index (RPA) 算法简单的取 PA 算法的结果的倒数，该算法认为在交通网络中度较小的节点对于交通网络的运输起着更大的作用，这也就是复杂网络当中所提到的弱连接效应。

$$S_{xy} =$$

$$(5.4)$$

$$\{ 52 \% : \text{基于链路预测的反相加偏好连接方法} \}$$

Reverse Add Preferential Attachment Index (RAPA): 与 RPA 类似，区别点仅在于将分母部分的两节点度相乘变为度的相加。该一变化的目的在于探索怎样的算法在进行交通拥堵序列的恢复中能起到更好的作用。

$$S_{xy} =$$

$$(5.5)$$

基于链路预测的反本地路径算法方法

Reverse Local Path Index (RLP): 该算法简单的将 LP 算法计算的分值进行求导，与链路预测算法中的区别在于承认运输网络当中的弱连接效应。

$$S =$$

$$A_2 + A_3$$

$$(5.6)$$

$$\{ 63 \% : \text{基于链路预测的三阶相似性方法} \}$$

Reverse Three order Similarity Index(RSA): 与 SA 算法向对应，将 SA 算法所计算的分值求导。

\$xy =

|3(x) 3(y)|

(5.7)

5.2 数据集描述

5.4.1 数据集描述

本章实验部分共有三类数据集，这三类数据集分别是BA网络（无标度网络），USAir（美国航空网络）和California Roads（加州路网）数据。三类网络中BA网络属于模拟网络，另外两种属于真实网络，三类网络均可表示交通运输网络。
{ 50%：现对这三类数据做简要的介绍。 }

5.4.1.1 BA网络模型

BA模型[57]是一种利用偏好链接机制的随机生成具有无标度特性的网络的算法，其中偏好链接的含义是：在进行某项资源的分配时，往往已经拥有很多资源的个体会有比那些比较贫穷的个体拥有更高的概率分配到资源。
{ 60%：也就是俗话所说的富者愈富，穷者愈穷。 } 它由Albert- Lszl Barabsi 和Rka Albert提出，因此被称为BA模型。
{ 55%：根据BA模型生成的无标度网络的度分布遵循以下形式的幂法则： }

$p(k) \propto k^{-3}$ (5.8)

{ 46%：本实验中利用 BA 模型生成的无标度网络作为一些实验数据。 } 生成网络的算法主要依据以下公式：

(5.9)

{ 53%：公式中 p_i 代表每新增一个节点，新节点与点 i 的相连的概率， k_i 代表点 i 的度数， } 也就是说 p_i 随着点 i 的度数所占比例增长而增长，也就是所谓的富者越富。

图5.1 BA网络样例

图5.1是我们利用 Matlab生成的BA网络的样例，如下图所示，我们可以看到在在图中的右上方，节点的边的密度明显要比其他部分高出许多，这是由于我们在画这幅图的时候按照逆时针的顺序从右上方方式画点，先画出的点本着富者更富的原则从而聚集了愈来愈多的边。我们生成这幅BA网络所用的参数为：总共300个点，每个点与之前的两个点相连接，因为每条边提供的度数为2，每个新点提供的度数为4，也就是对于最终的网络来说平均每个点的度数为4。

5.4.1.2 USAir网络

我们从pajek公开的复杂网络数据库当中获得美国航空网络数据，获取的网址为http://vlado.fmf.uni-lj.si/pub/networks/data/。该数据包含了美国的机场及相关航线信息。
{ 45%：网络当中包含节点数目为332个边 }

2126条，节点表示机场边表示机场之间的航线。} 值得注意的一点是该网络是一个向网络，当你可以从机场A买到直达机场B的机票之时，你也可以买到机场B到机场A的机票。在某种程度上来说，美国航空网络也是BA网络的一种形式，某些机场如洛杉矶机场、纽约机场相比其他机场拥有更多航线，这也是一类聚集现象，少数节点占有更多的资源。

5.4.1.3 California Roads网络

我们从stanford公开的复杂网络数据库当中获得了美国加州路网数据集，获取的网址为http://snap.stanford.edu/data/。加州路网包含了1,965,206个节点和2,766,607条边。节点表示路口，边表示道路。该网络同样为一个无向网络，当你可以从A节点到达B节点之时，相反的路径同样可达。考虑到该网络对于链路预测以及之后的流量模拟过程来说过于巨大，现我们需要对其进行处理。首先，我们从网络当中任意取一个节点作为中心节点。然后，我们以该节点为中心，按照网络的层次遍历算法朝外进行扩张。接着，根据我们对于网络尺寸的需求，当所遍历到的节点数目满足我们的需求之时，停止遍历。最后，根据遍历结果生成子图网络。

{ 41%：例如，我们以ID为100的节点作为中心节点，我们依次遍历100的邻居节点并将它们加入队列之中。} 若队列非空，出队队首元素，并加入该节点的邻居（若节点已经被访问，忽略该节点）直到访问节点数目满足我们的需求。在美国加州路网当中，如果我们选择ID = 800120D的节点作为中心节点，层次遍历的层数为30，我们可以获得一个包含1285个节点，1748条边的网络。

5.4.1.4 数据截图

//

(a) 加州路网 (b) 美国航空网络

图5.2 加州路网及美国航空网络原始数据

图5.2 (a) 是原始的路网美国加州路网数据，5.2 (b) 是美国航空网络数据，其格式均为纯文本格式。左侧数据第一列表示行号，2、3两列表示节点对所组成的边信息，第4列表示路径的长度。{ 45%：右侧数据相对讲点，每行的节点对表示网络中的一条边信息。}

5.4.2 数据集的网络统计特征

本节主要介绍的是三种网络当中的一些统计指标，这些统计指标如表5.1所示。

BA256表示BA网络，节点数目为256。BA512表示节点数目为512的BA网络，BA1024表示节点数目为1024的BA网络。{ 41%：BA网络当中新增加的节点均产生两条连边。} USAir即为美国航空网络。

表 5.1 数据集的网络统计特征

网络

点|V|

边|E|

密度D

聚类系数C

平均度[k]

平均距离[d]

度的异质性H

BA256

256

508

0.016

0.075

3.97

3.49

2.17

BA512

512

1020

0.008

0.052

3.98

3.68

2.87

BA1024

1024

2044

0.004

0.030

3.99

4.07

2.78

USAir

332

2126

0.0386

0.6252

12.8072

2.7381

3.4638

Cal Road 10

121

156

0.021

0.142

2.58

9.54

1.13

Cal Road 20

409

525

0.006

0.101

2.57

17.71

1.17

Cal Road 30

1285

1750

0.002

0.077

2.72

23.28

1.16

Cal Road 10表示以中心节点朝外遍历10层得到的网络。 Cal Road 20表示以中心节点朝外遍历20层，相应的Cal Road 30表示扩散层数为30。

{ 43 % : 如第三章所定义， $|V|$ 表示网络当中节点的数目， $|E|$ 表示网络当中边的数目。 } { 43 % : D表示网络的密度，C表示网络的聚类系数， $[k]$ 表示网络度的平均值， $[d]$ 表示网络当中两个节点间的平均距离， } { 48 % : H表示网络当中度的异质性，度的分布越是不平均，H值越大。 }

5.5 实验结果分析

以下是实验的结果部分，实验在美国航空网络、加州路网、BA三类网络之上进行。

5.5.1 加州路网上的结果分析

图 5.3 是我们在加州路网之上进行的实验。加州路网数据包含 1,965,206 个节点和 2,766,607 条边信息，由于网络的模拟过程耗时巨大，所以我们选取了该网络的部分区域进行实验。按照本章第 2 节所描述的处理方式，我们得到了一个包含 223 个节点和 291 条边的网络。横坐标表示 8 种不同的恢复算法，纵坐标表示对应算法在多轮实验当中表现最好的次数。

{ 43 % : 我们设置网络当中每个节点在单位时间内处理数据包的能力为 1 , 恢复单个道路需要 1000 个时间片。 } 每个时间片内节点按概率产生数据包，原点和终点均随机产生，游走路径为迪杰斯特拉最短路径。该设置同样适用于以下另外的三个实验。

图 5.3 加州路网上的实验结果

表 5.2 加州路网上的实验数据

算法

PA

RPA

RAPA

LP

RLP

SA

RSA

random

表现最好的次数

1152

1481

1330

846

1270

1091

1591

1239

相比于美国航空网络和BA网络，路网数据的度分布更为平均，大部分的度分布集中在[1, 2, 3, 4, 5]这几个数值之中，所以该网络通过网络的拓扑信息更加难以区分链路的重要性。但是，从10000轮的模拟之中，我们的实验依旧表明了RPA, RAPA, RLP, RSA算法表现优于随机恢复方式，随机恢复方式优于PA, LP, SA算法。这也就是说，反相关性算法要优于随机算法，随机算法优于链路预测算法。从图1当中，我们发现本文提出的基于SA的反相关性算法表现最为优异，这可能是在低的度异质性网络当中，我们利用了更多的邻居信息的，这样扩大了节点之间的差异性。详细数值见表5.2加州路网上的实验数据。

5.5.2 不同网络损毁比例对于实验的影响

图 5.4 BA网络上不同损毁比例对实验结果的影响

表 5.3 网络上不同损毁比例对实验结果的影响

算法

损毁比例

random

PA

RPA

RAPA

LP

RLP

SA

RSA

0.04

110

189

230

43

148

275

0.08

81

218

252

120

322

0.12

50

202

252

90

406

0.16

21

226

267

90

396

0.20

21

196

258

88

437

0.24

21

261

274

29

415

0.28

13

235

252

38

462

0.32

189

280

11

511

0.36

196

233

563

0.40

182

233

584

图5.4是我们在BA网络之上进行的实验。该网络包含256个节点和508条边。横坐标表示线路的拥堵比例，例如0.04表示有 $\text{int}(0.04 \times 508)$ 条边发生了拥堵，我们需要从这么多的边当中找到一条有序的恢复序列。Y坐标表示在1000轮实验当中某种算法表现最好的次数。另外，参数R = 10。

显示当中的交通网络可以近似看作是BA网络。从图5.9我们可以看出随着道路拥堵比例的提升，负相关性的算法表现越来越好，当损毁的比例达到0.4之时，在1000轮的实验当中负相关性类别的算法几乎总是能表现最好，其中本文提出的RSA表现最好。这主要是由于随着损毁边数目的增加，恢复序列的可能性成着指数型的增长，负相关性之外的算法很难蒙到更好的恢复序列。另外一个因素是，随着损毁比例的增加，道路恢复的时长也进一步增加，这就进一步的扩大了实验结果的差异性。详细数值见表 5.3网络上不同损毁比例对实验结果的影响。

5.5.3不同网络尺寸对于实验的影响

图 5.5不同尺寸的BA网络对实验结果的影响

表 5.4 不同尺寸的BA网络对实验结果的影响

算法

Accuracy

random

PA

RPA

RAPA

LP

RLP

SA

RSA

52

83

54

10

11

28

84

30

30

15

20

20

33

60

15

25

20

20

18

10

30

62

17

20

20

23

25

30

16

24

61

17

17

30

30

31

18

24

59

15

18

17

18

35

28

10

24

54

16

29

30

40

24

12

17

55

21

17

17

37

45

19

19

23

58

18

18

37

50

28

15

28

52

23

15

33

图5.5与5.4均是在BA网络之上进行的实验。为了验证不同的网络尺寸对于实验结果的影响。 { 42 % : 我们生成了节点数目分别为100 , 200 , 300 , 400 , 500的网络 , 这五种不同的网络的度的平均值均为2。 } { 44 % : 我们设置每个节点单位时间内产生数据包的概率为0.02 , 线路损毁的比例为0.05。 }

观察图5.10我们发现当网络的尺寸为100之时 , 各种算法的表现差异不大。这是因为较小的BA网络当中 , 各个节点的差异性并不大 , 所以算法很难区分不同边的重要性。然而 , 在BA网络当中 , 当网络规模达到一定程度之后 , 各个节点的差异性就显现出来 , 各个节点的度分布差异也将越来越大。 { 41 % : 随着网络尺寸的增

加，负相关性算法的优势便体现了出来。} 当网络尺寸达到200之时，我们的负相关性算法便完全优于其他方法，随着网络尺寸的增加，该优势进一步明显。对此现象，主要是因为随着网络尺寸的增加，BA网络的异质性进一步增加，各个节点的差异性进一步明显，所以算法可对这些节点作用进行更为精准的区分。此外，同实验5.5.2随着网络尺寸的增加，在一定拥塞比例的前提下，拥堵边的数目进一步增加，恢复序列的选择可能性进一步加大，这也是扩大实验区分度的一大因素。当然，损毁边的增加也会恢复的时间，该因素同样扩大着实验的差异性。

在现实的交通网络当中，网络节点数目通常不会低于200，所以负相关性算法可以在交通恢复序列的预测之中起到较好的作用，尤其是RSA算法。{ 43 %：详细数值见表 5.4 不同尺寸的BA网络对实验结果的影响。}

5.5.4不同大小的流量对于实验的影响

图 5.6 美国航空网络上不同大小的流量对于实验结果的影响

图 5.5 美国航空网络上不同大小的流量对于实验结果的影响

算法

Accuracy

random

PA

RPA

RAPA

LP

RLP

SA

RSA

100

125

125

120

140

120

100

150

120

200

108

37

191

239

34

111

19

261

300

101

14

191

229

18

137

11

294

400

128

11

211

214

12

143

272

500

83

11

221

218

12

97

11

347

{ 42 % : 图 5.6 是在美国航空网络之上进行的实验 , 美国航空网络包含 332 个节点 , 2126 条边。 } 该实验为了验证不同的流量产生大小对于实验的影响。 { 43 % : 横坐标 R 表示流量的产生速率 , 纵坐标表示 1000 此试验中某种算法表现最优的次数。 } 此试验中 , 我们设定网络的损毁比例为 0.05 。 从实验当中我们发现 , 在 R 值较小的情况下 , 负相关性算法表现明显优于一般的链路预测算法以及随机恢复算法。 随着 R 的增加 , 三类算法的差异性在缩小 , 但是负相关性类别的算法 RASA 依旧表现最为优异。 这可能是因为在网络流量较小时 , 恢复部分拥堵线路可很大程度的降低网络的拥塞率 , 然而 , 随着流量的增加 , 网络的拥堵现象日益严重 , 部分道路的疏导也未必能在全局上很好的缓解交通的压力。 尽管差异在减小 , 实验仍然证明了在不同的网络流量下 , 负相关性算法均优于随机算法。 详细数值见图 5.5 美国航空网络上不同大小的流量对于实验结果的影响。

5.6 本章小结

本章给出了3类交通拥堵恢复序列预测性算法，这三类散发分别是链路预测性算法，{51%：基于链路预测的反相关性算法，随机恢复算法。}

实验结果表明，根据链路预测思路变换得到的负相关性算法在拥堵序列恢复的预测之上明显优于随机的恢复方式，随机的恢复方式又优于原始的链路预测方式，其中在BA网络和加州路网之上RSA算法表现最为优异，在美国航空网络之上RAPA算法要优于其他算法。负相关性算法优于其他算法的原因可能是，当网络当中连接较大幅度的边发生拥堵之时，流量可分流至邻接边。然而，若是连接较小度的边发生拥堵，网络的连通度必然降低，周边流量也难以得到分流，所以这样的拥堵边长期存在的话会更大程度的提高网络的拥塞率。由于本质上负相关性算法优先恢复连接较小度的边，所以恢复过程中可更大概率的降低网络的拥塞率，这也可能是网络当中另外一种形式的弱连接效应。

第6章总结与展望

6.1 总结

{95%：世界上存在着形形色色的网络，合著网络、在线交友网络、生物网络、交通网络等等。}{100%：随着对这些对人类而言密不可分网络的越来越多关注，网络连接挖掘中链路预测研究的也得到迅猛发展。}{92%：本文旨在基于复杂网络结构相似性的链路预测算法的研究。}{96%：局部相似性指标都是在CN算法的基础上拓展的，本文也从一种角度上对CN类算法进行了改进，}{71%：对实际的无权网络进行预测表明改进思路的可行性与高精度。}{71%：随着我国经济的高速发展和城市化进程的加快，我国城市的交通出行需求日益增加。}{86%：尤其是在大城市，交通拥堵以及由此导致的交通事故的增加。}{85%：日益严重的交通问题，严重影响了城市的经济建设和运行效率，也给人们的工作和生活带来了种种不便与损害，交通拥堵问题已经成为制约城市可持续发展的一大瓶颈。}受到链路预测的启发，当交通网络当中出现多条线路的拥堵之时，我们根据链路预测算法对这些线路进行评分，按得分的高低依次恢复这些路段。

{56%：本文首先是提出了基于聚类系数的链路预测算法。}其次，运用链路预测的结构相似性思路，提出了交通拥堵恢复序列的预测方法。总的来说，本文主要的研究及成果包括：

1) 通过对现有9种局部相似性的链路预测算法分析，我们发现该类算法均基于节点间的共同邻居来对节点间的重要性进行评价，但该类方法忽略了共同邻居间的差异。通过分析我们发现聚类系数可以很好的表示节点间的差异性，我们将聚类系数这一网络指标加入9种经典的链路预测算法之中，得到了9种新的算法。在5种真实的网络数据集上的实验表明，该改进思路可在这5种网络之上较好的提升链路预测的准确性。先前关于链路预测的实验主要关注缺失边的恢复，本文的实验还考虑了算法对于错误连边纠正的情形，在该情形之下，改进后的算法表现依然优于原先的算法。除此之外，先前关于链路预测的实验大多将验证集的比例固定在10%这一数值之，本文中的实验部分将验证集和的取值范围扩大到了[0.044, 0.08, 0.12, 0.16, 0.20]这一区间，扩展的实验进一步表明了改进算法的健壮性。{73%：所以，总的来说，本文在讨论复杂网络结构相似性的基础上，根据节点的相似性发展了基于聚类系数的链路预测算法。}{86%：它们把局部的算法往更高阶的路径上拓展了一步，与全局算法相比，大大减少了运算量；}{91%：与局域算法相比，预测精度又有了一定程度的提高。}

2) 在交通拥堵恢复序列预测的研究中，本文提出了基于链路预测的方法来按序恢复拥堵的交通线路。这一点与先前的交通拥堵研究有着很大的不同，先前的研究都是力求提升网络的整体运载能力，而非在线路发生拥堵的

之后提出解决方案。其次，在两类真实网络和一类模拟网络之上，我们对于链路预测恢复算法，随机恢复方法和基于链路预测思路的反相关性算法进行了流量模拟，实验表明了反相关性算法的在交通拥堵恢复序列的预测方面的有效性，这也将链路预测的运用范围进行了进一步的拓展。

6.2 进一步工作

{ 43 % : 本文提出了基于聚类系数的链路预测算法以及利用链路预测的方式进行拥堵交通恢复序列的预测方法。 } 进一步研究工作可考虑：

1) 在链路预测算法方面，在后续的研究中我们可分析其他的网络特征（节点间距离、边介数等）对于链路预测的影响，{ 41 % : 并将该类特征融入到算法之中来进行算法准确率的提升。 }

2) 在交通拥堵恢复序列方面，本文尝试性的运用链路预测的思维解决恢复序列的预测问题，为了集中研究点，故将一些实际交通情况略去了。由于缺乏实际的交通数据，所以本实验是在真实数据的基础之上做的模拟实验，所以本实验是半模拟的过程，未来的研究当中可结合更多的真实流量数据进行实验的模拟，这样也可更大的增加算法的说服力。{ 47 % : 此外，本文的实验数据是基于无向无权的网络，有向含权网络有待进一步的验证。 } 本文提出了一种基于链路思维的交通拥堵恢复序列的预测，在此基础之上更多与真实场景相关的因素待进一步考虑。致谢

{ 86 % : 转眼间，三年的研究生生涯即将结束。 } { 49 % : 借此论文写作之机，我要感谢所有帮助、支持、信任过我的家人、老师、同学、朋友和同事。 } 你们的存在造就了现在的我，千万句的言语也不足以表示我对你们的感激，你们是我生命中最为重要的人。

{ 63 % : 研究生期间，我首先要感谢的是我的导师关信红教授。 } 在学术上关老师严谨求实，她丰富的学术经验指引着我的研究方向。在关老师的悉心教导之下，我初步掌握了做学术研究的基本技能，仅此一点而言可以说我的三年研究生生涯不虚此行。除了学术的指引，我想关老师对我最大的影响就是她的为人。关老师是位极其和蔼、宽容、谦逊的人民教师，学生们在您的教导下如沐春风。永远向您学习！

感谢复旦大学周水庚教授。虽然和周老师接触的机会并不多，但就是这几次少有的接触让我了解到了人的成功不是偶然而是必然这样的道理。周老师对于学术的热情与专研确确实实的震撼着我，您的学术精神时时刻刻在感染着我，向您致敬！

感谢同济大学青年教师张毅超老师。张毅超老师对我进行了非常具体的指导，可以说复杂网络方向就是张老师引我入门的。张老师带领我参加了多项学术会议，开阔了我的眼界；推荐给我很多优秀论文，让我在学术方向有目标性的前进；指导了我论文中的多项算法；并且，张老师对于我的小论文还进行了手把手的修改。我从张老师身上看到了一名优秀教师的品质，认真、踏实、勤奋、亲和，感谢张老师对我的指导！

感谢复杂网络讨论小组的各位同学，我们的复杂网络的讨论班已经开展了有两年的时间了。在这两年的时间当中，我从你们的报告当中获益良多，我不仅学习到了复杂网络相关的知识与技能，也从你们身上学习到了许多论文阅读的技巧。复杂网络讨论班上有曹志威、王佳晟、覃文杰、宋磊同学以及张毅超老师。

感谢实验室的师兄师姐，我从你们身上学习到了实验室文化以及做研究的入门方法。这些同学分别是李文根、吴秋阳、陈惠东、叶维帅、李维丹、卜宏达、王佳晟、卢赟、朱海泉、耿欣、褚翔伟、许中。

感谢几位同届同学与我共同度过研究生的适应期，怀念我们一道吃饭、上课当助教的时光。他们分别是陶汉、徐伟、裴胜兵。

感谢各位师弟师妹对我的支持与关心，他们分别是覃文杰、徐影、宋磊、姚恒、时运佳、葛声利。

{ 49 % : 感谢来参加我研究生毕业论文答辩的各位师兄弟姐妹。 }

感谢徐伟与王佳晨同学，感谢你们在计算机技术方面对我的帮助。

感谢实习期间的各位同事，从你们身上我学习到了计算机热门技术以及认真负责的工作态度。

{ 58 % : 感谢我的朋友，感谢你们一直以来对我的支持与陪伴。 } 你们让我感受到了人间的真情。

感谢我的家人，你们是我生命中的伟人！

{ 69 % : 最后，由衷的感谢在百忙之中评阅论文以及参加我的答辩的各位专家、老师。 } 感谢你们！

参考文献

[1] Watts D J , Strogatz S H. Collective dynamics of small-world networks[J]. nature , 1998 , 393(6684): 440-442.

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

[3] LL , Zhou T. Link prediction in complex networks: A survey[J]. Physica A: Statistical Mechanics and its Applications , 2011 , 390(6): 1150-1170.

[4] Vineyard C M , Verzi S J , Bernard M L , et al. A multi-modal network architecture for knowledge discovery[J]. Security Informatics , 2012 , 1(1): 1-12.{cite1}

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

[6] Watts D J , Strogatz S H. Collective dynamics of small-world networks[J]. Nature , 1998 , 393: 440-442.

[7] Christian M Schneider , Lucilla de Arcangelis , Han J Herrmann. Scale free networks by preferential depletion. arXiv: physics.soc-ph , 2011 , 1103.1396v1.

[8] Perotti J I , Billoni O V , Tamarit F A et al. Emergent self-organized complex network topology out of stability constraints. Phys. Rev. Lett. , 2009 , 103: 108701-108704.

[9] Scholz M. Node similarity is the basic principle behind connectivity in complex networks. arXiv: physics.soc-ph , 2010 , 1010.0803v1.

- [10] Milo R , Shen-Orr S , Itzkovitz S et al. Network motif: simple building blocks of complex networks. *Science* , 2002 , 298: 824-827.
- [11] Clauset A , Moore C , Newman M E J. Structure inference of hierarchies in networks. *Proceedings of the 2006 conference on statistical network analysis*.
- [12] Vito Latora , Massimo Marchiori. Is the Boston subway a small world network[J]. *Physica A* 314(2002): 109—113
- [13] Vamsi Kalapala , Vishal Sanwalani , Aaron Clauset , et al. Moore Scale Invariance in Road Networks[J]. arXiv: physics/0510198v2 [phys] CS.SOCph] 8 Mar 2006
- [14] Parongama Sen , Subinay Dasgupta , Arnab Chatterjee , et al. Smallworld properties of the Indian Railway network[J]. arXiv: condmat/0208535 v2 3 1 Dec 2002
- [15] R. Guimera , S. Mossa , A. Turtschi , et al. The worldwide air transportation network: Anomalous centrality, community structure, and cities global roles[J]. *LAN Amaral Proceedings of the National Academy of Sciences* , 2005—National Acad Sciences
- [16] 陆化普. *解析城市交通*[M]. 北京: 中国水利水电出版社 , 2001.9 [17] National Cooperative Highway Research Program. *The Benefits of Reducing Congestion*[C]. Cambridge Systematics , Inc. 2002.1
- [18] 2004年中国汽车工业年鉴。 中国汽车技术研究中心. 2004
- [19] 赵月. 复杂交通网络拥堵特性及控制方法研究[D]. 西南交通大学 , 2009.
- [20] SARUKKAI R R. Link prediction and path analysis using markov chains[J]. *Computer Networks* , 2000 , 33(1-6): 377-386.
- [21] ZHU J , HONG J , HUGHES J G. Using markov chains for link prediction in adaptive web sites[J]. *Lect Notes Comput Sci* , 2002 , 2311: 60-73.
- [22] POPESCU A , UNGAR L. Statistical relational learning for link prediction[C]// *Proceedings of the Workshop on Learning Statistical Models from Relational Data*. New York: ACM Press , 2003: 81-87.
- [23] OMADADHAIN J , HUTCHINS J , SMYTH P. Prediction and ranking algorithms for event-based network data[C]// *Proceedings of the ACM SIGKDD 2005*. New York: ACM Press , 2005: 23-30.
- [24] LIN D. An information-theoretic definition of similarity[C]// *Proceedings of the 15th Intl Conf Mach. Learn.* San Francisco , Morgan Kaufman Publishers , 1998: 296-304.

- [25] Ravasz E , Somera A L , Mongru D A , et al. Hierarchical organization of modularity in metabolic networks[J]. science , 2002 , 297(5586): 1551-1555.
- [26] Zhou T , LL , Zhang Y C. Predicting missing links via local information[J]. The European Physical Journal B , 2009 , 71(4): 623-630.
- [27] Adamic L A , Adar E. Friends and neighbors on the web[J]. Social networks , 2003 , 25(3): 211-230.
- [28] LL , Jin C H , Zhou T. Similarity index based on local paths for link prediction of complex networks[J]. Phys Rev E Stat Nonlin Soft Matter Phys , 2009 , 80(2): 046122.
- [29] Katz L. A new status index derived from sociometric analysis[J]. Psychometrika , 1953 , 18(1): 39-43.
- [30] Fouss F , Pirotte A , Renders J M , et al. Random-Walk Computation of Similarities between Nodes of a Graph with Application to Collaborative Recommendation[J]. IEEE Transactions on Knowledge Data Engineering , 2007 , 19(3): 355-369.
- [31] Brin S , Page L. The anatomy of a large-scale hypertextual Web search engine [J]. Computer Networks Isdn Systems , 1998 , 30(17): 107-117.
- Proc Natl Sci Acad USA , 2009 , 106(52): 22073-22078.
- [32] LL , Pan L , Zhou T , et al. Toward link predictability of complex networks.[J]. Proceedings of the National Academy of Sciences of the United States of America , 2015 , 112(8): 201424644.
- [33] CLAUSET A , MOORE C , NEWMAN M E J. Hierarchical structure and the prediction of missing links in networks[J]. Nature , 2008 , 453: 98-101.
- [34] GUIMERA R , SALES-PARDO M. Missing and spurious interactions and the reconstruction of complex networks[J].
- [35] Murata T , Moriyasu S. Link Prediction of Social Networks Based on Weighted Proximity Measures[C]// Web Intelligence , IEEE/WIC/ACM International Conference on. IEEE Xplore , 2007: 85-88.
- [36] Narayanan A , Shi E , Rubinstein B I P. Link Prediction by De-anonymization: How We Won the Kaggle Social Network Challenge[J]. 2011 , 42(4): 1825-1834.
- [37] Chua F C T , Lim E P. Modeling Bipartite Graphs Using Hierarchical Structures[C]// International Conference on Advances in Social Networks Analysis and Mining. IEEE , 2011: 94-101.
- [38] Traffic congestion in interconnected complex networks
- [39] Traffic fluctuation on weighted networks

- [40] T. Ohira , R. Sawatari , Phase transition in a computer network traffic model , Phys. Rev. E 58 (1998) 193195.
- [41] S. Boccaletti , V. Latora , Y. Moreno , M. Chavez , and D.- U. Hwang , Phys. Rep. 424 , 175 (2006).
- [42] [137] Z. Toroczkai , K.E. Bassler , Nature 428 (2004) 716.
- [43] [138] Z. Toroczkai , B. Kozma , K.E. Bassler , N.W. Hengartner , G. Korniss , preprint cond-mat/0408262.
- [44] R. Guimera , A. Arenas , A. Daz-Guilera , and F. Giralt , Phys. Rev. E 66 , 026704 (2002).
- [45] G.-Q. Zhang , S. Zhou , D. Wang , G. Yan , and G.-Q. Zhang , Physica A 390 , 387 (2011).
- [46] L. Zhao , Y.-C. Lai , K. Park , and N. Ye , Phys. Rev. E 71 , 026125 (2005).
- [47] D. De Martino , L. DallAsta , G. Bianconi , and M. Marsili , Physical Review E 79 , 015101 (2009).
- [48] G. Yan , T. Zhou , B. Hu , Z.-Q. Fu , and B.-H. Wang , Phys. Rev. E 73 , 046108 (2006).
- [49] X. Ling , M.-B. Hu , R. Jiang , and Q.-S. Wu , Phys. Rev. E 81 , 016113 (2010).
- [50] F. Tan and Y. Xia , Physica A 392 , 4146 (2013).
- [51] B. Danila , Y. Yu , J. A. Marsh , and K. E. Bassler , Phys. Rev. E 74 , 046106 (2006).
- [52] 李晓燕. 面向城市交通拥堵疏导的节点分流策略研究[D]. 广东工业大学 , 2014.
- [53] Zhang Guo Qing , Wang Di and Li Guo Jie. Enhancing the transmission efficiency by edge deletion in scale-free networks[J]. Physical Review E , 2007 , 76(1).
- [54] Xie Y B , Zhou T , Wang B H. Scale-free networks without growth[J]. Physica A Statistical Mechanics Its Applications , 2008 , 387(7): 1683-1688.
- [57] Albert-Lszl Barabsi and Rka Albert. Statistical mechanics of complex networks[J]. Reviews of modern physics , 2002 , 74(1): 47-94

{ 95 % : 个人简历、在读期间发表的学术论文与研究成果 }

个人简历 :

李星，男，1990年05月生。

{ 54 % : 2013年6月毕业于安徽大学计算机科学与技术学院，计算机科学与技术专业，获工学学士学位。 }

{ 69 % : 2014年9月入同济大学就读硕士研究生。 }

已发表论文：

{ 63 % : 1. 基于聚类系数的链路预测算法改进 }

2. Zhang Y , Li X , AzizAlaoui M A , et al. Knowledge diffusion in complex networks[J]. Concurrency Computation Practice Experience , 2016: n/a-n/a.

参与科研项目：

检测报告由PaperPass文献相似度检测系统生成
Copyright 2007-2017 PaperPass