****

**硕士学位论文**

**链路预测及交通拥堵恢复序列的预测**

姓 名：李星

学 号：1433384

所在院系：电子信息与工程学院

学科门类：工学

学科专业：计算机科学与技术

指导教师：关佶红 教授

二〇一七年二月

****

A dissertation submitted to

Tongji University in conformity with the requirements for

the degree of Master

**Link Prediction and Traffic Congestion Recovery Sequence Prediction**

Candidate: Li Xing

Student Number: 1433384

School/Department: College of Electronic and Information Engineering

Discipline: Master of Engineering

Major: Computer Science and Technology

Supervisor: Guan Jihong

February, 2017

|  |
| --- |
| **链路预测及交通拥堵恢复序列的**  **预测**  **李星**  **同济大学** |

**学位论文版权使用授权书**

本人完全了解同济大学关于收集、保存、使用学位论文的规定，同意如下各项内容：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，并采用影印、缩印、扫描、数字化或其它手段保存论文；学校有权提供目录检索以及提供本学位论文全文或者部分的阅览服务；学校有权按有关规定向国家有关部门或者机构送交论文的复印件和电子版；在不以赢利为目的的前提下，学校可以适当复制论文的部分或全部内容用于学术活动。

学位论文作者签名：

年 月 日

**同济大学学位论文原创性声明**

本人郑重声明：所呈交的学位论文，是本人在导师指导下，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，本学位论文的研究成果不包含任何他人创作的、已公开发表或者没有公开发表的作品的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。本学位论文原创性声明的法律责任由本人承担。

学位论文作者签名：

年 月 日

摘要

长期以来，网络的结构及其演化机制近都是复杂网络的一个基本研究问题。复杂网络的研究涉及到物理、计算机、生物、经济等多个学科。如何通过现有网络信息来预测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及其重要的研究点。

网络中的链路预测是指如何通过已知的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链接的可能性。链路预测是对于网络分析和数据挖掘的一项支撑学科。近年来，随着网络科学的快速发展，链路预测现在已经收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。

交通网络属于复杂网络的一种。交通拥堵已经成为现在社会最严重的社会问题之一，为了解决交通拥堵问题，人们已经提出了许许多多的办法，但是都无法避免拥堵的发生，一旦拥堵发生，还是会对我们的生活造成极大的危害，因此我们希望能找到一种办法可以在交通发生拥堵时迅速予以对策并有效地恢复网络的处理能力。

本文主要提出了对于链路预测算法的一种改良策略，并将链路预测方法运用到了对于实际交通拥堵问题的解决之上。本文所做的主要工作及贡献如下：

在链路预测算法的改进方面：

1）本文提出了一种基于网络节点中聚类系数的算法优化思路，并将该优化思路运用到 9 种经典的算法上，在 5 种真实数据集上的实验表明了该优化思路的可行性。

2）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。

3）传统的链路预测比较实验，对于测试集多选取 10% 这一固定值，本文为了说明算法的健壮性，针对不同大小的测试集进行了广泛的实验。

在利用链路预测方法解决交通拥堵问题方面：

4）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方面的研究。然而，再优良的网络设计也不能完全杜绝交通堵塞问题。基于链路预测的启发，本文提出了基于链路预测思维的交通拥堵序列恢复的算法，当网络一旦发生拥堵之后，根据我们的推荐算法可对交通线路有序进行恢复，从而在最短的时间内提升网络的运载能力。

5）本文在一种模拟网络及两种真实网络上进行了大量的实验，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。本文提出的基于链路预测的反相关性算法是 RPA、RAPA、RSA、RLP，这些算法都是基于节点的反相关性设计的，这在之前的研究当中少有提及。

**关键词**：链路预测；聚类系数; 错边识别; 交通恢复序列

**ABSTRACT**

**Key Words:**

第1章 绪论 5

1.1研究背景与意义 5

1.1.1 链路预测 5

1.1.2 交通拥堵问题 7

1.2论文研究问题 8

1.2.1链路预测算法 8

1.2.2交通拥堵恢复序列推荐 8

1.3 国内外研究现状 9

1.3.1 链路预测 9

1.3.2 交通拥堵问题 10

1.4 本文主要研究工作及贡献 12

1.4.1 基于聚类系数的链路预测算法研究 12

1.4.2 利用链路预测思维进行交通拥堵恢复序列的研究。 12

1.5 本文的章节安排 12

第2章 链路预测方法分析 15

2.1 问题描述 15

2.1.1 丢失边的预测 15

2.1.2 错误边的预测 16

2.2 数据集划分 16

2.2.1 随机抽样 16

2.2.2 逐项遍历 16

2.2.3 k折叠交叉检验 17

2.2.4 滚雪球抽样 17

2.2.5 其他方法 17

2.3 经典的链路预测算法 17

2.3.1 CN 指标(Common Neighbors) 18

2.3.2 Salton 指标 18

2.3.3 Jaccard 指标 19

2.3.4 Sorensen 指标 19

2.3.5 大度节点有利指标(hub promoted index, HPI) 19

2.3.6 大度节点不利指标(hub depressed index, HDI) 19

2.3.7 LHN 指标 20

2.3.8 AA 指标 (Adamic-Adar) 20

2.3.9 RA 指标 20

2.4 评价指标 20

2.4.1 AUC (area under the receiver operating characteristic curve) 20

2.4.2 准确度（Precision） 21

2.4.3 排序分（Ranking Score） 21

2.5 本章小结 22

第3章 基于聚类系数的链路预测算法 23

3.1 聚类系数的定义 23

3.2 基于聚类系数的链路预测算法改进 23

3.2.1 CCN 指标(Clustering Common Neighbors) 23

3.2.2 CSalton 指标(Clustering Salton) 24

3.2.3 CJaccard 指标(Clustering Jaccard) 24

3.2.4 CSorensen 指标(Clustering Sorensen) 24

3.2.5 CHPI 指标(Clustering  hub promoted index) 24

3.2.6 CHDI 指标(Clustering hub depressed index ) 24

3.2.7 CLHN 指标(Clustering LHN) 25

3.2.8 CAA 指标(Clustering Adamic-Adar) 25

3.2.9 CRA 指标(Clustering Resource Allocation) 25

3.3 数据集介绍 25

3.3.1 数据集描述 25

3.3.2 数据集的网络特征的统计 26

3.4 实验结果分析 26

3.4.1 参数设定 26

3.4.2 实验结果以及分析 27

3.5 本章小结 32

第4章 交通拥堵恢复方法分析 33

4.1 问题描述 33

4.2 解决思路 33

4.3 流量模型 34

4.3.1 吞吐量定义 34

4.3.2 衡量指标μ 35

4.4 本章小结 36

第5章 交通拥堵恢复序列算法 37

5.1 算法描述 37

5.1.1 PA算法（Preferential Attachment Index）[47] 37

5.1.2 LP算法（Local Path Index） [48,49] 37

5.1.3 SA算法（Three order Similarity Index） 37

5.1.4 RPA算法（Reverse Preferential Attachment Index） 38

5.1.5 RAPA算法（Reverse Add Preferential Attachment Index） 38

5.1.6 RLP算法（Reverse Local Path Index） 38

5.1.7 RSA算法（Reverse Three order Similarity Index） 38

5.2 数据集描述 39

5.2.1 数据集描述 39

5.2.2 网络特征统计 43

5.3 实验流程 44

5.4 评价方式 46

5.5 实验结果分析 47

5.5.1 加州路网上的结果分析 47

5.5.2 不同网络损毁比例对于实验的影响 49

5.5.3 不同网络尺寸对于实验的影响 50

5.5.4 不同大小的流量对于实验的影响 51

5.6 本章小结 51

第6章 总结与展望 53

6.1 总结 53

6.2 进一步工作 53

致谢 55

参考文献 57

# 第1章 绪论

## 1.1研究背景与意义

### 1.1.1 链路预测

在现实世界中，从神经系统到生态系统，从道路交通到 Internet, 从蚁群到人类社会，无不属于复杂系统的范畴。在这些系统内，往往相互作用的个体众多且作用种类复杂。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。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。这样在采样过程中难免有错误或冗余的链接出现。因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。

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。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。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。

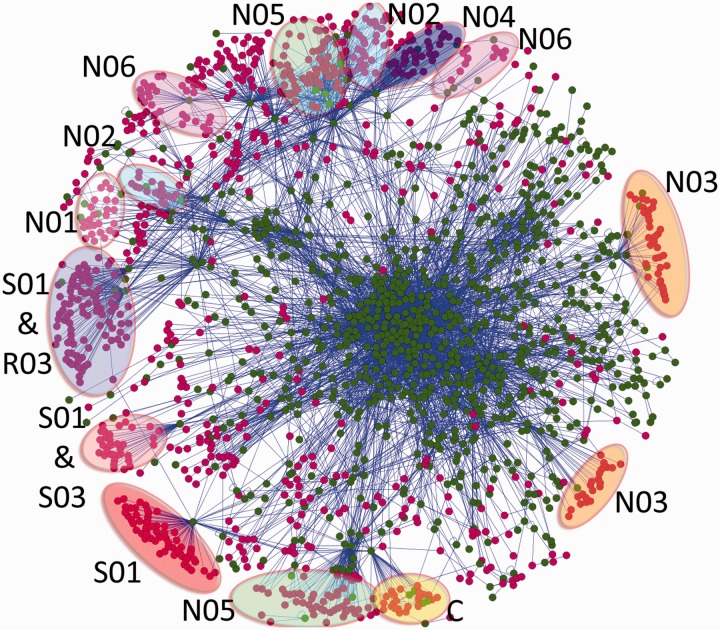


图 1.1 蛋白质相互作用网络

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，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。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，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。链路预测算法在实际生产生活中的可应用性是它收到广泛关注的重要原因。

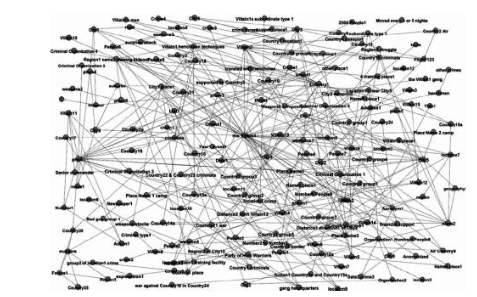


图 1.2 恐怖分子犯罪网络

链路预测的方法也可直接用于信息的推荐。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。

另外，它还可以作为评价网络演化机制的一种方式。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)[2] 、优先连接 (preferential attachment)[1] 、优先去除 (preferential depletion)[3]、稳定性限制(stability constraints)[4]、基于欧几里德距离的节点相似度 (node-similarity based on Euclidean distance)[5]等等。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。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。

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，这个问题仍然是历久弥新，开始在复杂网络领域内蓬勃发展。原因在于: 一、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，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。二、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。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。链路预测既然是对网络结构的一种预测，更深入地研究它与链路预测的内在关系，比如细致的分析网络的一些普适子结构(比如模体[6]、社团等)及研究某些网络的特殊结构(比如等级结构[7])，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

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。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。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[8-11]，所以在对交通网络拓扑特性进行系统分析的基础上，结合交通网络本身的演化特性，构建符合实际的交通拥堵模型，同时探讨网络节点在传播中重要性的度量，分析不同拓扑结构特性对拥堵的影响，进而提出有效地缓解和预防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。

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，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。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。

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。一般来说，城市交通拥挤的产生有三种方式:暂时的路障、永久能力瓶颈和网络某路段需求的随机波动。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

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。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， 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 7,900万 h和 3亿 L，折算的经济损失总额超过630亿美元。

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。 目前各国城市大气污染物主要是一氧化碳C0、碳氢化合物HC、氮氧化合物 N0。、浮沉等，而汽车是这些污染物的主要排放源(达35%左右)。研究发现，C0和HC的排放受行驶工况的影响很大，机动车排放c0和HC的数量随速度的增加呈负指数下降[12]。美国环境保护总署的一项研究表明，当干道车 速由10m.1e/h增加到20m.Ie/h时，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%[13]。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。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[15]。

## 1.2论文研究问题

### 1.2.1链路预测算法

随着链路预测这一学科的快速发展，链路预测在理论与实践当中日益发挥着更为重要的作用。近些年来关于链路预测的算法更是层出不穷，大部分新提出的算法均是基于网络结构的相似性。好的链路预测算法可以更好的预测网络当中的隐含信息以及网络的演化机制，本文的第一个研究内容便是希望找到一种提升链路预测准确性的方法。

### 1.2.2交通拥堵恢复序列推荐

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。

本文需要解决的问题之一便是，当交通网络当中出现多条线路的拥堵之时，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。例如，在城市道路之中，我们现在有3条道路出现了拥堵，由于资源的有限，我们不能一次性疏导这三条道路，然而我们现在的疏导顺序的选择有6种。这6种疏导顺序对于流量的恢复效果显然是会有区别的，我们希望做到的便是尽量找到一个最优的恢复序列，该恢复序列可最大程度上恢复整体网络的流量。

## 1.3 国内外研究现状

### 1.3.1 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。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。之后的很长一段时间，网络学科都没有很大的发展。1959 年，匈牙利的著名数学家 Erdös 和 Rényi 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[16]。随机网络的 出将网络的理论研究拓展到了更为广泛的科 研领域，比如社会学家也可以利用随机理论对于人类关系网络进行建模。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Barabási 在1999 出的无标度网 络[17,18]。小世界网络指网络中大多数的节点不直接相连，但彼此之间只需通过少量节点即可 以到达，即“六度分离”理论的诠释。无标度网络指网络中节点度的分布符合幂指数，即少数的节点具有很高的度分布而大多数节点的度相对较低。 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。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[19]。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。

链路预测具有重要的实践以及理论价值[20,21,22]。人类已知的蛋白质相互作用网络只占真实 网络的冰山一角，很多蛋白质之间是否有相互作用需要大量的实验来验证。酵母细菌的相互 作用网络已被观察到的只有 20%，人类网络更是只有 0.3%。由于验证蛋白质相互网络将消 耗巨大的人力以及时间的成本，所以如果能用链路预测的方式对可能存在的边进行预测，进 而指导实验，将对于人类的生命工程产生巨大的推动作用。社交网络中也存在信息丢失的情 况。比如 QQ 当中，我们好友列表中并不能涵盖我们全部的好友信息。链路预测此时便可以 根据现有好友列表向我们进行推荐可能认识的好友，进一步完善我们的好友列表 以及增强 用户对于产品的忠诚度。同样的道理，链路预测的算法也可以在商品推荐中得以应用。链路 预测的理论价值主要在于帮助我们更好的认识网络的演化机制。针对某种或某类网络，有许 多模型都 供了可能的演化机制，但由于刻画网络性质的统计变量很多，所以很难比较哪种 演化机制更好。有了链路预测，我们便得到了一个统一量化的指标，利用该演化机制构建链 路预测的算法，算法预测准确率高者更符合网络的演化，从而分析哪一个机制能够更好的刻 画网络的演化，这大大推进了网络演化机制的研究工作。

早期的链路预测主要基于机器学习和马尔科夫链。该类方法虽然具有较高的预测准确 率，但较高的时间复杂度限制了被预测网络的规模。同时，由于该类方法的分类往往需要获 取节点的属性信息，节点的属性信息往往很难获取即使获取节点信息也未必能保证信息的可 靠性。比如在社交网络之中，用户的年龄、性别、住址未必会如实填写，这便给该类基于节 点属性的算法造成了很大的障碍。由于基于马尔科夫以及机器学习方法的弊端，最近几年的 链路预测更多转移到了利用网络的拓扑信息进行预测。利用网络结构进行预测的方法主要分 为基于相似性的链路预测和基于似然分析的链路预测。基于相似性的链路预测的前 假设 是，如果两个点越相似那么这两个点存在连边的可能性越大。相比较基于相似性的链路预测， 基于思然分析的链路预测是一种更为复杂的框架，虽然可以取得较高的准确率，但同基于机 器学习的方法类似，当网络节点数目达到数千个之时，算法处理起来之时便会感到吃力。考 虑到计算的复杂度以及通用性，本文所提出的算法可以归纳到基于相似性的链路预测框架之内。

### 1.3.2 交通拥堵问题

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[23]。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[24]，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[25]。关于交通拥堵问题的研究主要是对于交通系统的规划与控制。网络的设计主要考虑的是交通网络拓扑结构的设计和资源在节点上的分配，因此相应的研究也主要分为两大类。以下我们就简要的介绍一下这两类研究的进展。

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。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会更有效的应对突发性的攻击[26]。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[27,28]。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[29]。

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的前提下。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。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[30]。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。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[31]。

关于交通网络的控制研究主要是指怎样的路由方案可很好的提升网络的运载能力。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[32]。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，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[33]。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。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[34]。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[35]。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[36]。

## 本文主要研究工作及贡献

本文的主要研究内容及工作分为以下两个大的方面：

### 1.4.1 基于聚类系数的链路预测算法研究

在本文中，我们 出了利用网络节点的聚类系数这一特征来 升基于相似性2017/2/26网络噪声纠错能力上的 升。除此之外，过往研究将测试集 的比例固定在 10%这一值之上，我们对此进行了扩展，我们将测试集的比例伸展到了 4%到 20%范围之间，进一步 升了实验的完整性，这一范围的变化也验证了我们算法的健壮性。

### 1.4.2 利用链路预测思维进行交通拥堵恢复序列的研究。

本文希望结合图论中的某些经典算法与链路预测算法，能够通过实验来探究链路预测在解决交通堵塞上的有效性，具体来讲本实验想要利用一些量化的标准来衡量交通路网的运载能力， 并在运输网络发生阻塞后，探究能否利用链路预测算法来提供一种解决方案，从而能够最有效地恢 复运载能力。在得出初步结论后，能够对课题进行拓展，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。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。交通网络也属于复杂网络的一种，我们分析了面对交通拥堵问题现有的解决方案，现有方案多基于事先规划的思想，但最好的规划也无法完全避免交通的拥堵问题。应对交通拥堵问题，基于链路预测思想的启发，我们提出了利用链路预测算法解决交通恢复序列的思想。本章最后交代了 本文的研究内容和组织结构。

第 2 章，链路预测方法分析。该章节首先对链路预测问题进行了定义，现有的链路预测主要是针对丢失边信息以及错误连边信息的预测。之后依次介绍了链路预测研究中的数据集划分、本文所使用的经典算法、链路预测准确性的评价指标。链路预测将数据集分为训练集和验证集两个部分，主要的划分方法有随机抽样、逐项遍历、k折叠交叉检验、滚雪球抽样等。算法的评价指标则主要有AUC、准确度和排序分三种。

第 3 章，基于聚类系数的链路预测。首先介绍了聚类系数的定义。之后介绍了基于聚类系数改进后的9种算法。之后是对于实验部分所使用数据集的介绍。最后在5类数据集之上进行了对比实验，对比的算法为改进前的9种算法和改进后的9种算法。

第 4 章，交通拥堵问题的研究。首先介绍了本文所研究的交通拥堵问题的定义，即本文想研究的问题是当出现交通拥堵问题，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。之后，本章介绍了先前研究是提升交通网络流量的方法。再之后，介绍了用于流量模拟的常见模型，紧接着介绍了本文提出的利用链路预测的思维进行交通拥堵的恢复序列推荐。

第 5 章，交通拥堵恢复序列算法。 本章首先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。之后介绍了本章实验所需要的3类数据集，其中模拟数据一种真实数据2种。紧接着介绍的是实验流程。最后是对于实验结果的讨论。

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。

本文研究的主要问题有两个。第一个大问题是改进的链路预测算法，对应于第二、三两章。第二个大问题是利用链路预测思维解决交通恢复序列问题，该问题对应于四五两章。六章内容当中最为核心的是第三章以及第五章的内容，这两张内容中包含了本文提出的算法、使用的数据集以及实验流程。

# 第2章 链路预测方法分析

本章主要介绍了链路预测的基础概念及相关技术。本章显示对链路预测问题进行了描述，之后对于链路预测之中数据集的划分做了简要介绍。链路预测包含众多算法，本章介绍了9种经典算法，这9种算法也是之后需要进行改进的算法。本章的最后介绍了关于链路预测算法的相应评价指标。

## 2.1 问题描述

定义 G（V,E）为一个无向无权的网络，其中 V表示网络中节点的集合，E表示网络中边的集合。网络中节点的数量为 N，若该网络为一个全联通网络，那么网络中边的数量应该为 N\* (N-1)/ 2,我们表示为全集 U。

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 Sxy。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。

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，为了预测算法的准确性，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 EP = E,且 ET EP = ∅ 。我们会将属于 U但是不属于 E的边称为不存在的边，在计算算法的准确性时，我们是基于训练集中的信息，验证集仅用作数据的验证。

如图2.1所示，其中图左侧是完整的网络，该网络拥有 9个节点，14条边。全集包含 36条边，我们在这 19条边中选出 3条作为验证集，如右侧图中虚线所表示的，剩下的 11条边就是训练集。之后我们就需要根据我们选用的算法将 25条未知边赋予一个分数值，并按降序排序。假如在全部的分数值中，我们的测试边的分数越高，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。

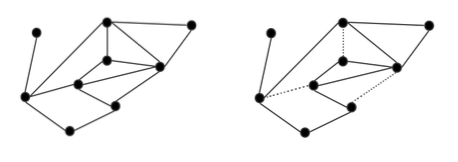


图 2.1 训练集和验证集的划分

### 2.1.2 错误边的预测

与丢失边的预测略有不同，为了验证算法的准确性，我们需要模拟错误连边的情况。我们将网络中添加的边用作验证集 EP，此时的训练集 ET = E + EP。在堆算法的预测准确性进行计算时，我们基于的网络拓扑结构包含中网络中原本存在的边以及新加入的扰动边。

## 2.2 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的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。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。

### 2.2.1 随机抽样

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

### 2.2.2 逐项遍历

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。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，求 M次准确率的平均值，该平均值即为我们算法的准确率。该方法相较于随机抽样法中每条边都有被选中的机会。

### 2.2.3 k折叠交叉检验

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。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。当 k等于 M时，该方法退化为逐项交叉检验法。

### 2.2.4 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。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。

### 2.2.5 其他方法

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。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。

## 2.3 经典的链路预测算法

本章所选取的算法是链路预测中最为经典也是最为广泛使用的算法。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。

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，兴趣相同等，我们则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。

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。由于节点属性的难易获取，所以利用节点属性进行链路预测便遇到了很大的障碍。

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。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很好的抓住了网络的结构特征，即网络的形成机制。

基于节点的聚类系数对于链路预测算法的改进中的几种算法便是基于拓扑相似性的算法，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。

### 2.3.1 CN 指标(Common Neighbors)

对于网络中的节点 vx，Γ(x)表示该节点的所有邻居，同样的 Γ(y)表示节点 vy的所有邻居集合。下式表示的便是节点 vx和节点 vy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[38]。

Sxy = | Γ(y)| （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2.2所示，对于点 A 和点 D 来说，根据定义 Γ(A) = {B,C},

Γ(D) = {B,C}，Γ(A) ∩Γ(D) = {B,C}。因此按照上述公式来计算，SAD=2。

而对于点 A 和点 E 来说，Γ(A)={B,C}, Γ(E)={B}，Γ(A)∩Γ(E)={B},得 SAE=1。

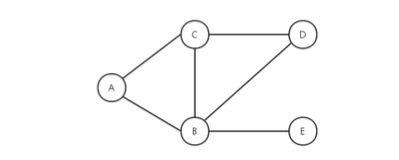


图 2.2 CN算法图示

### 2.3.2 Salton 指标

Salton指标又称作余弦相似性指标，kx和 ky分别表示节点 vx和节点 vy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领居数目一定的时候，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[39]。

Sxy = （2.2）

### 2.3.3 Jaccard 指标

Jaccard指标提出于一百多年前，相比于 CN指标，该算法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[40]。

（2.3）

### 2.3.4 Sorensen 指标

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[41]。

Sxy = （2.4）

### 2.3.5 大度节点有利指标(hub promoted index, HPI)

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于网络的连接[42]。

Sxy = （2.5）

### 2.3.6 大度节点不利指标(hub depressed index, HDI)

该定义与 HPI类似，知识分母部分略有不同，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于网络的连接[43]。

Sxy = （2.6）

### 2.3.7 LHN 指标

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[44]。

Sxy = （2.7）

### 2.3.8 AA 指标 (Adamic-Adar)

这是对于 CN指标的一种改进，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[45]。

Sxy = （2.8）

### 2.3.9 RA 指标

RA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。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。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[43]。

Sxy = （2.9）

注：缺图示

## 2.4 评价指标

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。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。

### 2.4.1 AUC (area under the receiver operating characteristic curve)

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，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，如果测试集中边的评分大于不存在的 边，那么加 0。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 当中分别取出一条边，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。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。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。

### 2.4.2 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。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。

### 2.4.3 排序分（Ranking Score）

排序分考虑了验证集在最终排序结果中的位置。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e 在排序中的排名。那么这条边的排序分为:

RS =

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。

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，本文中关于链路预测的准确度便是以AUC来进行衡量的。

## 2.5 本章小结

本章介绍了链路预测相应的技术。对于链路预测这一问题，我们分为了丢失边预测以及错误连边预测这两种情况。对于数据集的划分，我们分析了随机抽样、逐项遍历、k折叠交叉检验，滚雪球抽样这几种方式的区别，并决定在之后的实验当中选择随机抽样的划分方式。此外我们介绍了9种经典的链路预测算法。最后我们介绍了AUC、准确度以及排序分这样的3种链路预测指标，并决定选择AUC指标作为我们后续实验部分的评价指标。

# 第3章 基于聚类系数的链路预测算法

本章是我们所研究的第一个问题—链路预测算法的关键章节。该章节当中我们介绍了聚类系数这一网络概念，并提出了利用该网络指标对现有的9种经典算法进行改进的思路。之后介绍了实验部分所使用到的5类数据集。在最后的实验部分，我们在该5类网络当中进行了改进前后共18种算法的比较。

## 3.1 聚类系数的定义

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。针对网络中的任意一个节点vi，其聚类系数定义为该点所有邻居节点之间的连接数占领居间最大连边可能性的比例，其数学表达式为：

Ci = （3.1）

式子中 li表示节点 vi的邻居节点之间连边的数目，ki表示节点vi度的大小，Ci即为节点 vi的聚类系数大小。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。

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。如果节点 v1连接于节点 v2，节点v2连接于节点v3，那么节点 v3很可能与 v1相连接。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。基于大的聚类系数 促进网络的连接，所以本文认为可将聚类系数这一特征融入到现有算法之中提升算法对于部分网络的预测准确度。

注：缺图示

## 3.2 基于聚类系数的链路预测算法改进

对于上文中所提到的9种链路预测算法，加入聚类系数这一元素得到的改进后的算法为：

### 3.2.1 CCN 指标(Clustering Common Neighbors)

Sxy  = （3.2）

加入聚类系数后的CNN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，而在CNN指标看来，共同邻居节点的贡献度是不同的，聚类系数较高的共同邻居节点的贡献度较大。

### 3.2.2 CSalton 指标(Clustering Salton)

Sxy = （3.2）

该指标与Salton指标的区别在于分母部分，同CNN指标，CSalton指标将共同邻居节点的贡献度区别看待，同样认为较大的聚类系数的共同邻居节点的贡献度较大。

### 3.2.3 CJaccard 指标(Clustering Jaccard)

（3.4）

改进原理同CCN。

### 3.2.4 CSorensen 指标(Clustering Sorensen)

Sxy = （3.5）

改进原理同CCN。

### 3.2.5 CHPI 指标(Clustering  hub promoted index)

Sxy = （3.6）

改进原理同CCN。

### 3.2.6 CHDI 指标(Clustering hub depressed index )

Sxy = （3.7）

改进原理同CCN。

### 3.2.7 CLHN 指标(Clustering LHN)

Sxy = （3.8）

改进原理同CCN。

### 3.2.8 CAA 指标(Clustering Adamic-Adar)

Sxy = （3.9）

改进原理同CCN，CAA指标同时看重共同领居节点度大小和聚类系数的区别，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。

### 3.2.9 CRA 指标(Clustering Resource Allocation)

Sxy = （3.10）

改进原理同CAA。

## 3.3 数据集介绍

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两个数据集当中手机而来。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。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。

### 3.3.1 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。Footbal 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。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。

### 3.3.2 数据集的网络特征的统计

表3.1 数据集的网络统计特征

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 网络 | |V| | |E| | D | C | <k> | <d> | H |
| Elegans | 297 | 2148 | 0.0488 | 0.2923 | 14.4646 | 2.4553 | 1.8008 |
| Football | 35 | 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 |

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 。C 表示节点的聚类系数，上文已经给出了计算公式。<k>表 示网络中节点度的平均值。<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。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。

## 3.4 实验结果分析

### 3.4.1 参数设定

为了保证试验的准确定，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。

### 3.4.2 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3.1（a）为缺失边的发现，图3.1（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。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 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。随 着测试集比例的 升，改进后的方法同样优于原算法。

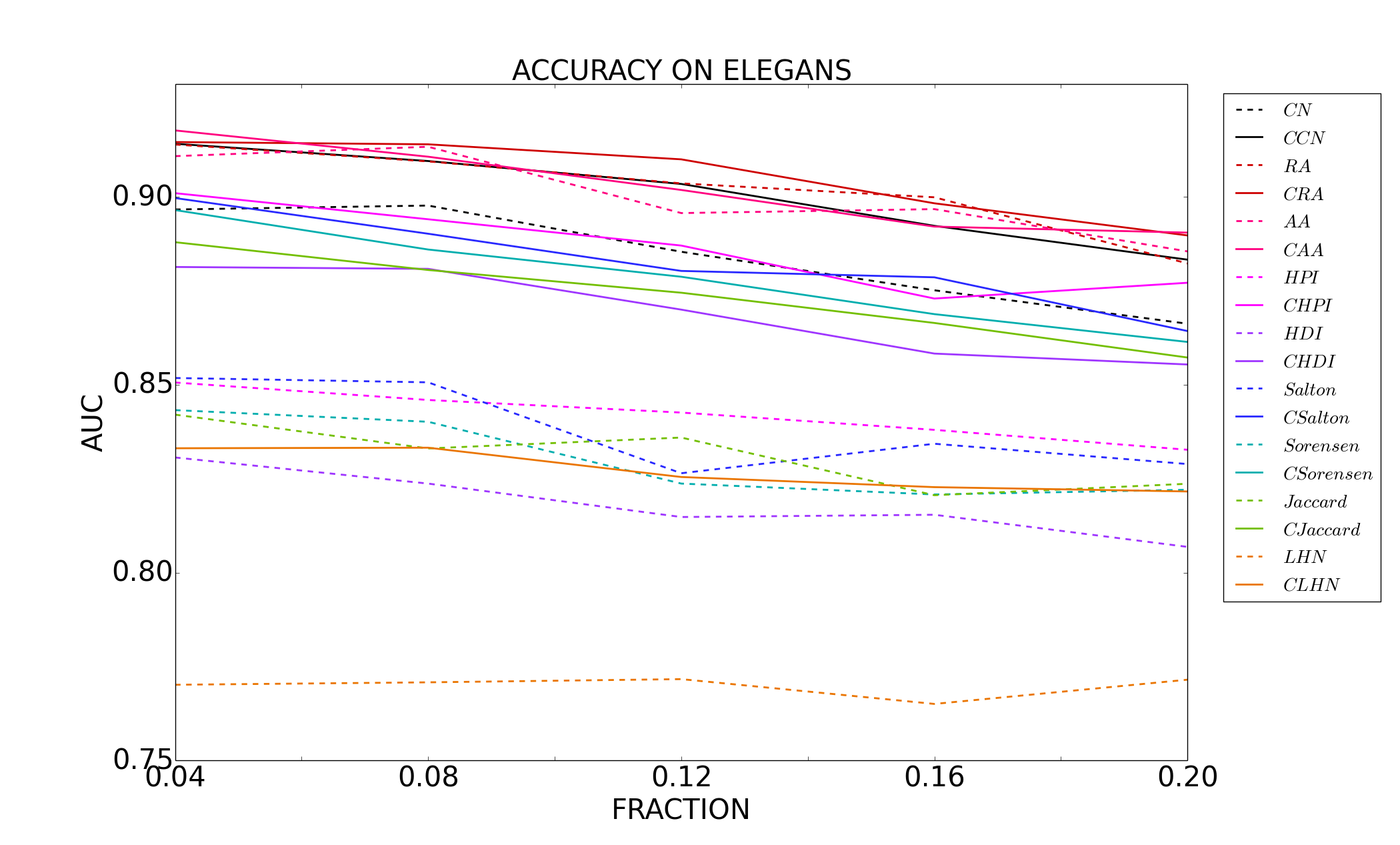


图 3.1（a） Elegans网络缺失边发现

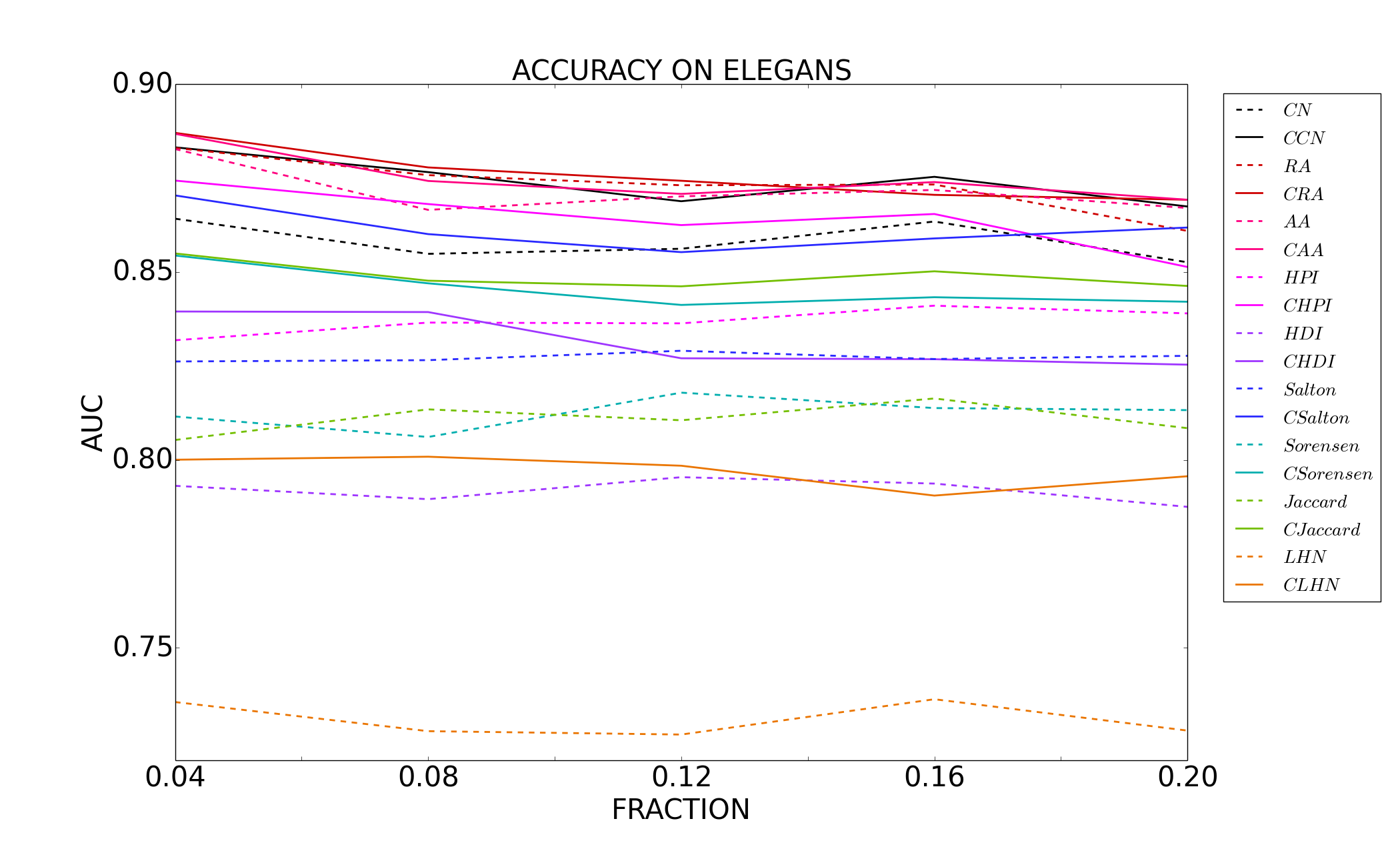


图 3.1（b） Elegans网络错误边边发现

(2)Football 网络的实验结果如图3.2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。

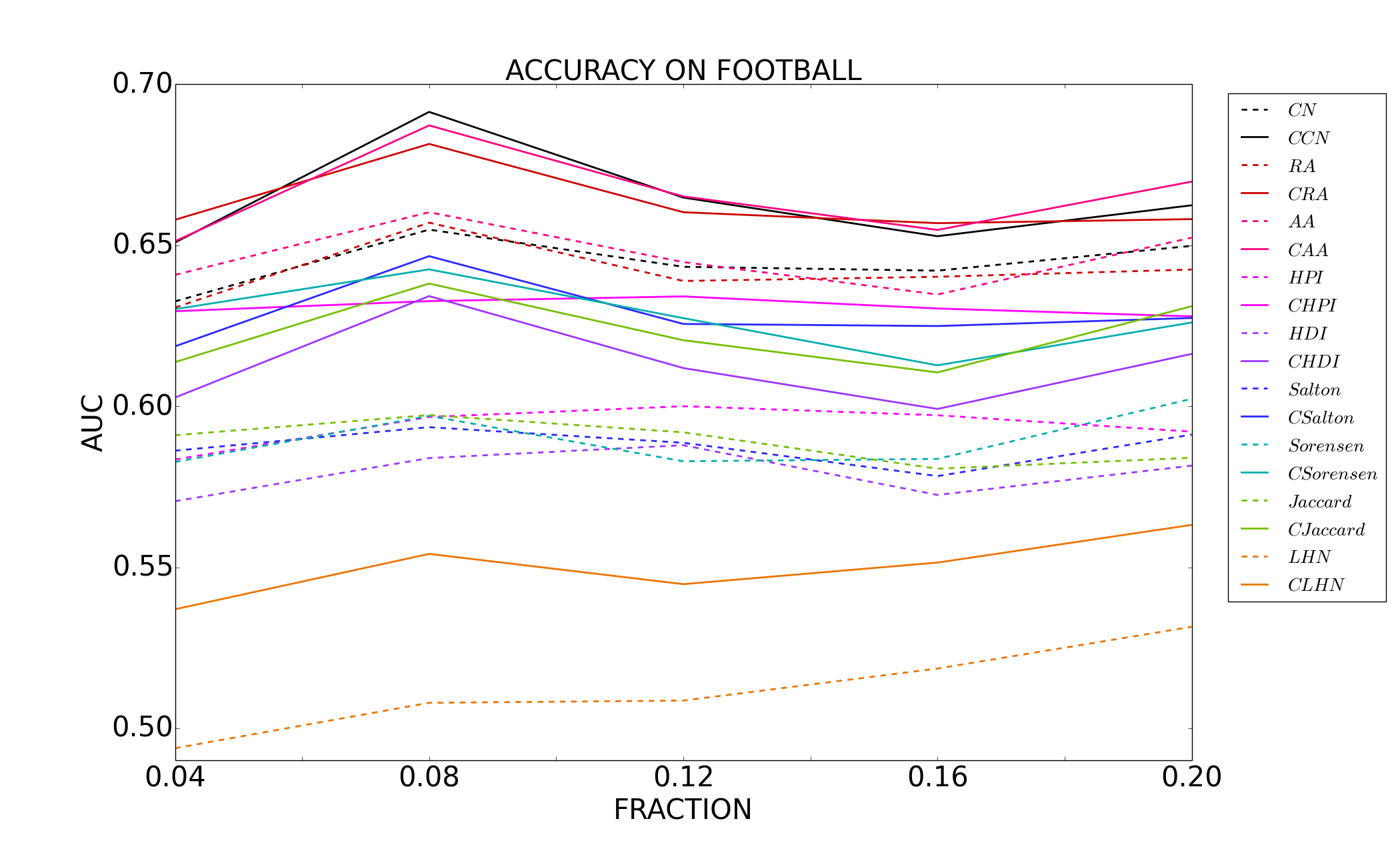


图 3.2（a） Football网络缺失边发现

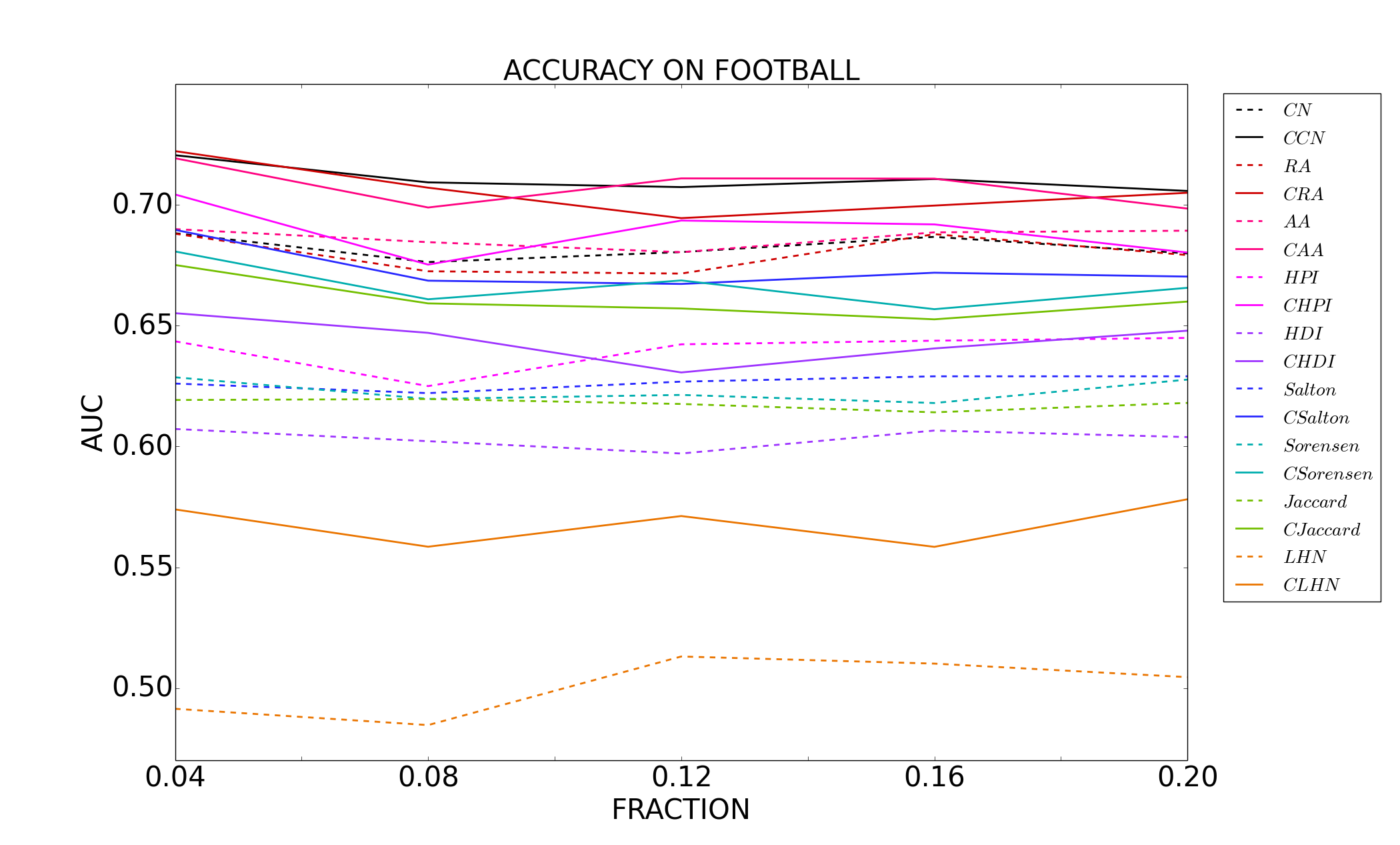


图 3.2（b） Football网络错误边发现

(3)Jazz 网络的实验结果如图3.3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。

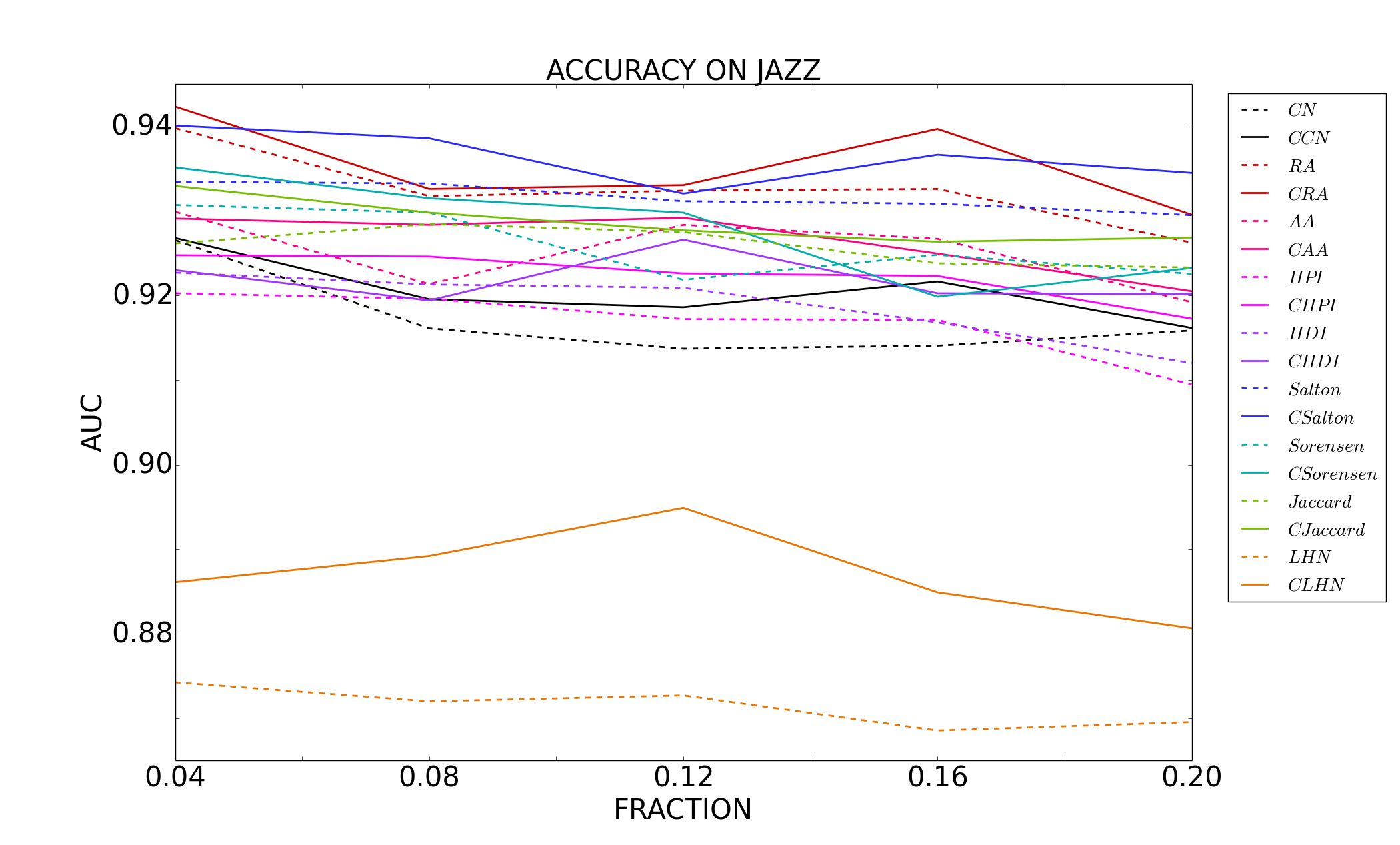


图 3.3（a）Jazz网络缺失边发现

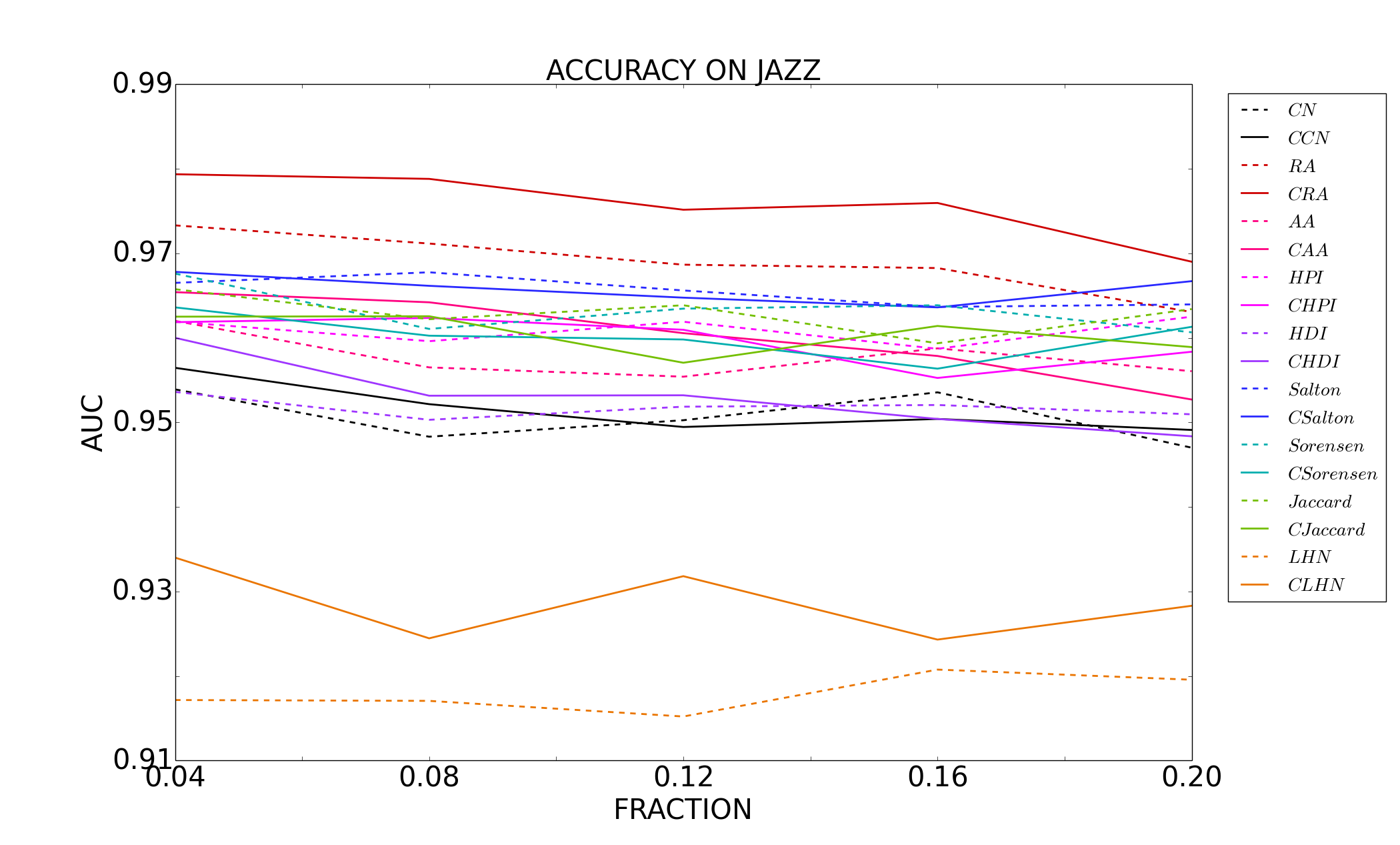


图 3.3（a）Jazz网络错误边发现

(4)USAir网络的实验结果如图3.4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，该 9 种算法在美国航空网络的预测上效果明显提升。

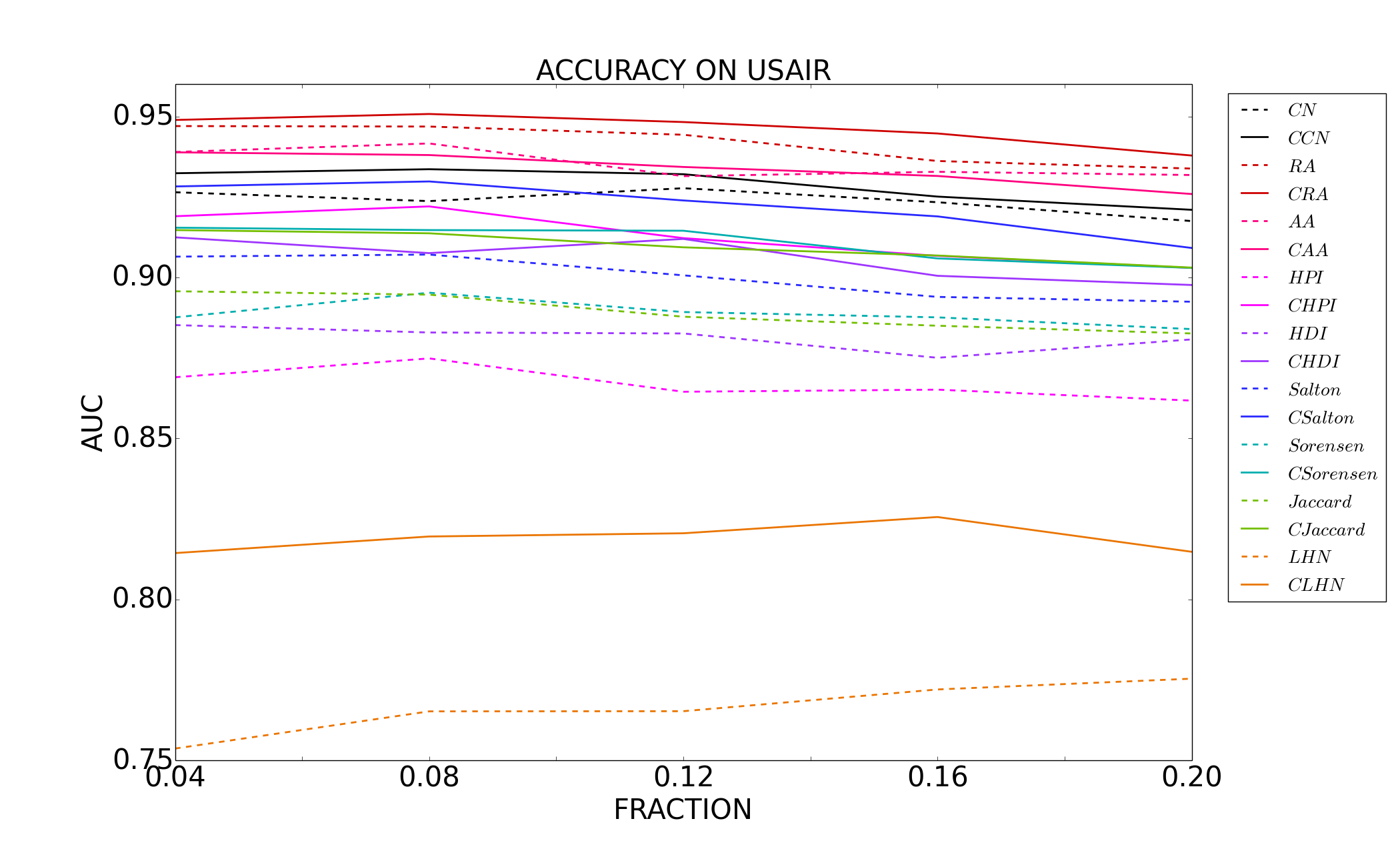


图 3.4（a）USAir网络缺失边发现

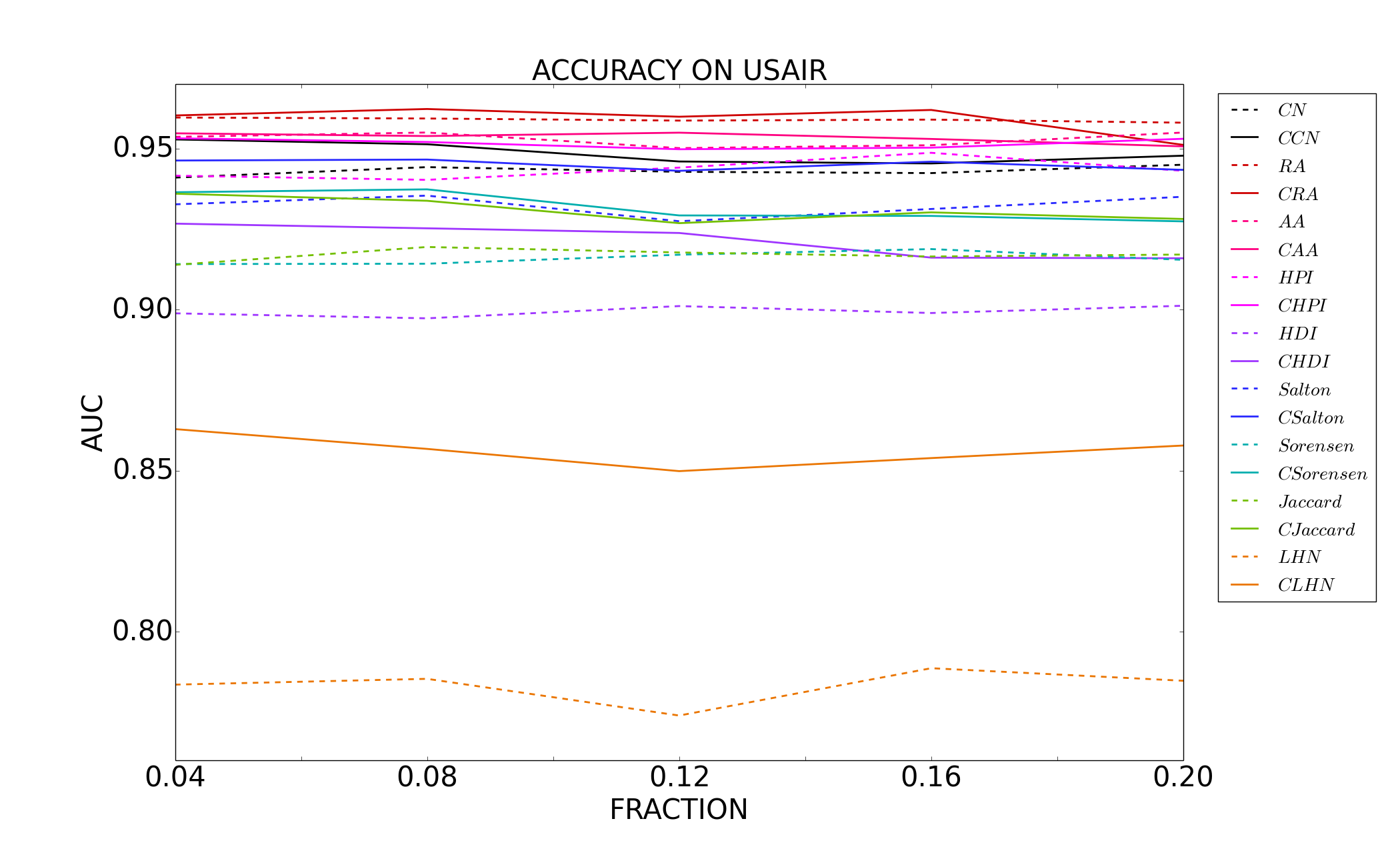


图 3.4（b）USAir网络错误边发现

(5)Karate 网络的实验结果如图3.5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。

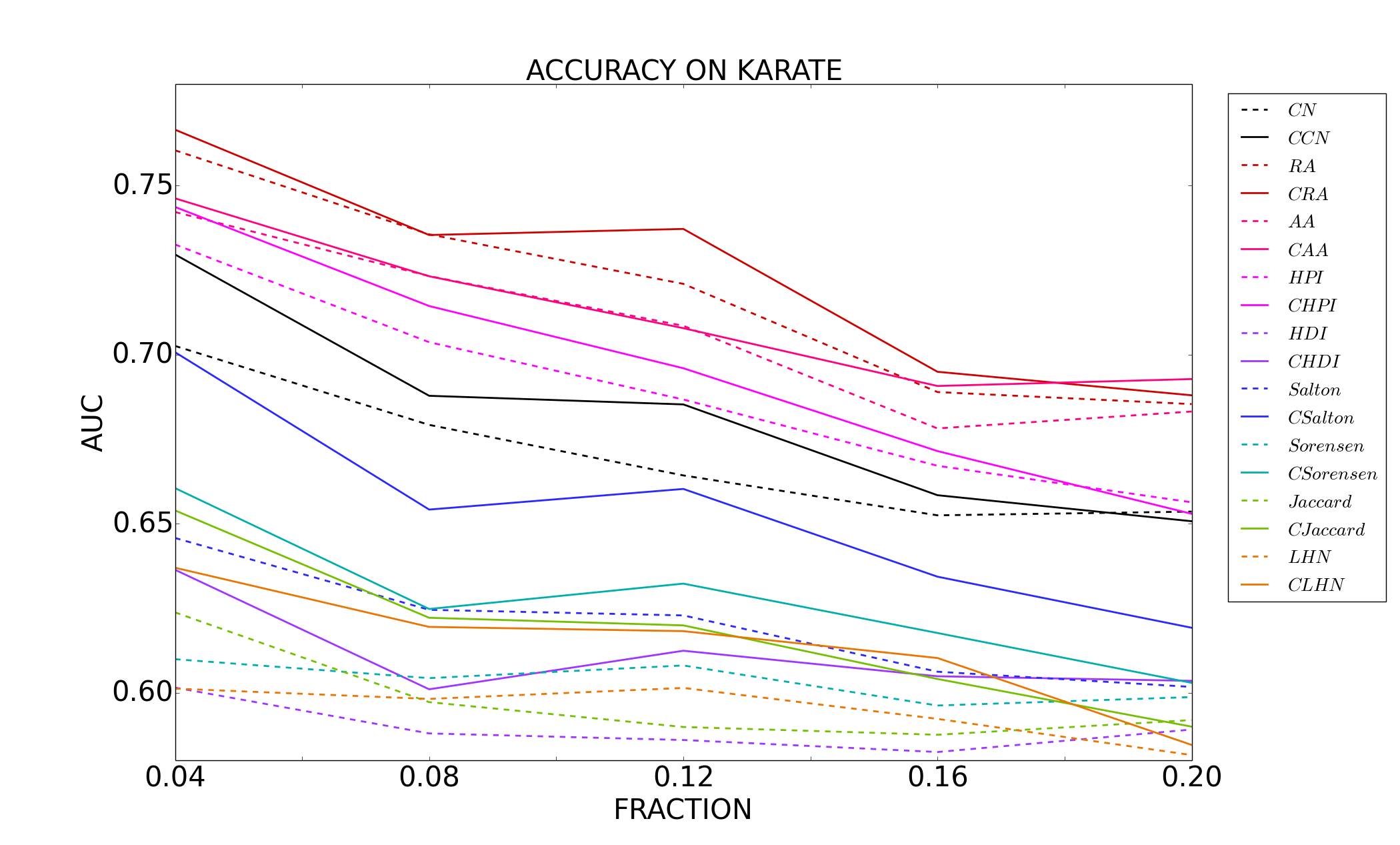


图 3.5（a）Karate网络缺失边发现

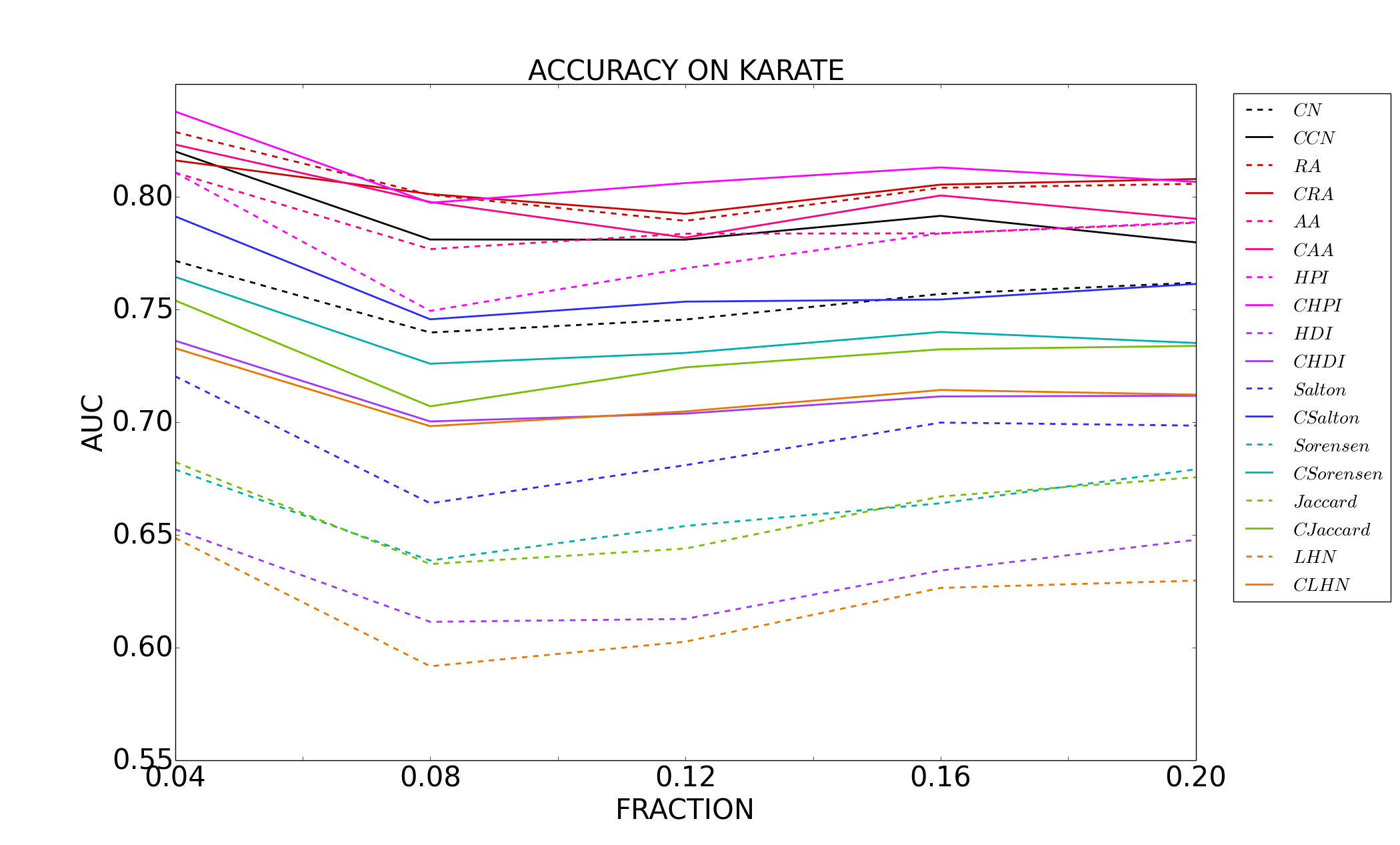


图 3.5（b）Karate网络错误边发现

## 3.5 本章小结

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。实验部分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。

# 第4章 交通拥堵恢复方法分析

本章主要介绍了我们所需要解决的问题及解决思路。为了验证第五章所提出的恢复方案，本章介绍了在交通流量模拟上一个广泛使用的模型。

## 4.1 问题描述

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。

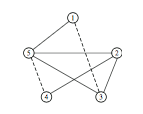


图 4.1 交通拥堵示例

## 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，我们队这些边进行评分，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，我们不会讲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[46]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 Ci，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 Ci。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。因此，数据包的路径选择在交通过程中也是很重要的。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。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，我们使用一个参量μ(R)来表示上面这段描述。

### 4.3.2 衡量指标μ

参数μ(R)的定义如下公式所示，

𝜇(𝑅) = （4.1）

在这里 W(t)代表在时间点 t 时网络中的数据包数量，ΔW=W(t+Δt)-W(t)，<...>代表长度为 Δt 的时间窗口内的平均值。换句话说，参数μ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μ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μ接近 1。

下面所展示的图片是我们从某此实验中记录下来的模拟过程中的μ的变化情况。我们看 到一开始μ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，μ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μ值大约为 0.07 左右，即网络中会残存 7%的数据无法处理。

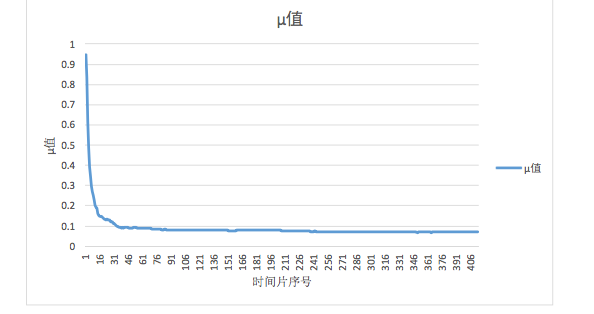


图 4.2 μ值变化(R=100，S=1.8，N=20)

随着 R 值得增长，交通网络的压力会越来越大，μ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μ值从 0 变到非 0 的临界 R 值来代表网络的吞吐量，记做 Rc。也就是说，Rc 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，Rc 越大代表交通网络的效率越高。

这里存在一个问题，R 值与μ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 μ值也就越大，但实际上临近 Rc 附近的 R 值所对应的μ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 R1 < R2，R1 所对应的μ1 并不一定小于 R2对应的μ2。这就造成了我们在寻找 Rc 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 本章小结

本章描述了交通网络拥堵恢复序列这样一个问题，紧接着我们提出了基于链路预测思维的解决思路，第五章的实验部分也将比较“反相关性”恢复算法、链路预测算法、随机恢复方法的效果。本章最后一部分介绍了现在在交通网络上的一种主流模拟方式，该方式以拥塞率这样一个指标来给出一个量化的比较标准。

# 第5章 交通拥堵恢复序列算法

本章首先介绍了7种链路预测相关的交通拥塞恢复序列的预测算法，根据不同的算法我们将得到不同的恢复序列。之后我们介绍了实验部分所用到的三类交通网络数据，这三类数据分别是美国航空网络、加州路网以及我们生成的模拟交通网络。实验部分，我们在三类网络之上对7种恢复算法以及随机恢复方式进行了比较。

## 5.1 算法描述

### 5.1.1 PA算法（Preferential Attachment Index）[47]

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，新增加的边连接节点vx和vy的可能性正比于vx和vy度的乘积。

Sxy = kx \* ky （5.1）

### 5.1.2 LP算法（Local Path Index） [48,49]

局部路径算法(local path)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α 参数用于调控三 阶邻居影响的大小。显然，当 α = 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。， ( An )xy 表示节点vx和vy之间路径长度为n的线路数目，A2矩阵中第x行第y列的元素表示的是节点vx到节点vy之间长度为2的路径数目，即vx和vy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。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。

S = A2 + εA3 （5.2）

### 5.1.3 SA算法（Three order Similarity Index）

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。Γ3(x)表示的是节点 x的一阶邻居、二阶邻居、三阶邻居的合集，Sxy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，力求进一步极高链路预测的准确性。

Sxy = |Γ3(x) ∩ Γ3(y)| （5.3）

**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。**

### 5.1.4 RPA算法（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。

Sxy = （5.4）

### 5.1.5 RAPA算法（Reverse Add Preferential Attachment Index）

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。

Sxy = （5.5）

### 5.1.6 RLP算法（Reverse Local Path Index）

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 = （5.6）

### 5.1.7 RSA算法（Reverse Three order Similarity Index）

与SA算法向对应，将SA算法所计算的分值求导。

Sxy = （5.7）

## 

## 5.2 数据集描述

### 5.2.1 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。现对这三类数据做简要的介绍。

#### 5.2.1.1 BA网络模型

BA模型[50]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。也就是俗话所说的“富者愈富，穷者愈穷”。它由Albert- László Barabási 和Réka Albert提出，因此被称为BA模型。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) ∝ 𝑘k-3 （5.8）

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：

𝑝𝑖 = （5.9）

公式中pi代表每新增一个节点，新节点与点 i的相连的概率，ki代表点 i的度数，也就是 说pi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”。我们根据以上的公式， 设计出了一个BA网络的生成程序。

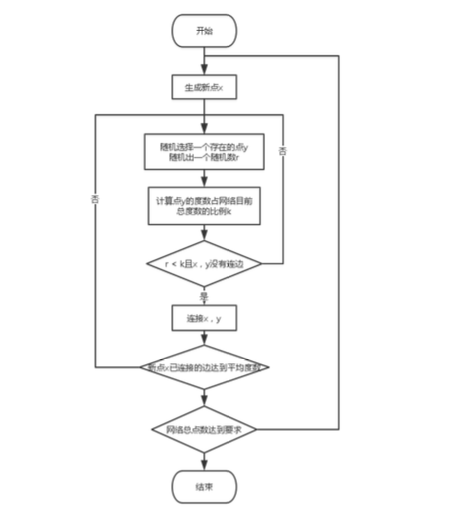
图5.1是我们生成BA网络时的程序流程图，我们在将新生成的点 x 与已经存在的点 y 进行连边时，随机出一个数字 r，然后计算点 y 的度数在网络中所有度中所占的比例，当我们随机出的数字 r 大于 y 的度数所占的比例时才将点 x 与点 y 相连， 这样一来，我们就实现了度数越大的节点连接新边的可能性就越大的要求，当r不大于y的度数所占比例时，就重新随机。如此循环，直到新点x连接了之前设定的平均度数的边，然后就 生成下一个节点，最后生成完整个 BA 网络。这样就实现了节点度越大的店，就越有可能聚集新边，到最后达到少数节点聚集了大量边，大多数节点只有很少边的情况。 

图 5.1 BA网络的生成流程

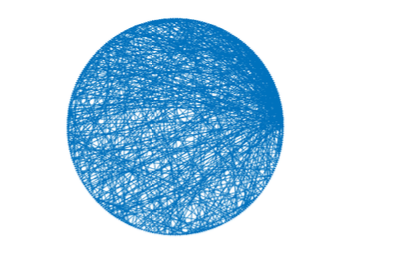


图 5.2 BA网络样例

图5.2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。 而图5.3是取自图5.2右上角的一部分，我们可以看到，这个节点有 15条边，远远 高出我们设定的每个点的平均度数。

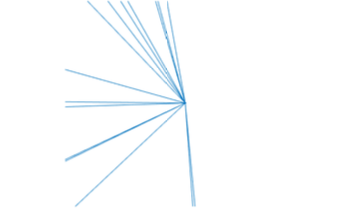


图 5.3 BA网络右上方大度节点取样

图 5.4是从图 5.3下方放大得到的一部分，我们发现这个区域中大部分节点的度数为 2或 3，低于平均度数 4，也就是说它只连接了自己的自己生成时的两条边。结合图 5.3所示的情 况，目前的网络是符合 BA网络的特性的:少数节点拥有大量连边，而大多数节点只拥有少量连边。

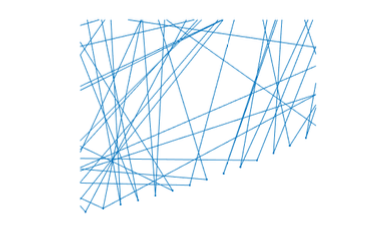


图 5.4 BA网络下方小度节点取样

#### 5.2.1.2 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。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。

#### 5.2.1.3 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。

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。

### 5.2.2 网络特征统计

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。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网络特征统计

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 网络 | |V| | | |E| | D | C | <k> | <d> | H |
| BA256 | 256  297 | | 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.

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。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，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。

## 5.3 实验流程

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。

如图5.5 所示，实验部分主要分为5个大的步骤。

**计算**

**结果**

**流量**

**模拟**

**恢复**

**序列**

**删除**

**边**

**创建**

**网络**

图 5.5 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（node1ID, node2ID）,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。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，具体过程见第4章第3节所提到的流量模型。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5.6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，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。重复上述过程，直至添加完成所有的待修复的边。

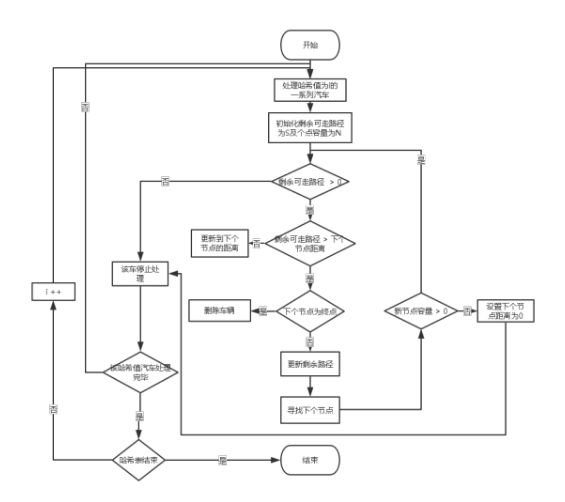


图 5.6 每个时间片内对数据包的处理流程

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。

流量模拟实验的大致流程如图5.7所示。

## 5.4 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μ < 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。

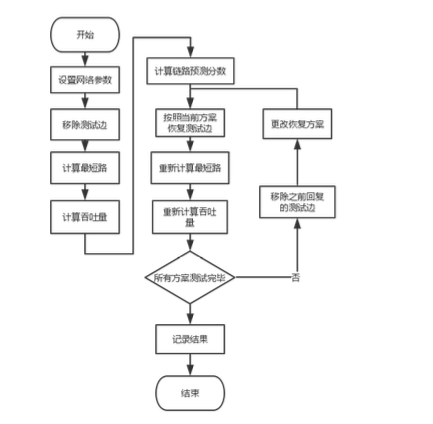


图 5.7 实验程序流程图

## 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 5.8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。

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。

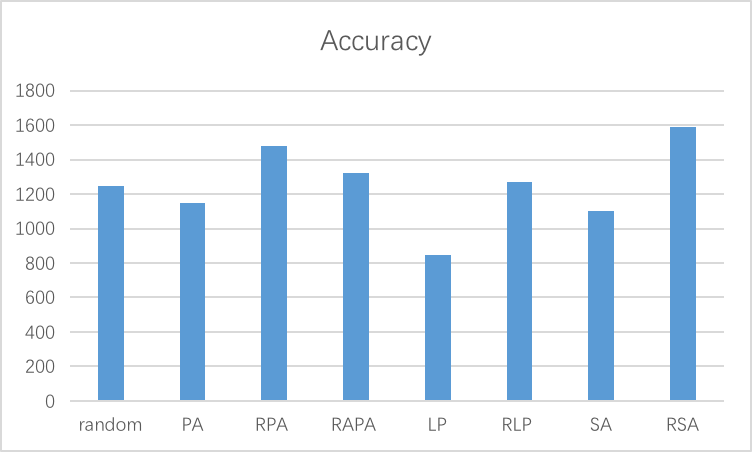


图 5.8 加州路网上的实验结果

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。

### 5.5.2 不同网络损毁比例对于实验的影响

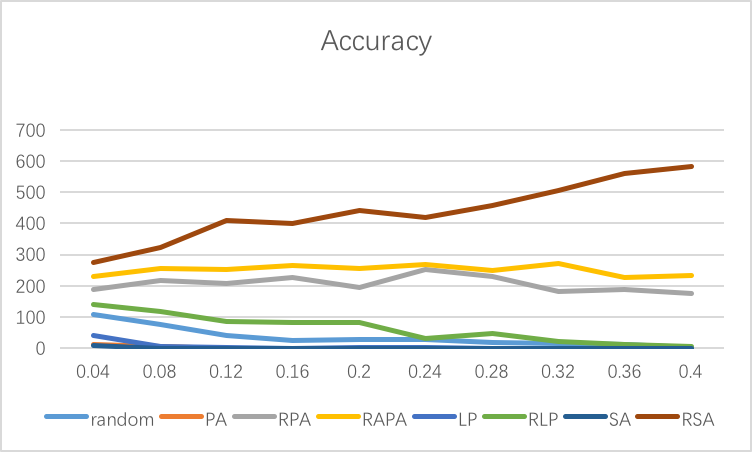


图 5.9 BA网络上不同拥堵率对实验结果的影响

图5.9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int(0.04\*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。

### 5.5.3 不同网络尺寸对于实验的影响

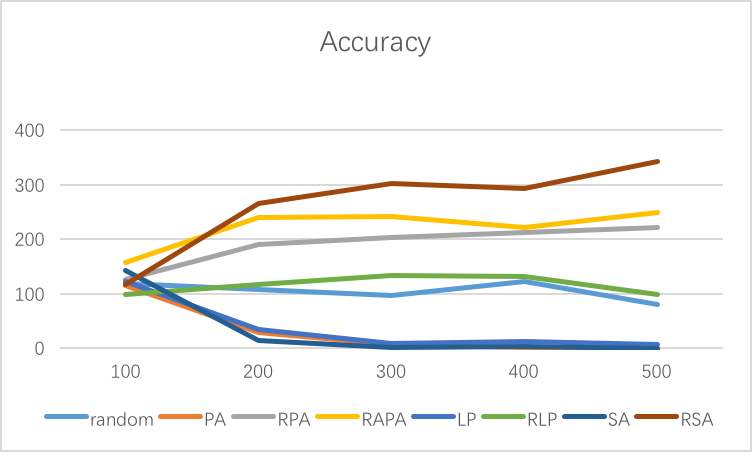


图 5.10 不同尺寸的BA网络对实验结果的影响

图5.10与5.9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。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。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。这是因为在较小的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的出来，各个节点的度分布差异也将越来越巨大。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。

### 5.5.4 不同大小的流量对于实验的影响

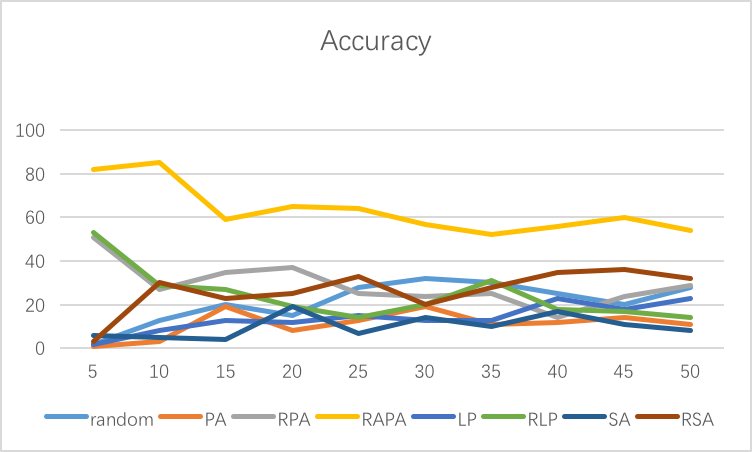


图 5.11 美国航空网络上不同大小的流量对于实验结果的影响

图 5.11 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。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。

## 5.6 本章小结

本章介绍了3种链路预测算法（其中SA算法是本文首次提出）并基于此3种算法变换得到的4种“负相关性算法”。本章将该7中算法加上随机恢复方法用于3类交通网络的拥堵序列恢复实验之中。

实验结果表明，根据链路预测思维变换得到的“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

本文主要做了两方面的研究，首先是提出了基于聚类系数的链路预测算法。其次，基于链路预测的启发，我们运用链路预测的结构相似性思维，提出了交通拥堵恢复序列的方法。

在基于聚类系数的链路预测算法研究中，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中新的算法。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。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。它们是把局域的算法往更高阶的路径上拓展一步，与全局域算 法相比，大大减少了运算量;与局域算法相比，预测精度又有了一定程度的提高。

在交通拥堵恢复序列的研究中，本文提出了基于链路预测的算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维的“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

在链路预测算法方面，本研究只简单的考虑了节点的聚类系数对于链路预测算法的影响，除此之 外，在未来的研究当中我们可将更多的网络特征融入到算法之中来进行算法的 升。本次实验的网络只有 5 种，更多数据的实验有待补充，我们不排除特定网络 之上，改进后算法的准确率不如原算法。本次实验网络的节点只有几百个，更大 规模网络的链路预测研究有待进一步深入。

在交通拥堵恢复序列方面，本文是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，故将很多实际情况略去了。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，所以可以说实验是半模拟的过程，随着越来越多交通数据被公布出来，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。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。本文首次提出了一种基于链路思维的交通拥堵恢复序列的预测，在此模型的基础上更多与真实场景相关的因素待进一步考虑。

致谢

转眼间，三年的研究生生涯即将结束。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。

研究生期间，我首先要感谢的是我的导师关佶红教授。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。

感谢来参加我研究生毕业论文答辩的各位师兄弟姐妹。

感谢徐伟与王佳晟同学，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。

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！

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

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

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

[4] 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.

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

[6] Milo R, Shen-Orr S, Itzkovitz S et al. Network motif: simple building blocks of complex networks. Science, 2002, 298: 824-827.

[7] Clauset A, Moore C, Newman M E J. Structure inference of hierarchies in networks. Proceedings of the 2006 conference on statistical network analysis.

[8]Vito Latora，Massimo Marchiori.Is the Boston subway a small world

network[J].Physica A 314(2002):109一113

[9]Vamsi Kalapala，Vishal Sanwalani，Aaron C1auset，et a1.Moore Scale

Invariance in Road Networks[J].arXiv:physics/0510198v2 [phys i CS.SOC—ph]8 Mar 2006

[10]Parongama Sen，Subinay Dasgupta，Arnab Chatterjee，et a1.

Small—world properties of the Indian Railway network[J]. arXi v:cond—mat/0208535 v2 3 1 Dec 2002

[11]R.Guimera，S.Mossa，A.Turtschi，et a1.The worldwide air transportation network:AnomalOUS centrality，community structure， and cities’global roles[J].LAN Amaral—Proceedings of the National Academy of Sciences，2005一National Acad Sciences

[12]陆化普.解析城市交通[M].北京:中国水利水电出版社，2001.9 [13]National Cooperative Highway Research Program.The Benefits of

Reducing Congestion[C].Cambridge Systematics，Inc.2002.1

[14]2004年中国汽车工业年鉴。中国汽车技术研究中心.2004

[15]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.

[16] Erdös P, Rényi A. On Random Graphs[J]. Publicationes Mathematicae, 1959, 6:290-291.

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

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

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

[20] Gao F, Musial K, Cooper C, et al. Link prediction methods and their accuracy for different social networks and network metrics[J]. Scientific Programming, 2015, 2015: 1.

[21] Li D, Zhang Y, Xu Z, et al. Exploiting Information Diffusion Feature for Link Prediction in Sina Weibo[J]. Scientific reports, 2016, 6.

[22] 刘宏鲲, 吕琳媛, 周涛. 利用链路预测推断网络演化机制[J]. 中国科学: 物理学 力学 天文学, 2011, 41(7): 81

[23]  Traffic congestion in interconnected complex networks

[24] Traffic fluctuation on weighted networks

[25] T. Ohira, R. Sawatari, Phase transition in a computer network traffic model, Phys. Rev. E 58 (1998) 193195.

[26] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.- U. Hwang, Phys. Rep. 424, 175 (2006).

[27] [137] Z. Toroczkai, K.E. Bassler, Nature 428 (2004) 716.

[28] [138] Z. Toroczkai, B. Kozma, K.E. Bassler, N.W. Hengartner, G. Korniss, preprint cond-mat/0408262.

[29] R. Guimer‘a, A. Arenas, A. D az-Guilera, and F. Giralt, Phys. Rev. E 66, 026704 (2002).

[30] G.-Q. Zhang, S. Zhou, D. Wang, G. Yan, and G.-Q. Zhang, Physica A 390, 387 (2011).

[31] L. Zhao, Y.-C. Lai, K. Park, and N. Ye, Phys. Rev. E 71, 026125 (2005).

[32]  D. De Martino, L. DallAsta, G. Bianconi, and M. Mar- sili, Physical Review E 79, 015101 (2009).

[33] G. Yan, T. Zhou, B. Hu, Z.-Q. Fu, and B.-H. Wang, Phys. Rev. E 73, 046108 (2006).

[34] X. Ling, M.-B. Hu, R. Jiang, and Q.-S. Wu, Phys. Rev. E 81, 016113 (2010).

[35] F. Tan and Y. Xia, Physica A 392, 4146 (2013).

[36] B. Danila, Y. Yu, J. A. Marsh, and K. E. Bassler, Phys. Rev. E 74, 046106 (2006).

[37] X. Ling, M.-B. Hu, J.-C. Long, J.-X. Ding, and Q. Shi, Chin. Phys. B 22, 018904 (2013).

[38] Lorrain F, White H C. Structural equivalence of individuals in social networks[J]. The Journal of mathematical sociology, 1971, 1(1): 49-80.

[39] Salton G, McGill M J. Introduction to modern information retrieval[J]. 1986.

[40] Jaccard P. Etude comparative de la distribution florale dans une portion des Alpes et du Jura[M]. Impr. Corbaz, 1901.

[41] Sorensen T. {A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons}[J]. Biol. Skr., 1948, 5: 1-34.

[42] Ravasz E, Somera A L, Mongru D A, et al. Hierarchical organization of modularity in metabolic networks[J]. science, 2002, 297(5586): 1551-1555.

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

[44] Leicht E A, Holme P, Newman M E J. Vertex similarity in networks[J]. Physical Review E, 2006, 73(2): 026120.

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

[46] 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).

[47]  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.

[48]  Zhou T, L L, Zhang Y C. Predicting missing links via local information[J].

[49]  The European Physical Journal B, 2009, 71(4):623-630. Ackland R, Ackland R. Mapping the U.S. Political Blogosphere:  Are Conservative Bloggers More Prominent?[C] 2005.

[50]  Albert-László Barabási and Réka Albert. Statistical mechanics of complex networks[J]. Reviews of  modern physics, 2002, 74(1): 47-94

**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01月生。

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。

2014年9月入同济大学就读硕士研究生。

**已发表论文：**

**参与科研项目：**