****

**硕士学位论文**

**链路预测及交通恢复序列的预测**

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

二〇一七年三月

****

A dissertation submitted to

Tongji University in conformity with the requirements for

the degree of Master

**Link Prediction and Traffic 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

March, 2016

|  |
| --- |
| **链路预测与交通恢复序列的**  **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

**关键词**：

**ABSTRACT**

**Key Words:**

目录

[第1章 引言 4](#_Toc438380793)

[1.1 研究背景与意义 4](#_Toc438380794)

[1.2 国内外研究现状 6](#_Toc438380795)

[1.3 本文主要研究工作 7](#_Toc438380796)

[1.4 论文组织结构 8](#_Toc438380797)

[第2章 手势密码触屏行为模式构建的关键技术 10](#_Toc438380798)

[2.1 距离计算算法 10](#_Toc438380799)

[2.1.1 曼哈顿距离 10](#_Toc438380800)

[2.1.2 动态事件规整算法（Dynamic Tine Warping, DTW） 10](#_Toc438380801)

[2.2 分类算法 12](#_Toc438380802)

[第3章 数据采集和预处理 15](#_Toc438380803)

[3.1 数据采集 15](#_Toc438380804)

[3.1.1 数据采集器 15](#_Toc438380805)

[3.1.2 数据集 16](#_Toc438380806)

[3.2 数据预处理 16](#_Toc438380807)

[3.2.1 数据清理 16](#_Toc438380808)

[3.2.2 数据变换 17](#_Toc438380809)

[3.2.3 归一化 17](#_Toc438380810)

[3.3 特征提取 18](#_Toc438380811)

[第4章手势密码触屏行为模式的构建和认证 19](#_Toc438380812)

[4.1 总体架构 19](#_Toc438380813)

[4.1.1 方法总体架构 19](#_Toc438380814)

[4.1.2 概念定义 20](#_Toc438380815)

[4.2 特征距离向量集构建 21](#_Toc438380816)

[4.2.1 参照特征向量集生成 21](#_Toc438380817)

[4.2.2 特征距离向量集生成 22](#_Toc438380818)

[4.3 分类器预测和训练 22](#_Toc438380819)

[4.4 实验分析 24](#_Toc438380820)

[4.4.1 实验环境和数据 24](#_Toc438380821)

[4.4.2 实验结果分析 24](#_Toc438380822)

[第5章 多用户单账号场景下的手势密码触屏行为模式的构建 27](#_Toc438380823)

[5.1 多用户单账号场景分析 27](#_Toc438380825)

[5.2 多用户单账号手势密码触屏行为模式的构建 27](#_Toc438380826)

[5.2.1 总体架构 27](#_Toc438380827)

[5.2.2 多用户聚类 29](#_Toc438380828)

[5.2.3 手势模型构建 30](#_Toc438380829)

[5.3 实验分析 30](#_Toc438380830)

[第6章 总结与展望 34](#_Toc438380831)

[6.1 总结 34](#_Toc438380832)

[6.2 进一步工作 35](#_Toc438380833)

[致谢 36](#_Toc438380834)

[参考文献 37](#_Toc438380835)

[**个人简历、在读期间发表的学术论文与研究成果** 40](#_Toc438380836)

# 第1章 绪论

随着移动计算能力以及智能手机占有量的指数级增长，触屏智能手机已经成为每个人日常生活不可替代的重要组成部分，如何正确识别移动终端用户身份的合法性是保障用户信息和财产安全的关键。传统的安全策略存在的局限性使其难以满足移动终端用户对于安全性的需求，用户触屏行为的研究和应用逐渐成为用户身份认证领域的热点。手势密码触屏行为认证通过采集用户完成手势密码时的触屏行为数据，对该部分数据进行分析和建模，针对每个用户建立独有的手势密码触屏行为模式，作为身份认证的依据。

本章首先阐明论文的研究背景以及研究的意义与价值，然后对本文研究内容的研究现状进行阐述和分析，随后介绍本论文的研究思路与主要研究工作，最后介绍本文整体的组织结构和内容安排。

## 1.1 研究背景与意义

在真核生物中，DNA序列需要通过严格的分级组装的方式包装和高度压缩后形成染色质结构。首先，DNA序列和蛋白质相结合，形成核小体。核小体进一步压缩形成更加复杂的纤丝状的染色质结构。最后，纤丝状的染色质被进一步压缩，形成能够被包装进细胞核的染色体结构。核小体是真核生物染色质的基本组成单位，由DNA缠绕在组蛋白上构成，各核小体串联而成染色质纤维。H2A、H2B、H3和H4这四种组蛋白，每一种组蛋白各二个分子，形成一个[组蛋白八聚体](http://baike.baidu.com/view/3872893.htm)，约200bp的DNA分子盘绕在组蛋白八聚体构成的核心结构外面，形成了一个核小体。其中，缠绕在组蛋白八聚体周围的核心 DNA（core DNA）约1.65 圈，约合147个碱基对；而相邻的核小体之间的自由区域（linker DNA）约为20～50个碱基的长度。核小体的分布影响了DNA的可接入性，因此在诸如转录调控、DNA复制和修复等过程中都扮演着重要角色。核小体是表观遗传机制中非常重要的组成部分。然而，核小体的形成机制仍然是一个未知的问题。核小体的形成和分布并不是随机的。要明确核小体对基因调控的影响，必须要先掌握核小体的分布规律。如何精确进行核小体定位是生物信息学一个非常热门的研究问题。近年来，对核小体定位的研究得到广泛关注，提出了很多核小体预测和定位的算法。然而如何高效准确地进行核小体定位依然是需要深入研究的问题。

目前，研究者们主要通过实验的方法来检测核小体的位置。实验方法的根本思路是用微球菌核酸酶来分解和消化两个核小体之间的连接DNA，从而得到互相分离的核小体，并且根据标识的位置进行核小体的定位。近几年，随着科技的发展，实验技术的提升，出现了很多其他的核小体定位的实验技术，例如染色质免疫共沉淀-芯片（ChIP-chip）、染色质免疫共沉淀-测序（ChIP-Seq）等。随着实验方法的发展，核小体的定位变得更加容易实现，准确度也更高。但是，这种通过实验技术来获取全基因组的核小体定位信息的方式需要大量的时间和经费。此外，随着更多不同物种的核小体位置信息的增加，如何存储通过实验方法得到的核小体的位置信息也成为研究者们面临的巨大难题。近几年来，Segal研究小组提出了基于序列特征的核小体定位理论预测模型，发现DNA序列能够为核小体定位提供重要的信息。此外，还有很多研究者发现，DNA的某些结构特征也和核小体密切相关。所以，对实验所获得的数据进行分析，使用计算方法来进行核小体定位对弥补生物实验技术的缺陷，辅助生物实验进行具有重要意义，对基因组中的核小体定位的理论预测已经成为生物信息领域热门研究课题。

核小体的分布影响了DNA的可接入性，因此在诸如转录调控、DNA 复制和修复等过程中都扮演着重要角色。在染色体的任何时期，核小体的结构都是不变的。核小体是细胞内基因调控的重要环节，是表观遗传机制的重要组成部分。核小体的定位预测有望为创新药物研究和临床医学带来实际的有价值的突破，具有很好的实际应用价值。

本课题旨在找到更高效更准确的核小体定位算法，提高预测和定位精度，解释某些特定的基因调控现象，为表观遗传机制提供一定的参考意义。

## 1.2 国内外研究现状

在真核生物中，DNA序列需要通过严格的分级组装的方式包装和高度压缩后形成染色质结构。首先，DNA序列和蛋白质相结合，形成核小体。核小体进一步压缩形成更加复杂的纤丝状的染色质结构。最后，纤丝状的染色质被进一步压缩，形成能够被包装进细胞核的染色体结构。核小体是真核生物染色质的基本组成单位，由DNA缠绕在组蛋白上构成，各核小体串联而成染色质纤维。H2A, H2B, H3和H4这四种组蛋白， 每一种组蛋白各二个分子，形成一个[组蛋白八聚体](http://baike.baidu.com/view/3872893.htm" \t "_blank)，约200 bp的DNA分子盘绕在组蛋白八聚体构成的核心结构外面，形成了一个核小体。其中，缠绕在组蛋白八聚体周围的核心 DNA （core DNA） 约 1.65 圈，约合147 个碱基对；而相邻的核小体之间的自由区域（linker DNA） 约为 20～50 个碱基的长度。核小体的分布影响了DNA的可接入性，因此在诸如转录调控、DNA 复制和修复等过程中都扮演着重要角色。然而，核小体定位是一个涉及 DNA、转录因子、组蛋白修饰酶和染色质重塑复合体等分子间相互作用的复杂过程。核小体的形成机制仍然是一个未知的问题。如何精确进行核小体定位是一个非常热门的研究问题。目前，研究者们主要通过实验的方法来预测核小体的位置。实验方法的根本思路是用微球菌核酸酶来分解和消化两个核小体之间的连接DNA，从而得到互相分离的核小体，并且根据标识的位置进行核小体的定位。近几年，随着科技的发展，实验技术的提升，出现了很多其他的核小体定位的实验技术，例如染色质免疫共沉淀-芯片（ChIP-chip）、染色质免疫共沉淀-测序（ChIP-Seq）等。此外，近年来，在实验数据的基础上，研究者们使用计算方法来进行核小体定位弥补生物实验技术的缺陷。在众多的影响核小体形成的因素中，包括染色质重塑子，特定的DNA结合蛋白，DNA本身的序列等已经成为近期生物信息领域的研究热点。目前已有的定位方法主要基于两大类，基于序列的核小体定位方法和基于结构的核小体定位方法。其中，基于DNA基于序列的研究方法有， Segal研究小组、Kaplan研究小组等小组提出的定位算法[1]-[7]。Trifonov等人首次将含AT和GC的二核苷酸和三核苷酸用来进行核小体的定位[8]。随后,有更多的研究学者发现呈现周期性的序列特征与核小体的形成有密切的联系。例如，G和C碱基的含量和体外核小体的分布的相关性高达50%[3]。基于序列特征的计算方法被越来越多的学者用来进行核小体的预测。但是，随后也有研究证明，DNA的序列特征并不是决定核小体组织形成的决定性因素[9]-[10]。之后，有学者就提出问题，DNA的结构化特征是否对核小体的形成有影响[11]-[13]。为了解决这个问题，一些研究学者开始对DNA序列的结构化特征和核小体的形成机制进行研究。核小体DNA数据库的一些物理化学性质，例如DNA蛋白质复合体的倾斜度和螺旋弯曲度已经被证实与核小体的绑定有显著关系[14]。基于roll-slide 模型，Tolstorukov等人发现，相邻碱基对的滑动对DNA超螺旋节距有很大的影响力，相邻碱基对的弯曲度对DNA的弯曲度起决定性作用。Miele等人发现基于二核苷酸的DNA弹性和固有的弯曲度可以用来分析核小体[13]。Morozov等人使用DNA伸缩能量函数构造了基于序列的核小体形成的生物物理学模型[15]。DNA晶体结构的二核苷酸的弯曲度也被发现和核小体DNA片段有关联[16]-[17]。对核小体晶体结构的研究分析显示碱基对的行为、核糖环的折叠和支柱扭曲共同影响着核小体DNA序列的结构变异[18]。这些研究表明，可能还存在着很多与核小体相关的结构化特征。因此，系统化的分析不同的结构化特征和区别对核小体形成有关的结构化特征变得非常重要，以便来理解核小体与基因序列的结构化特征之间确切的关系。越来越多的研究学者开始利用这些和核小体相关的结构化特征提出更多有效的新方法来进行核小体的预测，并达到了满意的实验结果。但是目前的核小体定位算法精度都还较低，大约在50%左右。主要存在三方面的问题：1、没有找到较好的特征进行核小体的定位。2、通过构造特征向量然后采用机器学习中的分类算法的过程中，采用的机器学习算法都不能达到很高的准确性。3、通过构造特征图谱然后寻找峰值的过程中，峰值寻找算法还不能很好判断峰值位置的问题。

（1）特征寻找和选择

核小体特征分为两类，序列特征和结构特征。序列特征包括K-mer、周期性、GC含量等；结构特征包括DNA变性、DNA弯曲度、Z-DNA等。目前大部分基于生物计算的核小体定位算法要么只采用序列特征，要么只采用结构特征，忽略了序列特征和结构特征的相关性。目前，已有的特征选择的方法有嵌入式、过滤式、封装式三大模型。嵌入式模型是将特征选择过程作为组成部分嵌入到学习算法理，在学习训练的过程中选择合适的特征，学习过程结束时，学习算法所用到的特征即为特征选择的结果。经典的算法是决策树分类算法，如ID3、CART以及C4.5。过滤式模型通常作为学习算法的预处理手段，独立于具体的学习算法。常用的特性评价函数如信息增益、类间距离、相关系数等等。过滤式模型在大规模数据或在线数据的处理上具有明显的优势。封装式模型直接利用学习算法的准确率作为特征或特征子集重要性程度的评价标准，相关算法有贝叶斯网络、支持向量机等。过滤式模型和封装式模型都存在各自的优势和缺点。过滤式模型效率高，但是最终的学习准确率可能不高。而封装式模型学习效果好，但是容易出现过拟合现象。

（2）相关机器学习算法

一些已有的核小体定位算法中，有一些算法是通过构造特征向量然后采用机器学习中的分类算法对DNA序列进行分类，分为有核小体的位置和没有核小体的位置。已有的机器学习方法在核小体定位上的应用有SVM[19]、HMM[24]、高斯混合模型[23]、Boosting[21]等。例如，但是目前为止，这些算法的准确率都不是很高，一般准确率都在50%左右。这些机器学习算法并不能很好的处理生物数据,并未针生物数据做一定的修改和适应。仍有很多的算法并未应用在核小体的定位上，例如J48、Decision Stump等，很多算法值得尝试。

（3）峰值寻找算法

目前已有的很多核小体定位算法大多都是通过构造特征图谱然后寻找峰值或者峰谷的位置来进行核小体的定位，例如DLaNe[24]、All- atom[28]等等。但是在峰值寻找的过程中，都不能避免局部最优问题，因而无法精确找到所需要的峰值或者峰谷。峰值寻找算法的精度大大影响核小体定位的精度。

## 1.3 本文主要研究工作

本文主要研究

本文的主要研究内容及工作分为以下几个方面：

（1）特征寻找和选择

至今为止，热点的生物属性仍然没有被完全了解，所以当前用来识别热点的相关的属性特征仍然是不够的。在众多特征中，需要根据生物特性和相应的模型进行删选。在本课题中，以序列特征为主，结合结构特征，采取序列特征和结构特征相结合的方法，并采用过滤式特征选择方法，以皮尔逊相关系数作为过滤条件，选择和寻找生物特征。

（2）机器学习算法的选择

相关机器学习方法较多，需要研究如何选择和集成相关算法以提高定位算法效率和定位精度。目前的核小体定位算法中，大多只采用了一种机器学习算法。本课题中将实验更多其他的机器学习算法，并集成相关算法以提高预测准确率。

（3）峰值寻找算法的优化

峰值寻找算法还存在不能很好判断峰值位置的问题，尚需进一步优化。本课题主要研究内容在于找到一种更加精确的峰值定位算法，解决局部最优问题，使得能够找打实际的所需的峰值（或峰谷）。峰值寻找算法的精度大大影响核小体定位的精度。

## 1.4 论文组织结构

本文的组织结构如下所列：

第一章主要介绍了研究的背景和意义，以及现有的针对手势密码触屏行为特征和认证的研究现状和基础；

第二章介绍手势密码触屏行为模式构建和身份认证过程中涉及的关键技术和算法，包括计算特征向量间距离的曼哈顿算法和DTW算法，筛选特征的主成分分析方法，用于分类的SVM及相关的参数优化算法；

第三章从数据采集、预处理、特征构建这三面入手。详细描述了Android系统上手势密码触屏数据采集器的开发，对采集来的原始数据通过数据筛选、数据格式统一、归一化等方法进行预处理，并基于预处理后数据进行手势密码触屏行为特征的构建。

第四章介绍手势密码触屏行为模式的构建和认证。首先描述了行为模式构建的总体架构，然后通过参照集特征向量生成、特征距离向量集构建以及SVM分类器等算法构建手势密码触屏行为模式并进行认证效果验证。最后，通过实验分析来获取触屏行为模式构建方案的误检率和漏检率，明确模型的认证效果。

第五章介绍的是多用户单账号场景下的手势密码触屏行为模式构建。通过k-means聚类算法区分同一账户下的不同用户，然后对每一个用户使用第四章介绍的手势密码触屏行为模式构建方法进行建模。认证时，主要匹配该账户下的任一用户的行为模式，即可认为其为合法用户。

# 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。

## 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’,该边的两个端点分辨为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。

### 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[0.044,0.08,0.12.0.16,0.20]这五个不同的值。

## 2.3 经典的链路预测算法

本文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QQ，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。

Sxy = | Γ(y)|

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。

### 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。

Sxy =

### 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。

### 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。

Sxy =

### 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。

Sxy =

### 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。

Sxy =

### 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。

Sxy =

### 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。

Sxy =

### 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。

Sxy =

此处缺张图

## 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)

### 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=

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。

## 2.5 本章小结

本章回顾了问题描述、数据集划分、评价指标以及被改进的算法，下章着重讲解算法如何改进以及实验部分。

# 第3章 基于聚类系数的链路预测算法

## 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 =

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， 升算法对于部分网络的预测准确度。

此处应该有图解

## 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：

### CCN 指标(Clustering Common Neighbors)

Sxy =

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。

### CSalton 指标(Clustering Salton)

Sxy =

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。

### CJaccard 指标(Clustering Jaccard)

改进原理同CCN。

### CSorensen 指标(Clustering Sorensen)

Sxy =

改进原理同CCN。

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

Sxy =

改进原理同CCN。

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

Sxy =

改进原理同CCN。

### CLHN 指标(Clustering LHN)

Sxy =

改进原理同CCN。

### 8) CAA 指标(Clustering Adamic-Adar)

Sxy =

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。

### 9) CRA 指标(Clustering Resource Allocation)

Sxy =

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 网络特征的统计

表1. 网络的统计特征

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 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 |

|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

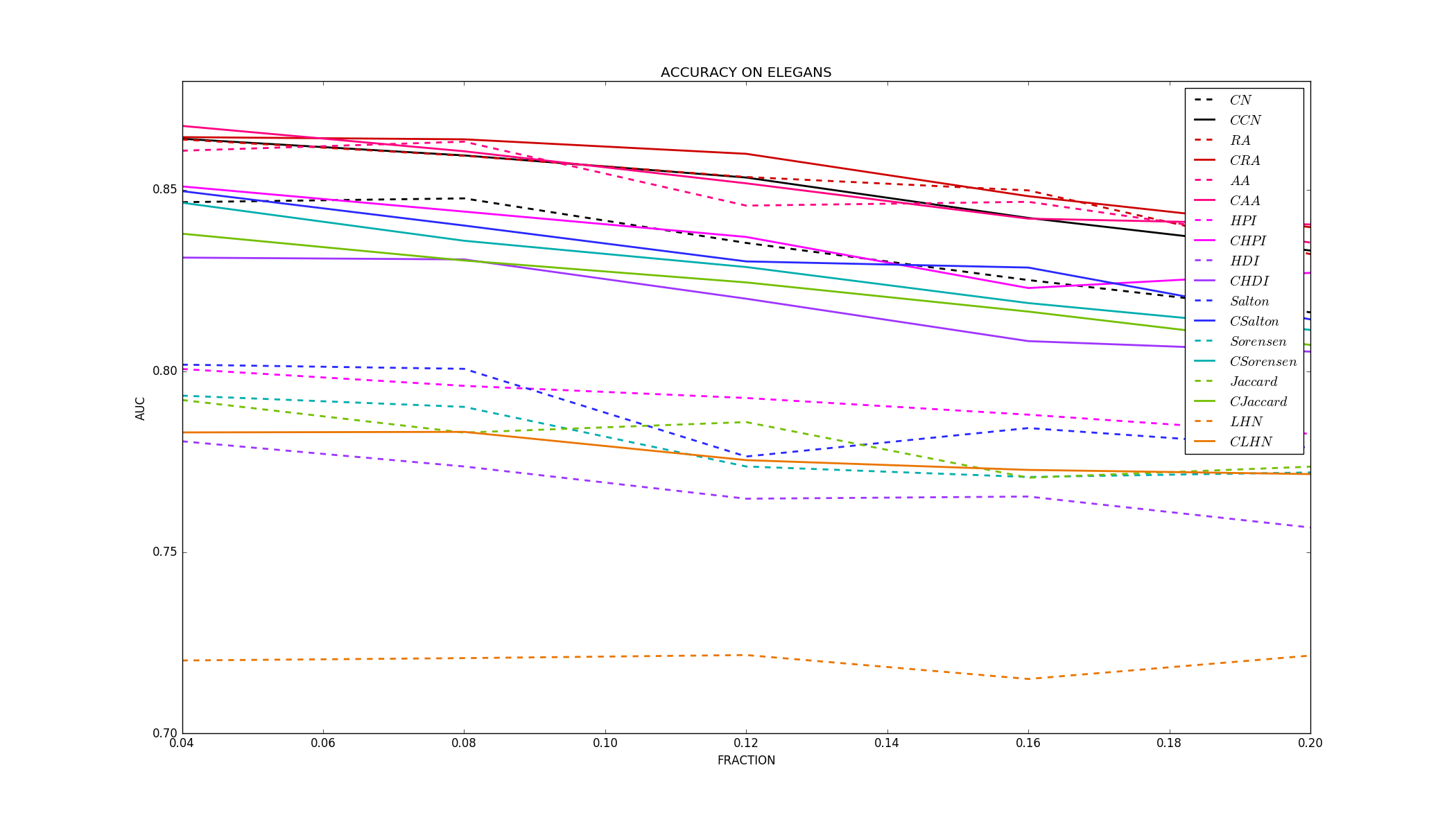
### 3.4.1 参数设定

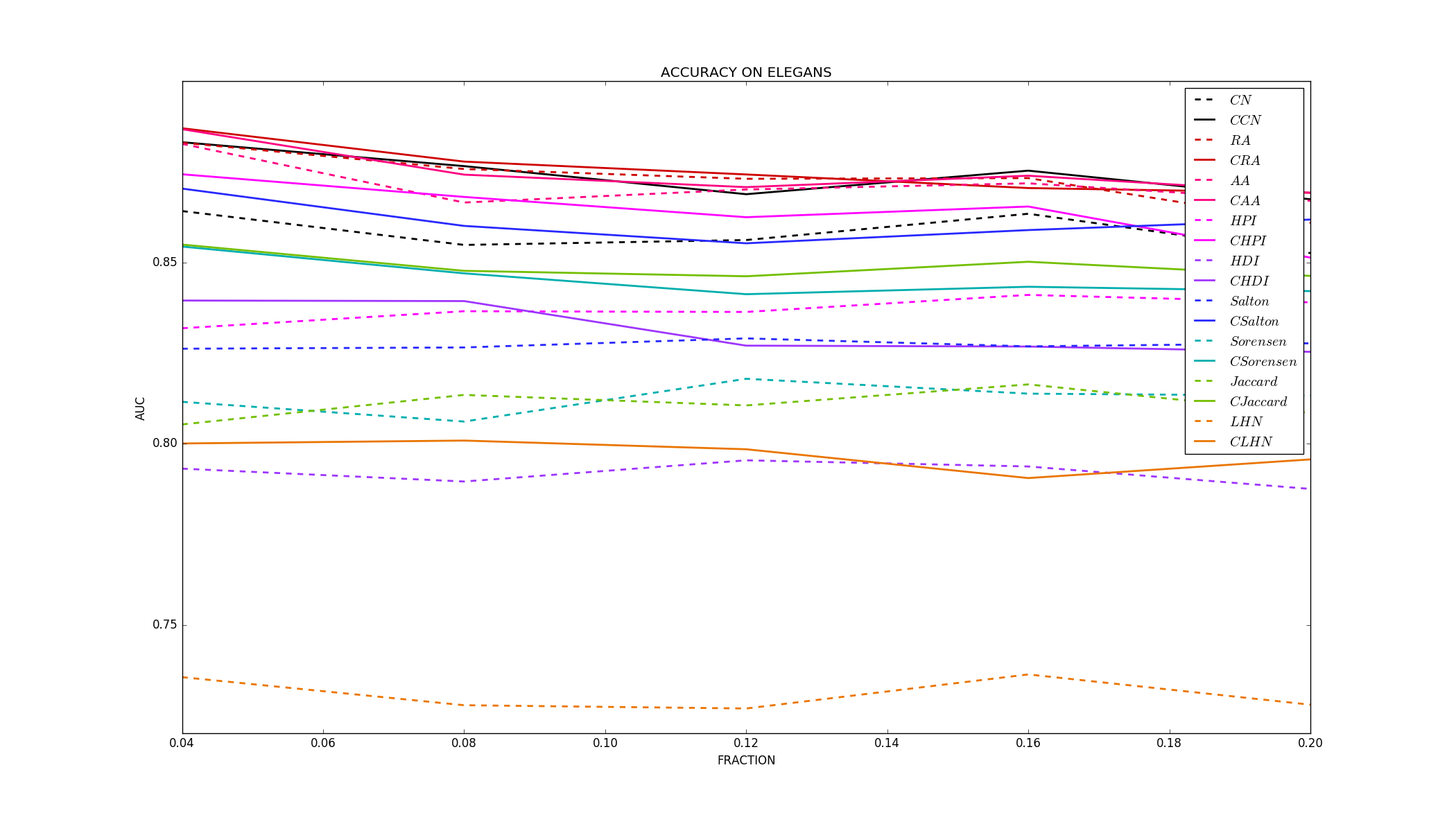
为了保证试验的准确定，当我们确定某一个 p 值之后，我们会对测试集进行 100 次划分，划分之间彼此不关联，这主要是防止测试集的随机性对试验结果造 成的误差。测试集划分好之后，我们将内层循环次数确定为 1000 次，也就是说 总循环比较的次数为 100000 次，该次数对应于上述 AUC 指标中的 n。

### 3.4.2 实验结果以及分析

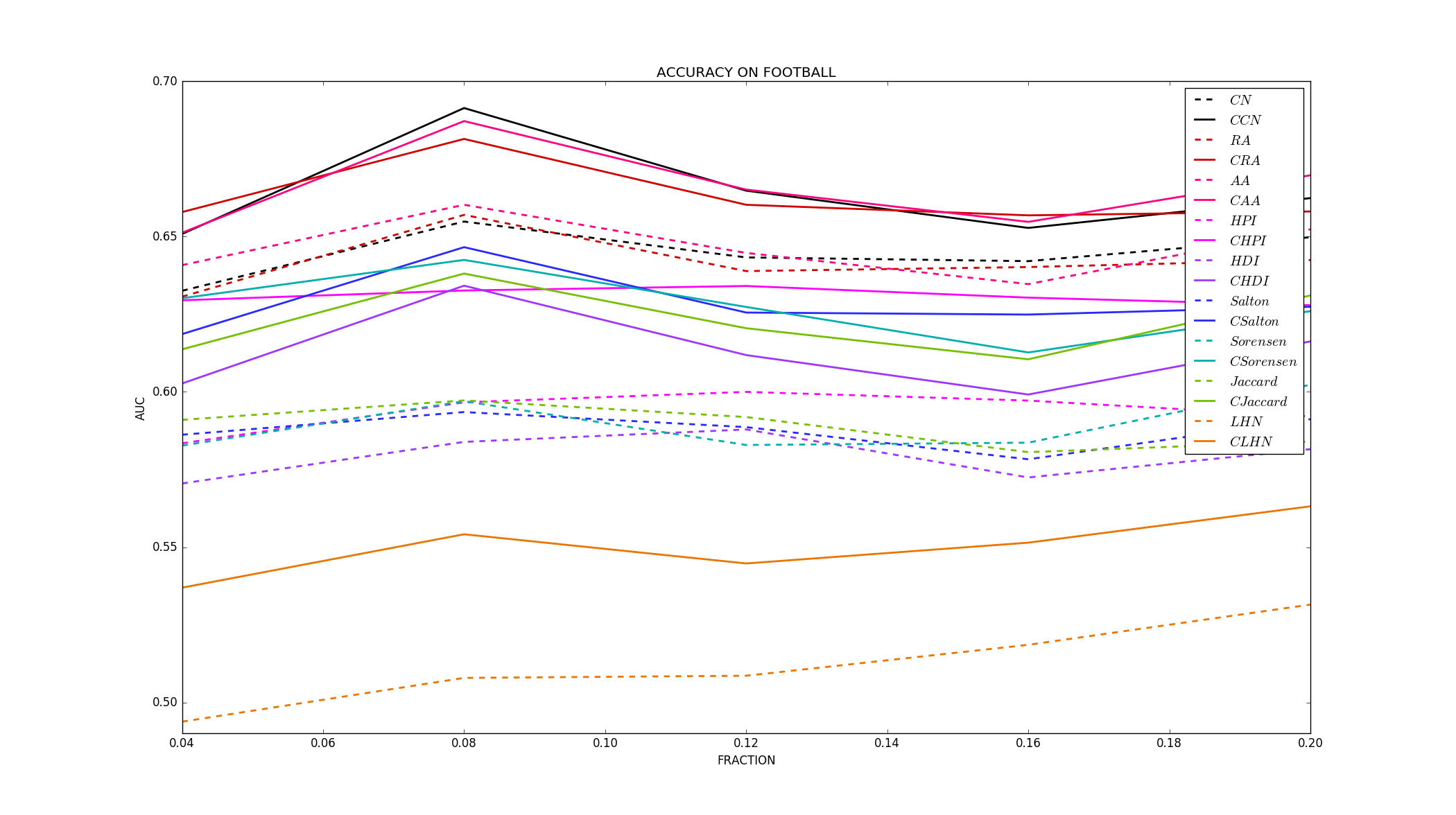
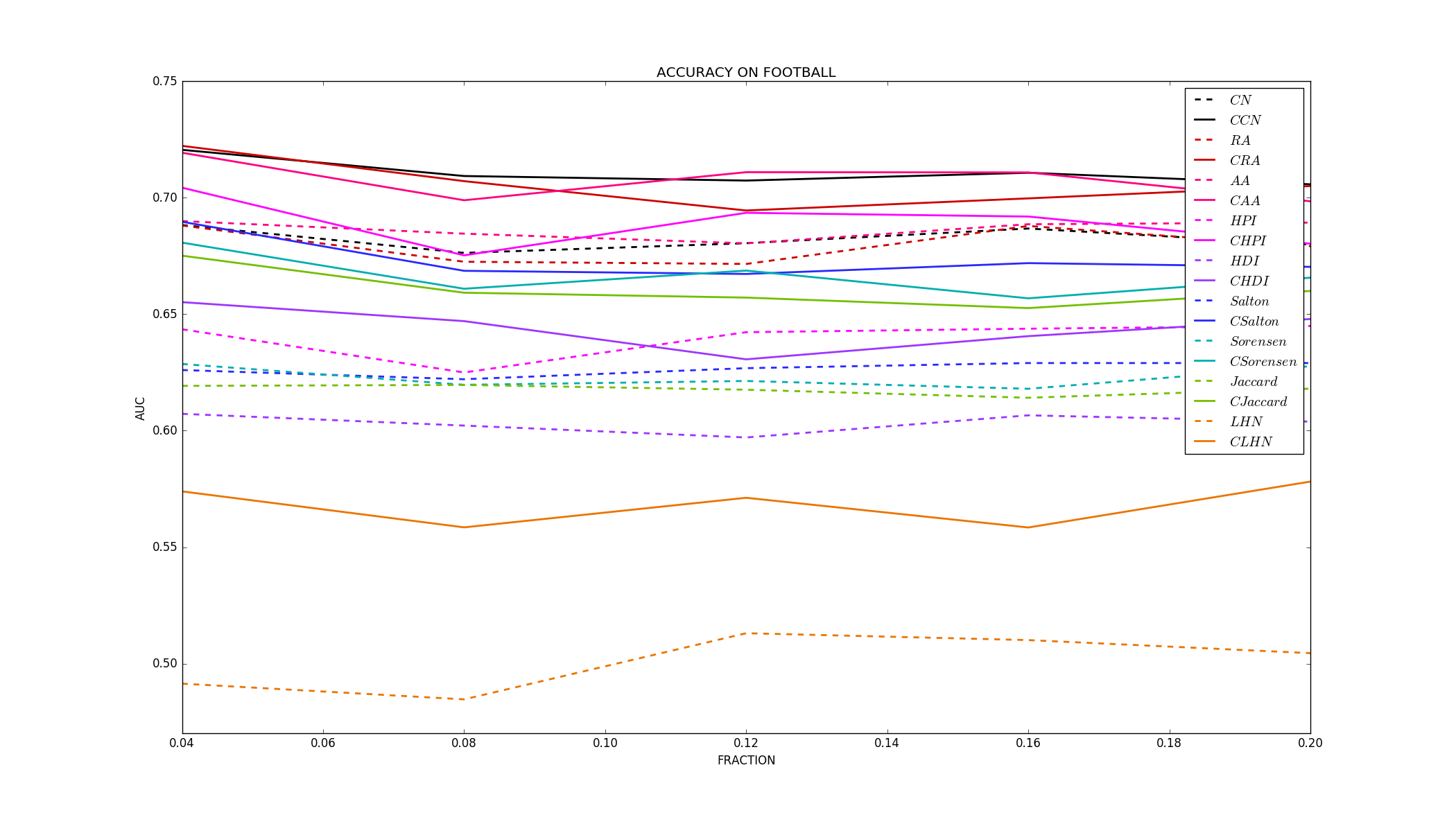
(1)Elegans 网络中缺失边的发现与错误边的纠正，左侧为缺失边的发现，右 侧为错误边的纠正。下图中共 18 种算法进行比较，其中原算法 9 种，加入聚类 系数改进后的算法 9 种。为了便于比较，我们将某一算法和改进后的算法一一对 应，这主要体现在绘图时线段的颜色，如下图所示 CN 和加入聚类系数后的 CCN 算法均由黑色线段表示，RA 和 CRA 均由红色表示。此外，原算法用虚线表示， 改进后的算法用实线表示。结果图的横坐标表示所选取的训练集的比例 p，纵坐 标表示算法的预测准确度。

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。

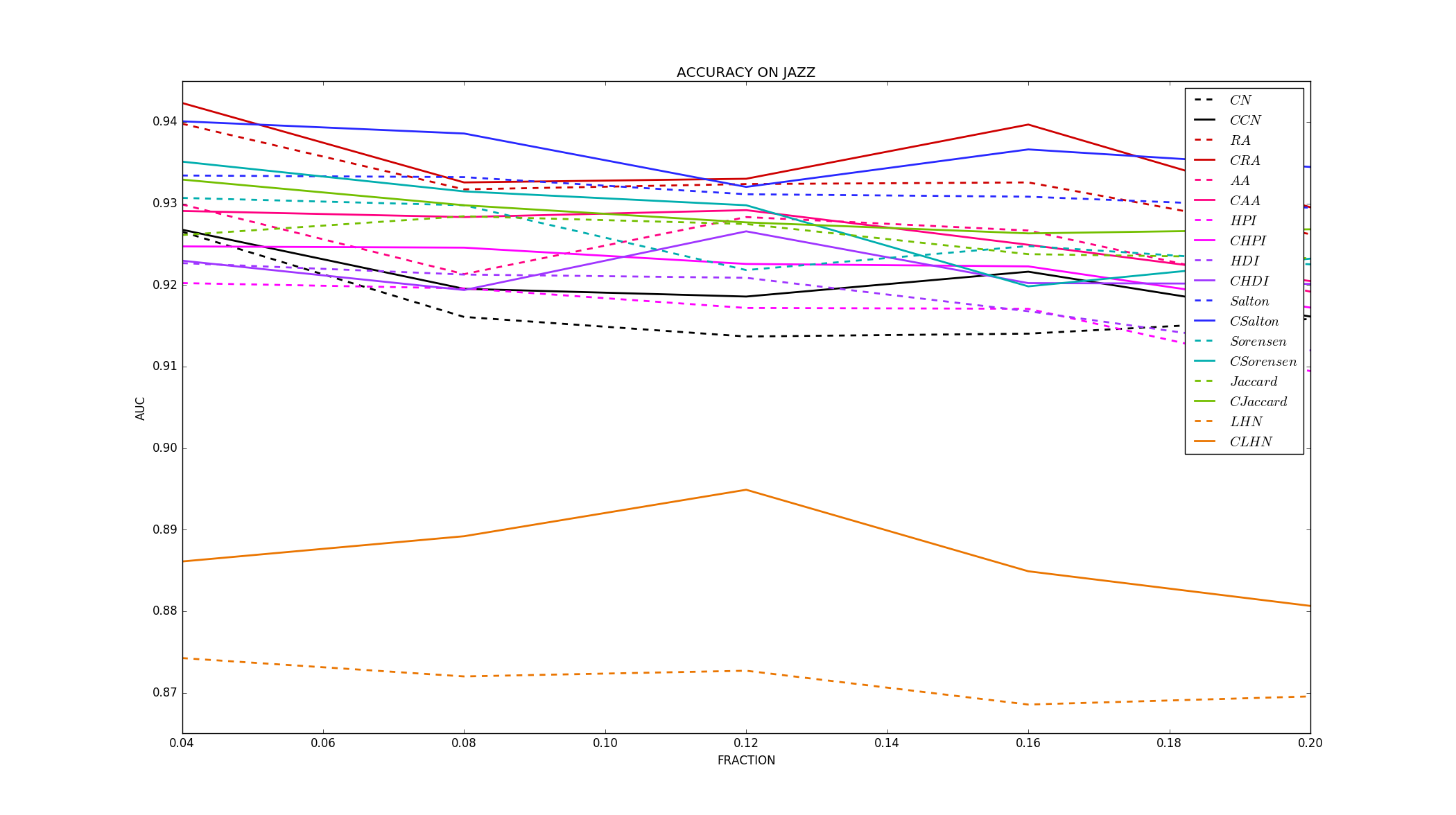


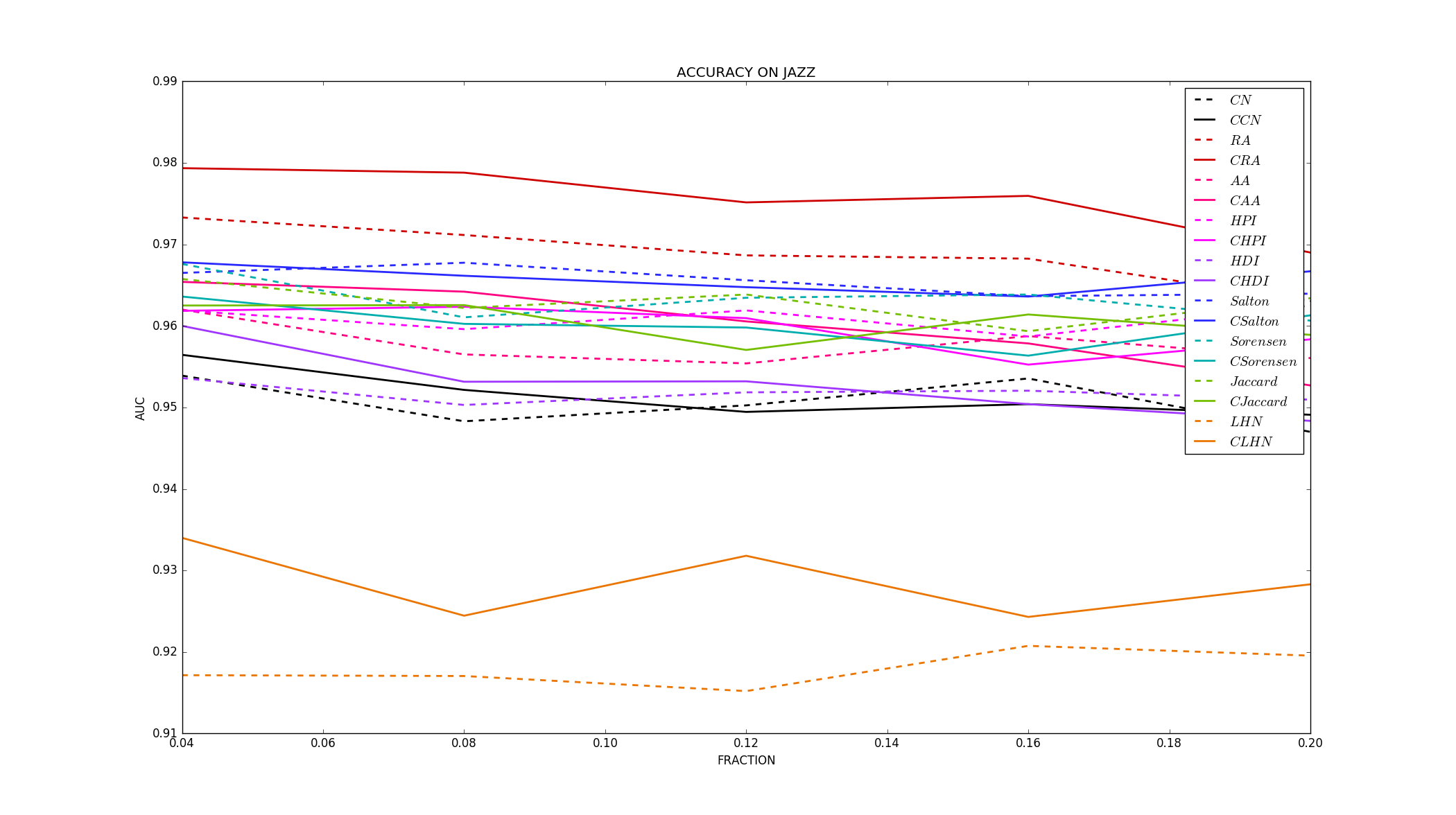


(2)Football 网络的实验结果如下图所示。实验结果表明改进后的算法在整体上明显 优于原先的算法，此外在链路预测的过程中我们发现当 p=0.08 时，预测的准确率达到 最高值。

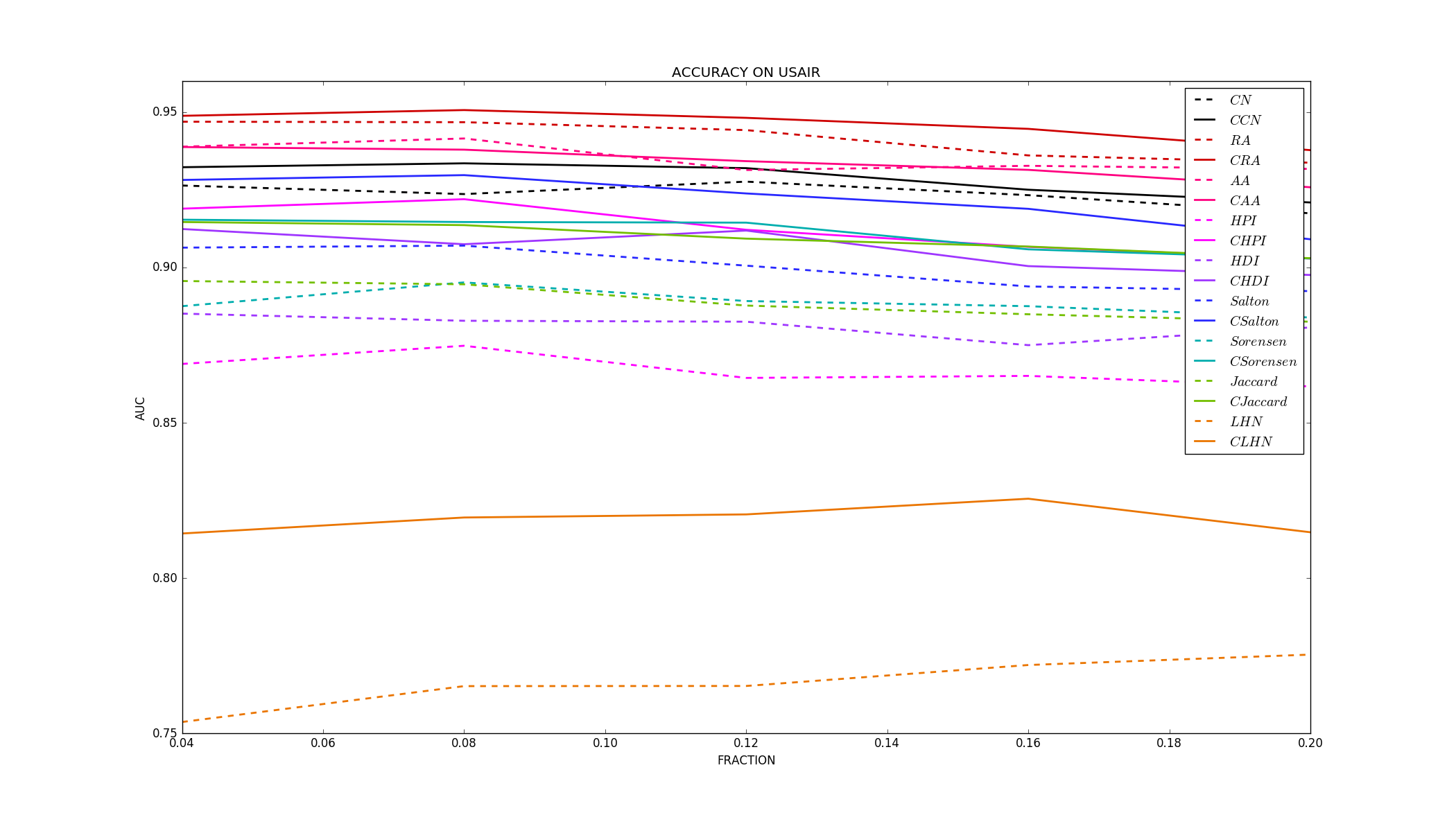
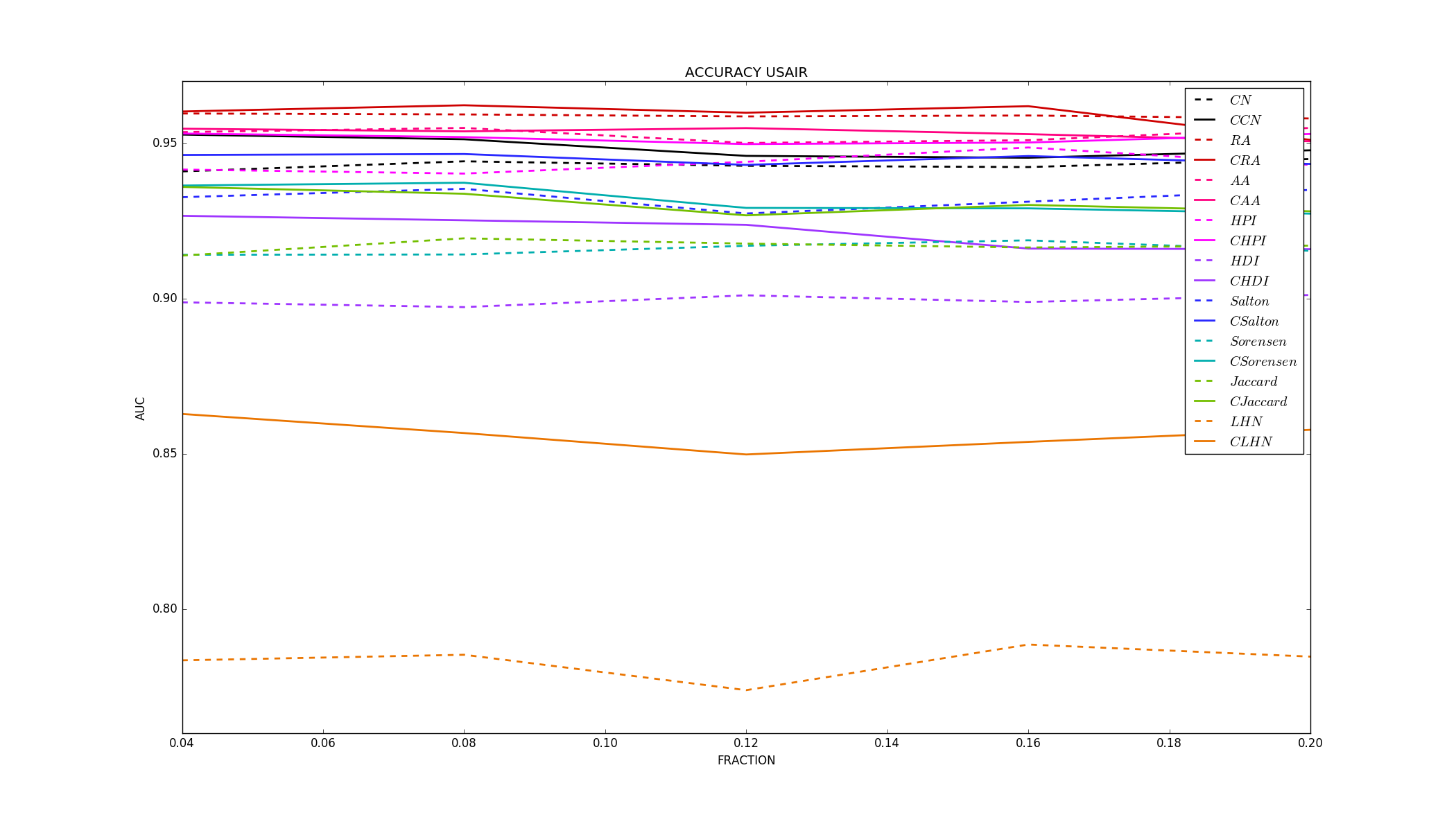


(3)Jazz 网络的实验结果如下图所示。改进后的算法只在少数几个点之上不能超越原 算法，整体表现仍然十分优异。

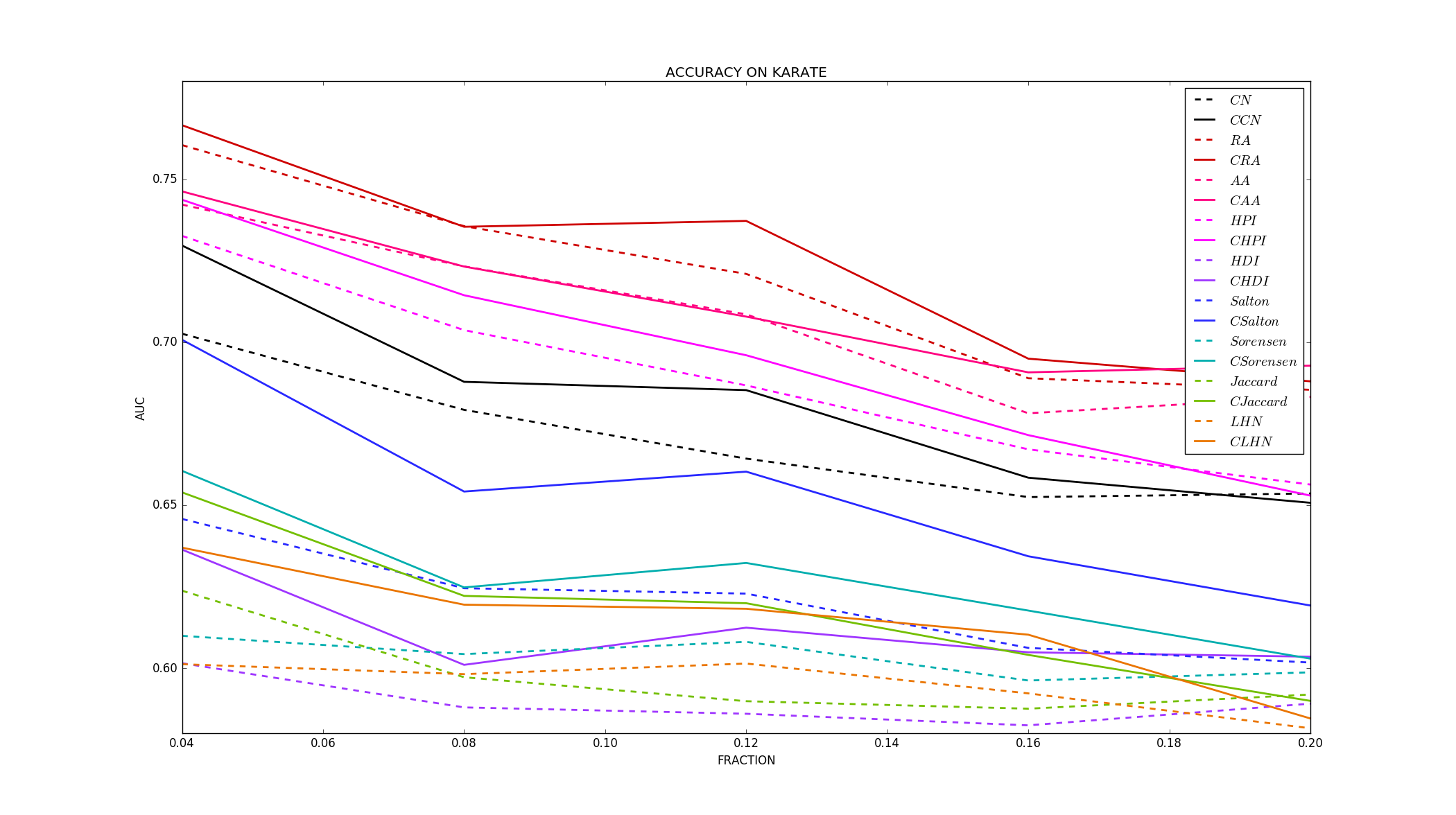


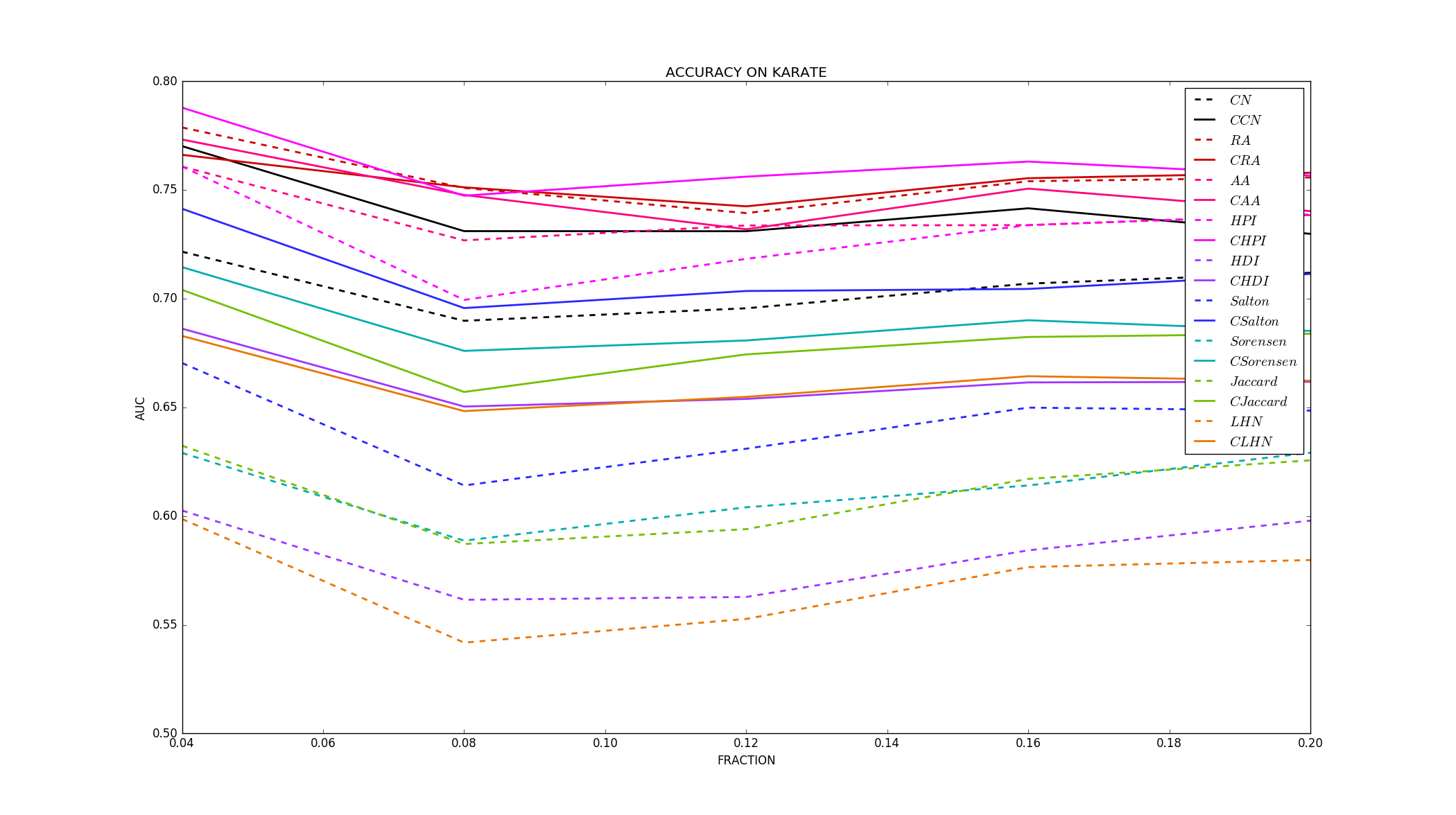


(4)USAir网络的实验结果如下图所示。该实验结果分层现象十分明显，改进后的算法 曲线明显在改进前算法的曲线之上，这表明加入聚类系数这一节点的局部信息之后，该 9 种算法在美国航空网络的预测上效果明显 升。



(5)Karate 网络的实验结果如下图所示。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。





## 3.5 本章小结

聚类系数表示节点邻居之间联系的紧密程度，较高的聚类系数表明邻居间较 高的连接概率。受该指标的影响，我们将其特性添加到现有的 9 种链路预测算法 之中得到了我们改进后的 9 种算法。通过在 5 种真实网络上的实验，我们发现了 改进后的算法在这 5 种网络上有着更好的表现。除此之外，针对不同大小的训练 集划分比例，我们进行了大量的比较实验，过往研究很少会探讨训练集 p 的大小 对于实验结果的影响。我们发现不同大小的训练集对于实验结果会产生影响，但 改进后的算法与原算法的相对大小不变。最后，我们的链路预测实验还包含了算 法在错误连边上的发掘能力，实验表明，若我们有算法 A、B，算法 A 相比 B 在 丢失边的发掘能力上可能会更好，但若换做是错误边的发现，那么 B 算法可能会 优于A算法。

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。

# 第6章 总结与展望

## 6.1 总结

## 6.2 进一步工作

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

[3] 中国信息产业网人民邮电报.三大移动支付安全问题凸显[R/OL]. <http://www.mpaypass.com.cn/news/201404/25115627.html>

[4] 360安全中心.2014年第一期中国移动安全支付报告[R/OL]. <http://policy.catr.cn/zcyj/201405/t20140504_1016393.htm>

**个人简历、在读期间发表的学术论文与研究成果**

**个人简历：**

卢赟，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：**