August 2014

Complex Network Analysis for Scientific Collaboration Prediction and Biological Hypothesis Generation

Qing Zhang
University of Wisconsin-Milwaukee

COMPLEX NETWORK ANALYSIS FOR SCIENTIFIC COLLABORATION PREDICTION AND BIOLOGICAL HYPOTHESIS GENERATION

by

Qing Zhang

A Dissertation Submitted in
Partial Fulfillment of the
Requirement for the Degree of

Doctor of Philosophy
in Engineering

at

The University of Wisconsin-Milwaukee

August 2014

ABSTRACT
COMPLEX NETWORK ANALYSIS FOR SCIENTIFIC COLLABORATION PREDICTION AND BIOLOGICAL HYPOTHESIS GENERATION

by

Qing Zhang

The University of Wisconsin-Milwaukee, 2014
Under the Supervision of Professors Susan McRoy, Hong Yu

With the rapid development of digitalized literature, more and more knowledge has been discovered by computational approaches. This thesis addresses the problem of link prediction in co-authorship networks and protein--protein interaction networks derived from the literature. These networks (and most other types of networks) are growing over time and we assume that a machine can learn from past link creations by examining the network status at the time of their creation. Our goal is to create a computationally efficient approach to recommend new links for a node in a network (e.g., new collaborations in co-authorship networks and new interactions in protein—protein interaction networks).

We consider edges in a network that satisfies certain criteria as training instances for the machine learning algorithms. We analyze the neighborhood structure of each node and derive the topological features. Furthermore, each node has rich semantic information when linked to the literature and can be used to derive semantic features. Using both types of features, we train machine learning models to predict the probability of connection for the new node pairs.

We apply our idea of link prediction to two distinct networks: a co-authorship network and a protein--protein interaction network. We demonstrate that the novel features we derive from both the network topology and literature content help improve link prediction accuracy. We also analyze the factors involved in establishing a new link and recurrent connections.

Acknowledgment

It's been an exciting and adventurous journey. I have been working with Professor Hong Yu, who is my major adviser from 2008 to 2013. I will forever be grateful for her guidance. She has always supported my research endeavors and believed in me. Her research advice and, more importantly, her perfectionism and persistence in her work, shaped my view of how to conduct research.

I am also fortunate enough to work with Professor Susan McRoy, who became my major adviser since 2013 after Professor Hong Yu moved to University of Massachusetts. I have got extremely valuable comments for my dissertation from her.

I would like to thank Weisong Liu, Shashank Agrawal, Ricky Sethi, Feifan Liu, and Balaji Polepalli Ramesh. It's always a pleasure to discuss with them and they always have great advice when I'm stuck. Professor Christine Cheng also provided valuable advice in my research questions.

This thesis is dedicated to my parents.

LIST OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

1.1 Digitalized Research Resources

Large-scale, multi-disciplinary literature repositories have become increasingly available, with publishers, government agencies, and industry service providers making considerable efforts to facilitate literature search and content mining. Notable examples include Thomas Reuters Web of Science [1] and Google Scholar. Google Scholar is Google's literature search engine, with data from individual authors, university repositories, and journal publishers. It "includes journal and conference papers, theses and dissertations, academic books, pre-prints, abstracts, technical reports and other scholarly literature from all broad areas of research" [2]. In addition, publishers have made efforts to improve content accessibility. Elsevier Scopus [3] provides a powerful literature search service.

Literature index services also exist for various disciplines. $CiteSeer^X$ [4] indexes citations in computer and information science and arXiv.org provides a publication posting and downloading service for multiple fields, including physics and mathematics.

PubMed [5], hosted by the National Institutes of Health in the United States, is the leading effort to incorporate biomedical research literature into a publicly accessible resource. It provides not only literature citation downloading (MEDLINE), but also a rich knowledge base for biomedical research. PubMed article IDs are widely referenced by other biomedical knowledge bases, such as Online Mendelian Inheritance in Man (OMIM) [6] and DrugBank [7], and thus made the mining of rich biomedical knowledge possible.

Biological knowledge bases are another important resource. Starting in the 1990s, biological research began actively adopting state-of-the-art computing technologies. Rich resources such as genome and sequence databases, as well as interacting protein databases, were created and have been constantly updated since. For example, the Biomolecular Interaction Network Database (BIND) is a database that stores biomolecular interactions, complexes, and pathway information [8]. The Protein Information Resource (PIR) provides a protein sequence database [9]. The Munich Information Center for Protein Sequences (MIPS) in Germany hosts repositories for genome and sequence data [10, 11, 12, 13, 14] in collaboration with PIR. UniProt [15] integrates sequence and gene function information, which integrates information from both literature curation and automatic classification from Swiss-Prot [16, 17], PIR, and other resources. Gene Ontology [18] provides a structured, controlled vocabulary that describes the roles of genes and gene products.

1.2 Linking Knowledge

1.2.1 Citation Networks

Citation relations between articles are one of the most important semantic relations and have been extensively studied. Citing one's work is usually considered an endorsement; therefore, citation counts are the primary quantified measure of a publication's impact. For example, the widely referenced Journal Impact Factor [19, 20, 21, 22] is based on the citation count a journal's publications receive during a certain period. The H-index [23], which is based on the citation count distribution over an author's publications, is also frequently used, for example, by Google Scholar, as a measure for quantifying an author's impact. These measures, due to the importance of judging research work, have also rise to their fair share of debate [24, 25, 26, 27, 28, 29]. Citations also provide a new source for the text mining community. Citations are an important resource for text mining tasks such as summarizations of target papers, named entity disambiguation, named entity extraction, and relation extraction [30, 31, 32].

Linked publications not only provide citation counts with respect to one another but also draw a comprehensive picture of one or more disciplines if we look beyond a network's immediate neighbors. For example, clustering methods have been used to identify sub-domains within a field, such as sustainability [33] and organic LED technology [34]. Community structure is also studied by applying modularity maximization to well-cited articles [35]. Work has also been done to identify research evolution, such as in the discovery of DNA theory [36], and research streams and diversity [37]. In the legal domain, case documents often cite previous cases to support a particular judgment and thus citation network analysis helps to identify related cases for a given issue [38]. Pattern citation networks have also been used to detect innovation [39]. Degree distributions in citation networks have also been studied and scale-free phenomenon have been frequently observed [40][41].

1.2.2 Co-authorship Networks

Researchers are linked together by co-publishing papers, thus creating co-authorship networks. An author's authority is widely analyzed based on the co-authorship network. De Castro et al. [42] defined the Erdos number, the distance between two authors as the number of nodes in the shortest path between them, and studied the co-authorship network that included the famous mathematician Paul Erdos. The smaller the number, the greater the impact of the author's reputation. Instead of measuring the distance to Paul Erdos, Nascimento et al. [43] extended this work to calculate the "centrality" score for each author, which is the average of the distance between a given author and any other author in the co-authorship network. Liu et al. [44] employed a PageRank algorithm to rank authors.

Co-authorship networks, similar to citation networks, are an important source to study any given research discipline. In particular, it represents the collaboration. Research collaboration has been analyzed through co-authorship networks, such as individual-level collaboration, organization-level collaboration [45, 46], international collaboration [47, 48, 49], and academia–industry collaboration [50]. Trends in research collaboration (inter-/intra-disciplinary, inter-/intra-institutional, and international/national) are important evidence for understanding the outcome of policy and inspiring new initiatives.

Co-authorship networks have been an important resource for social network study, since they are cleanly formatted and the establishment of connections (the co-authors of a paper) is primarily under the author's control instead of random. Physicists use such networks to understand nature's complex systems. Co-authorship networks have been found that are scale-free, small-world, clustering, and assortative networks [51, 52, 53, 54]. Studies on social group evaluation also use co-authorship networks [55].

1.2.3 Protein–Protein Interaction Networks

Protein–protein interaction networks are an important resource in the discovery of new interactions and understanding disease mechanisms. Since the late 1990s, high-throughput computational approaches have also been heavily used to create protein interaction networks. A yeast protein network was built with over 2,000 interactions between more than 1,000 proteins curated from the literature [56]. STRING is a database of functional associations between proteins [57]. BioGRID is a database that incorporates physical and genetic protein interactions that were manually curated from primary literature [58]: As of 2013, it has archived over 500,000 manually annotated interactions from more than 30 model organisms [59].

Protein networks help to understand cellular mechanisms. For example, the essentiality of a (high-degree) protein can be examined by removing it from the network and observing changes in cell function [60].

1.3 Network Analysis

From the late 1990s to the early 2000s, researchers paid increasing attention to patterns of network growth and link establishment. Scale-free networks [61]and small-world networks [62, 63] are arguably the two most important findings. In a scale-free network, also known as preferential attachment, network growth is modeled as a "the rich get richer" pattern (power-law). In other words, the chance of a node receiving a new link is proportional to the number of its existing links. A small-world network, on the other hand, models navigation behavior in the network and shows that one can navigate from one node to the other with a few (counter-intuitively small number of) steps. Furthermore, there are clustering effects (the neighbors of a node tend to connect among themselves) in real-world networks, such as social networks.

Link analysis is one of the most important research areas in network analysis. PageRank [64] and HITS [65] are two pioneer works for ranking nodes in a network. PageRank models a user surfing the Web. A page's importance is simply the probability of a Web surfer landing on it. On the other hand, HITS calculates an authority score and a hub score for each node in the network. A high-authority node means it is pointing to many high-authority hubs and vice versa. The two algorithms are extensively studied and extended.

Link prediction has been attracting more and more attention. The success of friendship-based social network applications, such as Facebook, has motivated the extensive academic study of co-authorship networks, which have long been a standard data source for social network analysis due to their availability and clear semantic context [66].

1.4 Tasks

1.4.1 Citation and Co-authorship Network System

Citation and co-authorship networks are very important for information retrieval and author collaboration studies. However, few resources with full text, citation, and co-authorship networks are available in the biomedical field. First, we use Elsevier's full text collection to build a citation network and identify biomedical related articles. Second, we determine network and co-authorship network characteristics, such as power-law distributions, the size of connected components, and temporal characteristics.

We create a citation and co-authorship network database from Elsevier's full text collection and MEDLINE records. We link the Elsevier articles by citation information with the assistance of information retrieval, which helps to eliminate unlikely candidates. Similarly, we map the Elsevier articles to the MEDLINE records, thus identifying medically related literature. The author names are disambiguated by the authority database.

Furthermore, we use the MEDLINE resource to automatically annotate PubMed citations and train conditional random field models to parse the full citations into fields, such as title and author.

1.4.2 Scientific Research Collaboration Prediction

Biomedical research demands a high degree of collaboration. Collaboration recommendations are therefore important to facilitate the increasing need to find potential collaborators. Collaboration recommendation is often cast as a link prediction task and existing works are primarily based on topological predictors such as the number of mutual collaborators. However, our hypothesis is that other semantic features can be informative for collaboration prediction.

We first study first-time collaboration to predict the collaboration between two researchers who have never collaborated together. We use co-authorship to represent collaboration and use existing co-authorships as training instances to create supervised machine learning models, namely, naive Bayes, logistic regression, support vector machine, and random forest models. Given a new pair of authors, the models predict the probability of the two collaborating together in the future. We develop novel features of author research profiles, as well as network structures.

We subsequently explore the recurrent collaboration problem, since research collaborations, unlike making connections in social media, are discrete events and two researchers may collaborate only once or multiple times. We hypothesize that recurrent collaborations have patterns and therefore machine learning models can be applied to predict them.

1.4.3 Protein–Protein Interaction Prediction for Hypothesis Generation

Protein–protein interactions are important in understanding disease mechanisms. Traditionally, biologists come up with new interaction hypotheses by searching the literature or studying the results of previous experiments, both of which are time-consuming when validating a hypothesis. We propose using a link prediction technique to automatically predict the most likely interactions that have not yet been discovered (i.e., not reported by the literature). Specifically, we formulate the task as a link prediction problem in a protein–protein interaction network. Existing interaction pairs are considered positive examples and we assume that proteins that have never interacted are negative instances. The learning features are derived from both the literature and network structures. We used supervised machine learning, naive Bayes, naive Bayes multinomial, logistic regression, and support vector machine models.

1.5 Contribution

This study's contribution is threefold. First, we create a resource for a biomedical citation network together with full text articles. This provides a resource for both text mining and biomedical research communities for tasks such as information retrieval, literature searches, and research evolution. Second, our in-depth study of scientific collaboration identifies features used in predicting research collaboration. Third, the proposed protein–protein interaction prediction approach is highly scalable and can be used to generate hypotheses for biologists and to build large-scale protein–protein interaction networks.

1.6 Thesis Outline

- Chapter 2 describes how full text biomedical citation and co-author networks are constructed.

- Chapter 3 describes the work of citation parsing using a conditional random field model.

- Chapter 4 presents the work of predicting first-time research collaboration between two researchers.

- Chapter 5 describes the prediction of recurrent research collaborations.
- Chapter 6 describes how to use a machine learning approach to generate a new hypothesis on protein–protein interactions.
- Chapter 7 summarizes the thesis.

Chapter 2

CiteGraph

2.1 Introduction

With the increasing volume of full-text biomedical articles are available online, citation recognition and analysis has become more and more crucial for many text-mining applications. Citations play an important role for both the rhetorical structure [67] and the semantic content of articles [30] and have proven beneficial to many text mining tasks, including information retrieval, extraction, summarization, and question answering. Citation analysis can provide us with insights into the underlying landscape of scientific publication, and useful resources for retrieving, categorizing and evaluating articles, authors and institutions [19]. By knowing which articles reference and are referenced by a particular paper, we can better understand how that paper fits into the larger network of research. This can help us to understand the connections between research topics, and allow us to find related work and judge the authority of an individual paper. Much in the same way hyperlinks transformed the World Wide Web from a set of static documents into a vibrant and interesting network, properly utilizing citation information can produce a scientific knowledge resource with utility far beyond the sum of its parts.

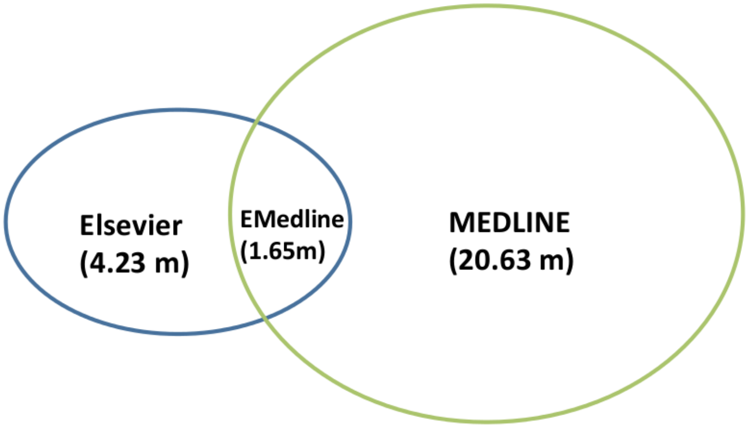


Figure 2.1: The overview of the Elsevier, EMedline and MEDLINE datasets.

In this chapter, we developed and evaluated CiteGraph, a fully implemented pipeline system that builds a large-scale citation network from a large collection of biomedical full-text articles. Currently, the CiteGraph network encompasses 4.23 million articles taken from the Elsevier collection spanning disciplines ranging from physics to economics. In particular 1.65 million MEDLINE indexed articles are identified from them. The CiteGraph network aligns co-authorship, bibliographical, institutional and citation information into a single cohesive network resource. It also assigns PMIDs, the unique identifier in the MEDLINE collection, to the corresponding articles. The data of the Elsevier and the MEDLINE collections are shown in Figure 2.1. Since the articles in the Elsevier collection contain full citation information, the overlap between Elsevier and MEDLINE, named as EMedline here represents a subset of 1.65 million biomedical articles from which we built the EMedline CiteGraph network. We evaluated the CiteGraph on linking articles by citation and identifying MEDLINE articles. The system achieved F-1 scores of 0.99 and 0.98 respectively. We also present further analysis of properties of the network and its citation-related characteristics. This chapter is based on the publication [68].

The overview of the network is shown in Figure 2.2. For the purpose of this work, we define an incite of A as an article that cites A and an outcite of A as an article A

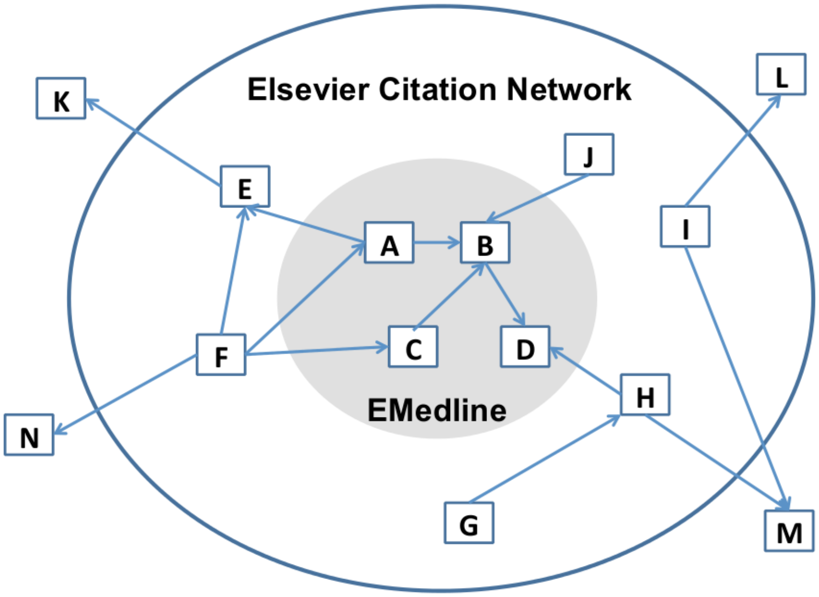


Figure 2.2: An illustration of the CiteGraph network. Each square represents a unique article. A directed link represents a citation relation. The CiteGraph network encapsulates citation relations extracted from the Elsevier data set (outer circle). The set of articles contained in the MEDLINE dataset and the links between them is EMedline network (the inner circle).

cites. For example, in Figure 2.2, the incites of B are A, C and J, the outcite of B is D. CiteGraph maps citations and PMIDs to documents within the network, and merges citations to out-of-network documents in order to form a coherent picture of the links between articles. In this chapter, we calculate precision and recall scores for the citation mapping, PMID mapping and citation merging tasks separately, achieving F1 scores of 0.99, 0.98 and 0.86 respectively.

CiteGraph is fully automated, which makes the future development of larger-scale networks possible. While some manually created networks do contain citation information, this work is the first to report on the development and evaluation of a fully automated system for building large-scale citation networks. These networks will also be made available to researchers, allowing this work to have a substantial impact on both the text-mining and biomedical fields.

2.2 Background

2.2.1 Citation Networks

Citation networks have been built in many domains. In physics, Bilke and Carsten [40] investigated a citation network built using data from publications of high-energy physics ranging from 1975 to 1989, observing, among other features, that the number of citations followed a power-law distribution. Redner[41] studied the citations from 110 years of Physics Review publications, creating a network consisting of 353,268 articles. Chen and Rener[35] used the same network to study the community structure of physics subfields by applying modularity maximization on well-cited articles to discover sub-groups. Small[69] used data from the Information Science Institute to create a map of scientific articles covering 23 disciplines by using co-citation clustering. Discipline sub-topics and interdisciplinary pathways were also identified. Kajikawa and Takeda[34] analyzed citation network of research papers in the field of organic light-emitting diodes (OLEDs). Topological clustering methods are used to cluster articles based on citation relations and identified research disciplines. Emerging research topics can also be identified through the use of recursive clustering. In a study of citations related the scientific development of absorptive capacity field, Calero-Medina and Noyons[37] identified main paths in a citation network to study research stream and diversity. Greenberg[70] observed citation bias, amplification and invention in a network of PubMed articles that shared a particular belief. The work showed that citation distortion might occur with literature that has been accumulated over long periods. Link analysis methods such as HITS[65] and PageRank[71] have been applied to citation networks to identify and distinguish high quality articles [72][73][74].

Citation networks have been used for broad types of documents. In the legal domain, case documents often cite previous cases to support a particular judgment.

These citations are categorized by legal issues and defined by guidelines. Zhang and Koppaka[38] described a semantic-based legal citation network tool capable of creating legal citations networks for a given issue. Csardi et al[39] used a patent citation network to detect innovations.

2.2.2 Literature Repositories

Large-scaled literature repositories have been built. $CiteSeer^X$[4] is a digital library of scientific literature, and search engine, primarily focusing on computer and information science domain. It autonomously creates a citation index, and provides citation statistics including the number of citations for a given paper, and a list of top cited articles and authors. $CiteSeer^X$ also allows searching of articles that cite a given paper. Google Scholar[75] is another popular literature search engine that collects papers from multiple disciplines. The number of incites is used as part of its ranking algorithm.

2.3 Methods

In this study, we describe the system CiteGraph that builds the citation and co-authorship network. It consists of a matching algorithm that maps each citation to its article and the corresponding PMID. The implementation of the matching algorithm resulted in four distinct components: *citation mapping* matches a reference to its corresponding article and subsequently establishes the citation links between two articles in the CiteGraph network and *PMID mapping* assigns a unique PMID to each article node. *Citation merging* component merges citations that are pointing to the same article that is outside of the collection. In addition to the two components, *author name disambiguation* component disambiguates authors.

2.3.1 Data and Preprocessing

The data that CiteGraph used comprises of 4.23 million full-text articles from the Elsevier data and 20.63 million MEDLINE records. Each MEDLINE record comes with an XML file with fields including Title, Journal, Author, and Year, as well as the MeSH terms assigned to the record. Every Elsevier article also comes with an XML file. Unlike the MEDLINE, the Elsevier article does not include MeSH terms, but has citations -with fields similar to the MEDLINE record - that the article cites. Accordingly, CiteGraph parsed both the MEDLINE and the Elsevier XML files by field, assigning the parsed citations to their files and indexed them with the open source information retrieval tool Apache Lucene [76].

2.3.2 Matching Algorithm

The CiteGraph network contains three entity types: article, citation and PMID. Each entity type has common fields (e.g., title, journal, year). All three of the following tasks are based on a matching algorithm that uses these fields to resolve co-reference entities. The matching algorithm operates on three main fields: title, author list and journal, applying different equivalence tests for each field. Two entities are matched if two or more of their fields are equivalent.

Title: We found title varies in its expression. For example, named entities including chemical and gene names were often represented differently in a citation when comparing it to the title in the original published article. We therefore developed an approximation approach for matching two titles. Specifically, two title fields are considered equal if one of following conditions is met: 1) the set of tokens contained in one title field is a subset of the tokens in the other, or 2) the number of tokens common to both fields is more than 80% (we veried the range from 30% to 100% and found 80% show best performance) of the size of the larger of the two fields. This ad-hoc approximation approach works quite well, as demonstrated in our preliminary evaluation.

Author List: To identify whether the two author lists are identical, we first compared the surnames of the authors; we discarded first names as we found them to introduce significant noise. The author list fields are considered equal only if the set of surnames are equivalent, or if the set of surnames in one list is fully contained in the surname set of the second.

Journal: Due to the fact that many journal citations are given using the journal initials, or abbreviated names (e.g., "Mech. Dev", "J. Neurosci." and "JAMA"), journal name initials, rather than full titles, were compared. Stop words, such as *of* and *the* were removed. If the number of common initials in the journal titles was greater than 80% of the tokens in the longer journal name, they were considered equivalent. The figure of 80% was determined through empirical evaluation ranging from 30% to 100%.

2.3.3 Citation Mapping

The citation-article mapping task attempts to link citations to their corresponding article nodes. For each citation, the title, author list and journal name are extracted and used to create a query. A list of 20 candidate documents is retrieved, and each returned document is checked using the matching algorithm. 20 candidates are required as citation information is usually less complete than the meta-data stored in the Elsevier index (30% of the citations in the network do not provide an article title, and 12.2% of the citations do not have journal information). An alternative approach would be to use article information to query the citation index and attempt to find all citations that refer to a particular article. However, in practice it was found that due to the high percentage of missing fields, this required a much larger candidate list in order to maintain recall levels, and was therefore less efficient.

2.3.4 PMID Mapping

We identify PubMed articles in our network so they can be analyzed separately as the EMedline network. For each article node, the appropriate data fields are retrieved from the Lucene index and used to construct a query to retrieve candidate articles from the MEDLINE index. The matching algorithm is then used to check the ten most relevant candidates. If a match is found, its PMID is mapped to the article node.

2.3.5 Citation Merging

Unlike citation mapping, the citation merging task does not have a complete article upon which to map each citation. Instead, citations that reference articles outside of the network are merged into a single citation object. Due to the less complete and consistent nature of the citation information, this is the most difficult task among the four. However, it is important, as it allows us to determine when, for example, two articles share a citation, even if the cited article does not exist in our network.

In this task, both entities submitted to the mapping algorithm are likely to have fields with missing data, so it is important to retrieve a large number of candidates. For each citation C, a query is constructed with all available field information, and the top 3000 candidates are retrieved. We choose 3000 based on our preliminary study, which showed a maximum of 3000 citations to a single article. Each candidate is compared pairwise with C using the matching algorithm. Matched nodes are merged into a new citation C', which may have more complete information than C, and can be used for subsequent steps.

Table 2.1: EMedline Network Statistics

Measures	EMedline Incites	EMedline Outcites	Total incite	Total outcite
Average	3.85	3.85	26.66	32.34
Max	4977	829	37065	3730
Min	0	0	0	0
Std	10.94	5.87	78.64	32.03

2.3.6 Author Name Disambiguation

Author names are frequently ambiguous; the same name may refer to different authors. We therefore developed the *author name disambiguation* component. For implementation, we used the Author-ity database[77], which has disambiguated all authors in the MEDLINE database. With the Author-ity database, we identified a total of 1.37 million unique authors in our EMedline network.

2.4 The Network and Evaluation

The aforementioned four CiteGraph components allow us to create citation networks by linking articles and co-authors. Two networks were created by CiteGraph: the Elsevier network, consisting of all article nodes and citations extracted from the Elsevier dataset; and the EMedline network, a subset of the Elsevier network consisting of only the articles which were assigned PMIDs and the citations between them, as illustrated in Figure 2.1.

Using both our Elsevier and the MEDLINE datasets, the Elsevier network contains 4.22 million articles, and 106.25 million citations, while the EMedline network has 1.65 million articles, and 6.35 million citations. The average EMedline network node cites to 3.85 EMedline and 32.34 Elsevier nodes. EMedline article nodes are also on average cited by 3.85 EMedline nodes and by 26.66 Elsevier nodes. Table 2.1 shows the characteristics of the EMedline network.

The ratio of internal citations (links between EMedline nodes) to total citations (links to EMedline nodes from all Elsevier nodes) in the EMedline network is 0.14, which is similar to work [41] which uses the 110 years of Physics Review (PR) article collections, where the internal-citation ratio is as small as 0.2 for well-cited elementary particle physics publications.

There is no gold standard for citation resolution, and citations are written with the intention of being human readable. Therefore evaluation of the three tasks detailed in Section 2.3 (author name disambiguation was excluded for our evaluation as it was completely based on Authori-ty database, whose evaluation was reported (98% recall) by the referenced article [77]) is carried out by human judges. In each of the following evaluations, seven Human evaluators are provided with entity (article, citation or MEDLINE record) pairs, and are asked to determine whether the two entities refer to the same article. Each evaluator provides judgments on 20 instances of each task. 25% of the instances are double annotated in order to evaluate inter-annotator agreement.

Citation Mapping: For each citation in the network, a list of potential article mappings is created by selecting Elsevier articles that have either their *title* or *authorList* field equal to the corresponding field of the citation, as determined by the matching algorithm described in section 3.2. Each evaluator is presented with a list of 20 citations randomly selected from the set of citations with at last one potential mapped article. For each citation, the list of potential mapped articles is presented, and the evaluator must select which article, if any, corresponds to the citation in question. This establishes a set of 110 user-curated citation mappings. Precision is measured as the number of citations correctly mapped divided by the number of citations presented that receive a mapping. Recall is calculated as the number of citations correctly mapped divided by the total number of correct mappings found by evaluators. Precision and recall for this task were calculated as 1.00 and 0.96 respectively, resulting in an F1 score of 0.98.

Table 2.2: Evaluation result.

Task	Precision	Recall	F1	Inter-annotator agreement(Kappa)
Citation mapping	1	0.96	0.98	1
PMID mapping	0.99	0.99	0.99	1
Citation merging	0.82	0.91	0.86	0.89

PMID Mapping: The evaluation for PMID mapping is performed using the same tool as the citation mapping. Evaluators are presented with information for 20 articles that have at least one candidate PMID mapping, and the list of potential PMID mappings. The evaluators are then asked to choose which PMID, if any, corresponds to the given article. For this task, both precision and recall were found to be 0.99, for a resulting F1 score of 0.99.

Citation Merging: As the citation merging task does not have a set article upon which to map, it is necessary for the evaluators to evaluate all possible merges for a given citation and select all of the candidates which should be merged. In order to reduce the number of pairs evaluators will need to compare, citation nodes that have all of their fields equal are first merged, and the evaluation is only completed on candidate pairs with at least one missing or non-equivalent field. The evaluators are then presented with a citation node, and a list of potential candidate nodes for merging. They select all candidate nodes that refer to the same article as the original citation node. Precision is measured as the number of candidate pairs correctly merged divided by the total number of pairs merged. Recall is measured as the number of candidate pairs correctly merged divided by the total number of merges identified by the evaluators. Precision and recall for this task were calculated as 0.82 and 0.96 respectively, for an F1 score of 0.86.

Autuhor Name Disambiguation: We also disambiguated the author names using Authori-ty database. 1.39 million authors from 1.19 million articles were disambiguated.

2.5 Network Analysis

2.5.1 Citation Network

The number of articles over the number of incites is often observed to follow a power-law distribution [78][79]. Figure 2.3 shows the plot of EMedline network, with $\alpha = 0.85$, and $\beta = -2.40$ for $logy = \alpha + \beta logx$, where x is the number of inciting citations in an article, and y is the frequency of the articles that have inciting citations x.

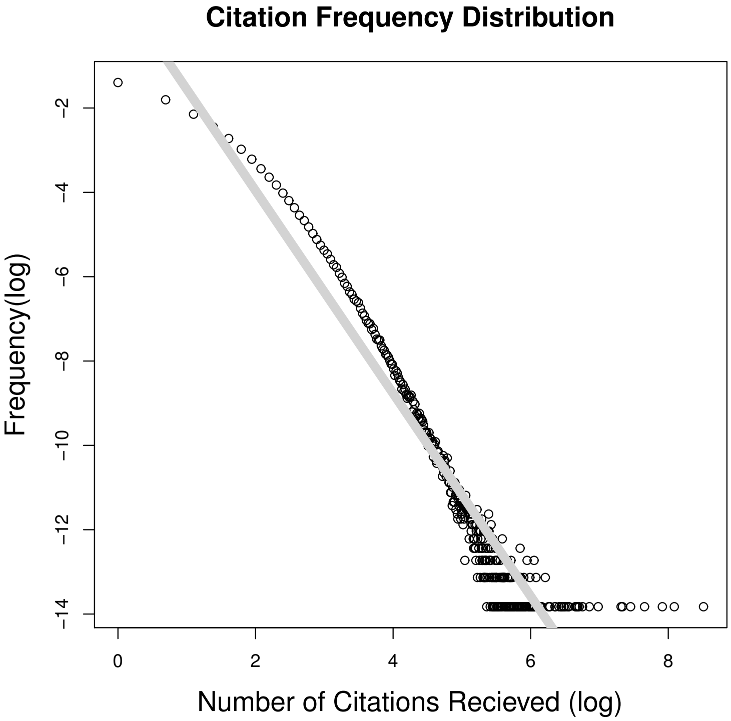


Figure 2.3: EMedline citation frequency distribution. The line is the linear regression of the plot, let it be $logy = \alpha + \beta logx$, the linear regression produces $\alpha = 0.85$, and $\beta = -2.40$. The citation frequency is therefore follows power-law distribution

2.5.2 Co-authorship Network

2.5.2.1 Basic Analysis

Density. As described in [69] , density measures the general connectedness of a network. It is defined as the number of links in the network divided by the number of links in a complete graph with the same number of nodes. Given a graph G with N nodes and $L(G)$ links, the density D is defined as:

$$D = \frac{2 * (|L(G)|)}{N(N-1)}$$

In this co-authorship network, which has 7,206,814 links and 1,376,779 nodes, D=7.6e-06. As expected, the network is extremely sparse.

Clustering coefficient describes how well the neighbors of a vertex are connected among themselves. As described in [69] the clustering coefficient of vertex v in a graph G is defined as

$$\frac{|E'|}{k(k-1)/2}$$

V' is the set of vertices of the direct neighbors of v, $k = |V'|$. E' is the set of edges among vertices in V'. The overall clustering coefficient of G is the average of each node. The nodes with zero indegree and outdegree are assigned 0 as the coefficient.

The average clustering coefficient of the network is 0.6798. The co-authorship network in [44] also has 0.67. Similarly ACL SIGMOD co-authorship [80] has clustering coefficient 0.69. It is highly likely a small world graph according to other study [69], but a random graph is needed to verify it. It is not included in this study.

Connected component is analyzed and 25,510 components are found. Top 10 components with largest size are shown in Table 2.3. The largest component has size 1.27 million, and the second largest one has 36 nodes. The co-authorship network has

1.37 million unique authors, and therefore 92.7% authors in the network are connected in the largest component. The study of [69] showed in their co-authorship network of 1567 authors, the largest component had 599 (38.2%) nodes. Figure 2.4 shows the number of component sizes, excluding the largest component.

Table 2.3: Top 10 largest connected components

Component No.	Size
1	1269989
1755	36
529	30
2473	26
1307	25
1263	24
1146	23
5061	23
2886	23
5132	22

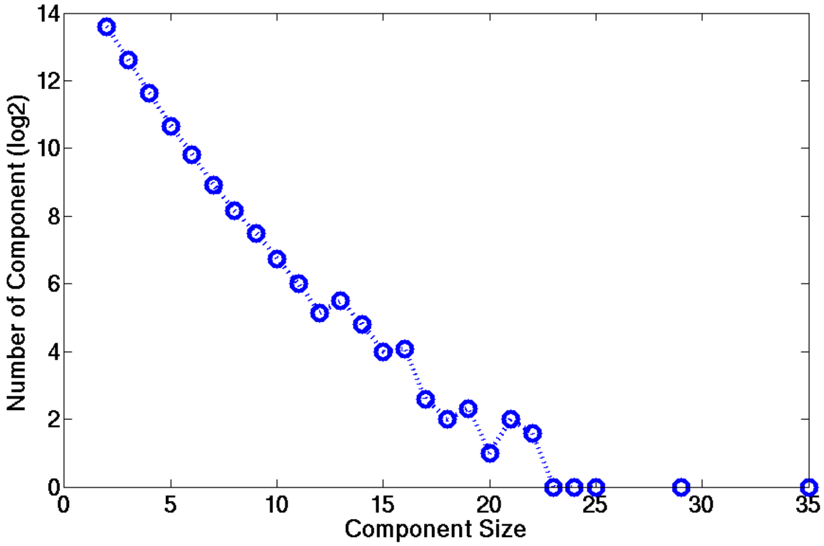


Figure 2.4: Connected component. The largest componnet is excluded from the plotting as its size (1.2 million authors) will skew the figure.

Table 2.4: Statistics of Network Measures

Measure	Mean	Median	Std	Max	Min
Component size (No.1 excluded)	3.1756	3	1.8280	35	2
Clustering coef.	0.6798	0.8095	0.3542	1	0
Number of co-authors	11	6	14	671	0
Co-authorship year span	1.5211	1	1.5762	35	1

2.5.2.2 Temporal Analysis

The time span of co-authorships is an indicator of the strength of the collaborations. As shown in Figure 2.5, co-authorship spans from 1 to 35 years, while 83.7% of author pairs just appear once. There are 766,834 and 113,640 co-authorships with five and ten years span respectively.

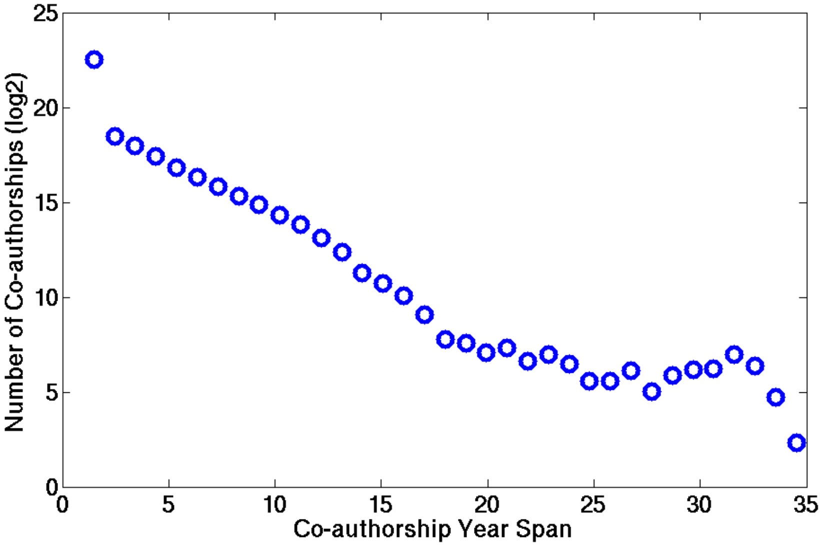


Figure 2.5: Co-authorship Span

We were also interested in analyzing the relations of author, publications over the year offset such as average number of co-authors for an author in the i^{th} year of his/her publication history. Therefore we calculated the percentage of outside institute co-authors, citations received from EMedline/Elsevier, number of publications as well as number of co-authors per publication. We subtracted the earliest year when an author published in this collection from current year of interest to get the year offset. There are only 0.24% authors who have a publication history longer than 16 years in this collection, so the data is not as representative as the rest. Hence we used year offset from 0 to 16 for the analysis.

As shown in Figure 2.6, Average number of co-authors per publication is generally increasing, which suggests an author tends to have more co-authors as he/her gets senior (a). Similarly the percentage of co-authors outside of institution is increasing, suggesting in general authors have broader collaborations when becoming senior (b). Number of publications every year is also increasing, and peaks at 14. It shows an author gets more and more productive along his/her career (c). Number of citations per publication decreases, and the reason could be that earlier works have more time to receive citations than the newer ones (d). The funding level, policy and culture changes along the time are also possible reasons for the trend in (a) and (b).

2.6 Conclusion

In this chapter, we described our efforts on building large citation networks using both the Elsevier and the MEDLINE datasets. We developed a citation matching algorithm and implemented four components that match a citation to its corresponding article, identifies the MEDLINE indexed articles and disambiguates author names. Our system - named CiteGraph - incorporates over 4 million Elsevier articles, and 1.6 million related articles in it have been identified. The evaluation demonstrated a F1 score ranging from 98% - 99% for different components. With the CiteGraph networks, we subsequently conducted preliminary graph analysis, including citation frequency over publications and co-authorship network topological statistics. Our analysis demonstrates that the CiteGraph networks reflect the general characteristics

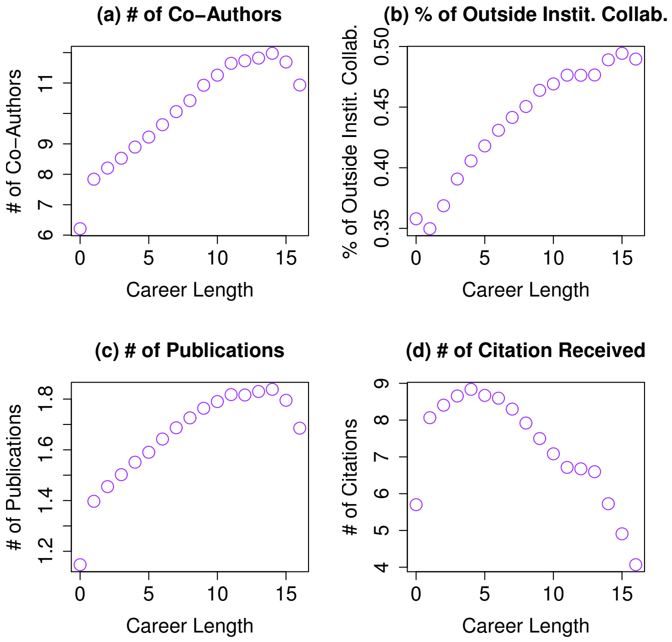


Figure 2.6: Author temporal analysis. (a)Average number of co-authors per publication is generally increasing, which suggests an author tends to have more co-authors as he/her gets senior. (b) Similarly the percentage of co-authors outside of institution is increasing, which shows authors have broader collaboration along his/her career. (c) Number of publications every year is also increasing, and peaks at 14. It shows an author gets more and more productive. (d)Number of citation per publication decreases, and the reason could be that earlier works have more time to receive citations than the later ones.

of existing networks in other domains. In addition, the temporal analysis shows the researcher tends to have more co-authors per publication and more outside institution collaboration along the career. The limitation of this work is due to the incomplete dataset, as CiteGraph is built on a subset of MEDLINE.

The CiteGraph network provides the medical informatics community a new resource on text mining. The system can be used to combine the citation network and co-authorship network together to analyze links between them. Such combination and analyses may lead to important discovery, including document ranking, author ranking, and author collaboration pattern detection.

Chapter 3

Parsing Citations Using Conditional Random Fields

3.1 Introduction

As more and more full-text biomedical articles become open-access, there is a great need to move beyond merely examining abstracts and to develop text mining approaches that apply to full-text articles. Citations are used ubiquitously in biomedical articles; for instance, we found an average of 34 citations for the 160,000 full-text biomedical articles in the TREC Genomics Track text collection [81]. Citations play important roles for both the rhetorical structure and the semantic content of the articles, and as such, citation information has shown to benefit many text mining tasks including information retrieval, information extraction, summarization, and question answering.

For example, citation indexing has been used to associate citing articles with cited ones, and the associations have been used to score Science Citation Index to measure the impact factors of scientific journals and articles [82]. Two articles can be considered as “related” if they share a significant set of co-citations and a study that incorporated this model has shown to improve information retrieval [83]. The number of times a citation is cited in a paper may indicate its relevance to the citing paper [84, 85]. Citances (or the citation sentences) sometimes represent the condensed semantic content of the documents they identify [86, 87] and have been used to extract scientific fact[30] and for summarization [87]. In addition, citation indexing has been used to model the evolution of author and paper networks [88]and research collaboration[89].

In order for text mining systems to benefit from citation information, one must automatically identify citations from full-text articles and extract their fields, including Author, Title, Journal and Year. Citation parsing automatically parses a full citation into its fields. Co-authorship is common in the biomedical domain, and it is important for a citation parser to identify the information of each author, including given name and surname. Separating an author's surname from the given name will enable a text mining system to separate two different authors (e.g., "John Smith" and "Smith John") who share the same names.

Citation parsing is challenging because citations come with different formats that are rooted in either different requirements by different publishers or non-standardized formats introduced by authors. The following examples illustrate some variations in citation format:

[Example 1] Yu, H and Lee M. 2006. Accessing Bioscience Images from Abstract Sentences. Bioinformatics. Vol 22 No. 14, pages e547-e556.

[Example 2] Hong Yu and Minsuk Lee. Accessing Bioscience Images from Abstract Sentences. Bioinformatics. Vol 22 No. 14, pages e547-e556. 2006.

[Example 3] Yu H, Lee H. 2006. Accessing Bioscience Images from Abstract Sentences. Bioinformatics: 22 (14), e547-e556.

First, the order of fields may vary. As shown in Examples 1 and 2, the publication year (2006) may appear before or after the title. There are also different ways to present publication volume and issues, for example, “Vol 22 No. 14” and “22 (14)” in Example 2 and 3. Author names also come with different format (e.g., “Yu H”, “Hong Yu”, “Yu, H” or “H. Yu”); Journal names vary as well. For example, an article that is published in “American Medical Informatics Association Fall Symposium” can be referenced as “AMIA”, “Proceedings of AMIA”, “Proceedings of the American Medical Informatics Association Fall Symposium”, “Proceedings of the American Medical Informatics Association (AMIA) Fall Symposium”, or “Proc AMIA” These variations pose a significant challenge for developing a natural language processing system that automatically parses citations into fields, since we can see from the example each field may have different formats.

In this chapter we report the development of a natural language processing system that automatically parses a citation into its fields (i.e., Author Given Name, Author Surname, Title, Source, Year, Volume, First Page, and Last Page). Although citation parsing is not new, and tools for doing it have been widely used in $CiteSeer^X$, few tools are publicly available and little work has been evaluated for citation parsing in the biomedical literature. This chapter is based on the publication [90].

3.2 Background

The Institute for Scientific Information (ISI) constructs citations and indexes scientific articles to produce multidisciplinary citation indices including the Science Citation Index (SCI). However, the database is mostly built manually. Similarly, HighWire Press[91] and other open journal projects link citations across journal sites. It is unclear whether they provide tools to automatically recognize citations. An Autonomous Citation Indexing (ACI) system automatically locates articles, extracts citations, identifies identical citations that occur in different formats, and identifies the fields of citations and has been developed and implemented by CiteSeer (now

CiteSeer^{X}) [92, 93]. Citation parsing is a subtask of ACI.

Earlier work in citation parsing uses manually crafted rules and external databases. CiteSeer is such an application that is built upon a set of heuristic rules (e.g., the format of any citation is uniform within one article) and external databases of author names and journal names. They reported a performance of 80.2% for Title, 82.1% for Authors, and 44.2% for Page Number. Wei (2007) [94] reported a similar approach. Besagni and Belaid (2004) [95]also developed a rule-based system. Their system assigned each field a tag (e.g., alphanumeric string and common word, capital initial) that resembles a part-of-speech tag. Rules (e.g., ”when author names are given, it is always at the beginning of the reference”) were used to detect each field. They evaluated their system on a dataset of 2,575 citations that came from 64 articles randomly selected from 140 journals, and reported that 75.9% references were completely parsed.

Powley and Dale (2007) [96] observed that the pattern of a citation mentioned in the body of a full-text article (e.g., ”Yu and Lee, 2006”) is consistent with the citation pattern in the reference section, and accordingly they developed rules to capture the author name. They reported 92% precision and 100% recall, although the system works well only when the citation is presented by author name and year, not by other common patterns (e.g., the digit format (e.g., ”[1]”)). Other related work link citations to the same reference that differ in format[97].

Statistical and machine-learning approaches have been reported for citation parsing. Takasu (2003) [98] developed a Hidden Markov Model (HMM) for citation parsing and reported over 90% accuracy on an evaluation data of 1,575 citations. BibPro [99] is a citation parser that is based on sequence alignment techniques. Specifically, they used sequence alignment tool BLAST to find the most similar citation fields for a citation, and subsequently parsed the citation and reported an accuracy of 97.68%. Kramer et al (2007) [100] developed Probabilistic Finite State Transducers (PFSTs) for citation parsing. Based on the observation that the content of a citation field is independent of other fields, they built an FST model for each citation field. An evaluation on the Cora dataset that serves as a common benchmark for accuracy measurements yields a field accuracy of 82.6%. Peng and McCallum (2006) [101] parsed citations with Conditional Random Fields (CRFs), and reported an F1-score ranging from 76.1% (for Publisher) to 99.4% (for Author), although their system is not made available to the public. ParsCit [102] is an open-source CRF-based citation parser that has been used by $CiteSeer^X$ [4]. The model was reported to be trained on 200 reference strings sampled from computer science publications. Their performance is 95% for F1-score on Cora dataset. Most systems (except for ParsCit and the system developed by Kramer et al [100]) described above were not available publicly. None of the systems developed was evaluated in the biomedical domain. Furthermore, none of the systems described above extracted author's surname and given name as we do in this study.

3.3 Methods

We developed the citation parsing system based on conditional random fields model [103], which are probabilistic models that relate to the HMM. The CRF model has an advantage over the HMM in that it relaxes strong independece assumptions [104], and as a result, has shown to work well with biomedical sequence data that frequently comes with variations. For example, research has found that CRFs performed the best in biomedical named-entity recognition tasks[105]. In the following, we first define the problem then describe our experiments on applying CRFs for citation parsing.

3.3.1 Problem Definition

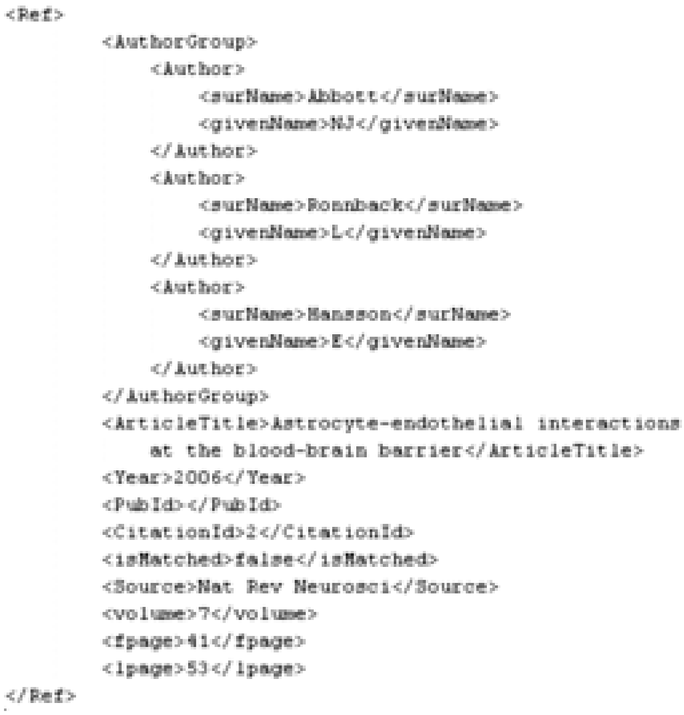
We define a full citation as a citation that incorporates four or more of the following fields: Author (further separated by Surname and GivenName), Title, Source (i.e., journal, conference, or other source of publication), Volume, Pages (we further separated the page information by FirstPage, and LastPage), and Year.

3.3.2 Data

Creating training and testing data set that can be used for applying supervised machine learning is one of the key components of our work. We automatically extracted the training and testing data from articles in the PubMed Central (PMC). PMC is a free digital archive of biomedical and life sciences journals. As of February 2009, the PMC Open Access article collection has incorporated a total of 794 journals and 12,537 full-text articles. We found that the average number of articles per journal was 154, with standard deviation 1,302, the maximum number of articles per journal was 32,881, and the minimum was 1. We found that for certain articles, PMC provides two formats of a full-text article: the parsed XML format from which we can extract the citation fields, and the HTML format from which we can extract the original full citation. Figure 3.1 and Example 4 show the two formats of a citation. The articles with both the XML and HTML representations were selected to extract citations to use as the data for this study.

[Example 4] Abbott NJ, Ronnback L, Hansson E (2006) Astrocyte-endothelial interactions at the blood-brain barrier. Nat Rev Neurosci 7:41-53

In order to ensure that our data was representative, we randomly selected 2% of the articles with the XML format from every journal. If a journal incorporated 49 articles or fewer, we randomly selected one article from that journal. Our selection resulted in a total of 2,988 articles, from which we were able to identify 672 articles that have the corresponding HTML citation format. Those 672 articles were used



```
<Ref>
        <AuthorGroup>
            <Author>
                <surName>Abbott</surName>
                <givenName>NJ</givenName>
            </Author>
            <Author>
                <surName>Ronnback</surName>
                <givenName>L</givenName>
            </Author>
            <Author>
                <surName>Hansson</surName>
                <givenName>E</givenName>
            </Author>
        </AuthorGroup>
        <ArticleTitle>Astrocyte-endothelial interactions
            at the blood-brain barrier</ArticleTitle>
        <Year>2006</Year>
        <PubId></PubId>
        <CitationId>2</CitationId>
        <isMatched>false</isMatched>
        <Source>Nat Rev Neurosci</Source>
        <volume>7</volume>
        <fpage>41</fpage>
        <lpage>53</lpage>
</Ref>
```

Figure 3.1: Gold Standard XML.

to extract full citations and their labeled fields, and they incorporated an average of 41 citations (Min 1, Max 333, and Standard Deviation 33). The total number of citations is 27,606, which were used as the tagged citation data for training and evaluation.

We chose Mallet [106]as our CRF package for implementation. The overview of the citation parsing system is shown in Figure 3.2.

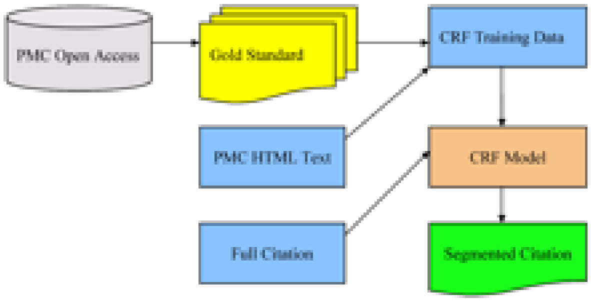


Figure 3.2: The Overview of the Citation Parsing System.

3.3.3 Learning Feature

We use Abner's [105] default features to train the CRF model. Abner's features capture the morphological characters of each token, for example, whether the token incorporates a capital letter, all capital letters, a digit, all digits, and other symbols (e.g., Roman and Greek characters and "-"), as well as the length of the token. We found that the features can be effective in discriminating different fields. For example, a combination of upper- and lower-case letters will lead to the detection of Author (”Hong”).

3.3.4 System and Evaluation

We used the 27,606 citations extracted from PMC for a 10-fold cross-validation to take the advantage of it: All observations are used for both training and validation, and each observation is used for validation exactly once. We split each citation into a series of tokens (each token contains a word or punctuation) and tagged them following the tag definition style used by the CoNLL shared task [107]. We defined the beginning of each field, inside the field, and OTHERS by ”B-<fieldname>”, ”I-<field name>” and ”O”, respectively. For example, B-TITLE refers to the word at the beginning of a title. The tags of FirstPage and LastPage are FPAGE and LPAGE. Tages SN and GN denote Surname and GivenName. We evaluated the test result for each field by recall, precision and F1 score, defined as F1 = (2 *Precision * Recall) / (Precision + Recall). Recall is the number of correctly predicted fields divided by the total number of annotated fields, and precision is the number of correctly predicted fields divided by the total number of predicted fields. These measures are per-entity.

Table 3.1: Recall, Precision, and F1 score for automatic citation field recognition.

Field	Precision	Recall	F1
TITLE	0.9546	0.9524	0.9535
SOURCE	0.8722	0.8703	0.8713
YEAR	0.9964	0.9977	0.9971
SurName	0.9951	0.9977	0.9964
GivenName	0.9794	0.9797	0.9796
VOL	0.9915	0.9932	0.9924
FirstPage	0.9966	0.9944	0.9955
LastPage	0.9991	0.9940	0.9965
OVERALL	0.9794	0.9796	0.9795

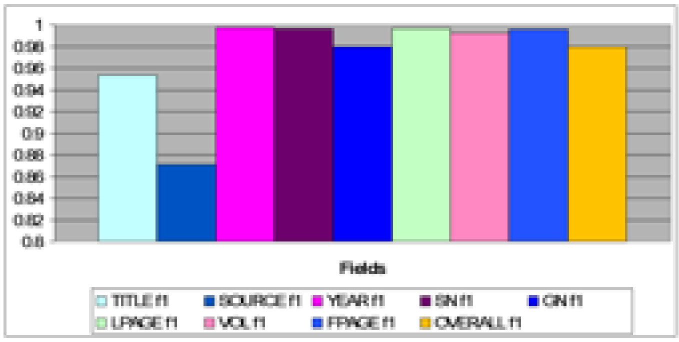


Figure 3.3: The F1-score of each field.

3.4 Results

Table 3.1 shows the recall, precision, and F1-score of the 10-fold cross-validation. The overall F1 score is 0.9795, and the YEAR, SurName, VOL, FirstPage and LastPage all have F1 above 0.99. On the other hand SOURCE has lowest F1 score 0.8713. Figure 3.3 is the corresponding F1 score of each field and the overall score.

3.5 Error Analysis and Discussion

In order to understand the source of errors, we examined those fields that were predicted to be wrong and summarized the patterns in them. For example, one pattern that our system wrongly predicted is that a word from the SOURCE field is a word from the TITLE field. We identified a total of 101 such patterns. We found that

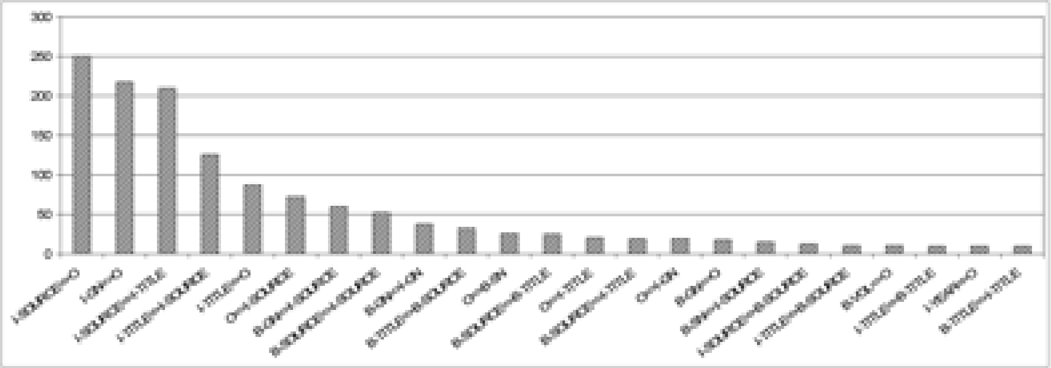


Figure 3.4: The number of wrongly predicted fields by pattern.

84.6% of the total number of errors fell into a total of 23 patterns. Figure 3.4 shows the number of instances of those patterns.

We further manually examined the source of errors for frequently-occurring patterns. We found that one source of error was introduced by the ambiguity of the period symbol ".". For example, in SOURCE, the period not only represents the boundary of a source (e.g., "Nature Neuroscience."), but also appears in the abbreviated version of journal titles (e.g., "Nat. Neurosci."). The ambiguity resulted in some inconsistent tagging in the training data as well. Similarly, the ambiguity of the period applies to other fields, for example, the period that appears in the AUTHOR field (e.g., "Haynes, J. D.). Our system sometimes confuses SOURCE with TITLE. We speculate that this error was in part introduced by the fact that some full citations have a missing title, and some full citations do not separate the title from the source. For example, for "'Colquhoun,D.; Sigworth, F.J. Fitting and statistics of single-channel records analysis Single-Channel Recording 1983.pp.191-263.", the "Single-Channel Recording" is the source and directly follows the title.

We reported in the related work that has shown different approaches for related tasks. We found that our approaches achieved the highest performance among all. We did not implement other approaches because it would be expensive for us to re-implement the systems. Our system was trained in the bioscience domain and certain word features are domain-specific. As a result, we speculate that our system may not work well in other specialized domains such as mathematics and physics. However, the methods are generalizable and one can apply the same methods we developed to data in another domain.

3.6 Conclusion

In this study, we developed an open-source package that automatically parses a citation into its fields in full-text biomedical articles. We applied and implemented a supervised machine-learning system based on Conditional Random Fields (CRFs) for citation parsing and report 97.95% F1 score to parse a citation into a total of eight fields. Our results show that CRFs are efficient machine learning model for citation parsing.

Chapter 4

Link Prediction for Research Collaboration Recommendation

4.1 Introduction

Millions of researchers contribute to biomedical research, collectively publishing tens of millions of research papers. These researchers are interlinked through a network of publications. Since biomedical research is a growing interdisciplinary field, it requires successful collaborations, which usually generate high impact work [108, 109, 110, 111]. Such collaborations span the physical and quantitative sciences, as well as basic, translational, and clinical research. It has been found that the average number of collaborators in the biomedical field is twice that in physics and more than four times that in mathematics [51]. Remarkably, the ground-breaking work of the complete sequencing of Yeast (Saccharomyces cerevisiae) genes involved over 600 scientists from North America, Europe and Japan [112].

The importance of scientific collaboration has accelerated the development of researcher profile platforms, most of which focus on facilitating institutional collaborations. Such platforms, including Harvard Catalyst Profile [113], SciVal Experts

[114], and ProQuest Pivot [115], integrate research and collaboration information - including publication history, co-authorship connections, research topics, and funding information - making it easier to find potential collaborators. In addition, semantic Web resources, such as VIVO [116], have been developed to provide a general scheme to describe researcher profiles so that these can be embedded in particular applications. Similarly, online communities, such as BiomedExperts [117], allow users to upload their personal profiles and help them make new connections.

Few systems, such as Harvard Catalyst, however, have the functionality of recommending collaborators automatically. Such services are, however, important for researchers, especially junior researchers, whose work depends upon successful collaborations. Traditional ways of finding a collaborator, such as socializing at a scientific conference or being introduced by a mutual colleague, are found to be increasingly insufficient. Today, emails and social networking allow researchers to establish varied collaborations, many of which do not require a face-to-face meeting. For example, one of the author's (Yu) group has collaborated with hundreds of biomedical researchers who were approached by email [118]. However, such approaches are ad hoc and may not lead to long-term collaborations.

We formulate research collaboration prediction as a link prediction problem in the co-authorship network. Since joint publication is one of the most effective representations of collaboration, co-authorship indicates a collaborative relation. This problem is illustrated in Figure 4.1, where author s has collaborated with authors a, c, and e and we would like to know the probability that s will collaborate with b, f, and d.

Link prediction has been studied in social networks. Liben-Nowell and Kleinberg [66] reviewed various topology predictors for link prediction. They ranked author pairs by a particular predictor and considered top-ranked potential collaborations in the future. Al Hasan, et al. [119] explored supervised machine learning approaches to classify author pairs using models trained by previous collaborations. They explored

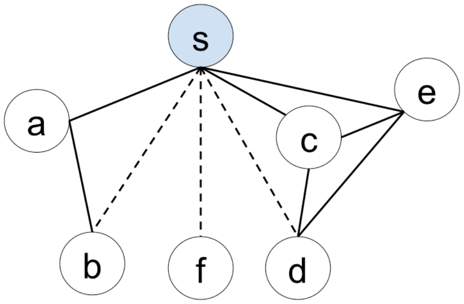


Figure 4.1: An illustration of automatic research collaboration recommendation. The graph shows a co-authorship network in which the nodes are authors and the links represent co-authorship. The solid lines represent existing co-authorships. Our study is to build a computational model to predict whether author s will collaborate with authors b, f, and d based on their existing research and collaborations.

features such as keyword overlap between two authors' publication histories and the shortest path between two authors. Backstrom and Leskovec [120] proposed an approach that uses supervised machine learning to predict the weight of a connection that later guides a random walker starting from a particular node in the network. The stationary probability that the walker lands on an author is the chance of establishing the connection. All the aforementioned approaches have limitations, however. The topological prediction approaches [66] are made by a single network feature, ignoring the combined effect of different network factors. The hybrid models [120], although theoretically sound, are computationally expensive, limiting their real world applications. Although the supervised machine learning approach [119] is robust, the features thus far explored by others have been limited.

In this chapter, we built an Automatic Research Collaboration Recommendation (ARCR) system to predict potential first-time collaborations. ARCR was built on supervised machine learning models. Our contributions are: Firstly, we explored novel learning features derived from the semantic content of an author's research profile. Our computational approaches and learning features are computationally inexpensive, making them applicable to the big data challenge of scientific collaboration recommendation. Secondly, we evaluated our approaches on various datasets reflecting data sparseness, which is a common problem in the real world. Finally, we analyzed important factors including research interest and mutual collaborators that contribute to biomedical collaborations.

This chapter is based on the publication [121].

4.2 Background

Three types of link prediction approaches have been reported. Topology-based prediction utilizes network structure, including the connectivity and similarity of neighbor nodes. Liben-Nowell and Kleinberg [66] built co-authorship networks in physics and explored topological predictors, including the Jaccard and Adamic/Adar coefficients, preferential attachment, and random walk for co-authorship prediction.

Supervised machine learning approaches for link prediction have also been developed. Al Hasan et al. [119] explored supervised machine learning models, including naive Bayes and support vector machines (SVMs). They explored topological features (e.g., the number of common co-authors) and simple semantic features (e.g., the overlap of the keywords of two authors publication profiles). Sun et al[122] extensively studied topological features in the heterogeneous networks consisting both co-authorship and citation relations to predict co-authorship in DBLP databases. A Markov random field model[123] and a Markov logic model[124] can also be used for link prediction, as they both have the ability to model relations[125]. Huang, et al.[126] modeled collaborations as a Poisson process and developed a stochastic model to predict them.

Hybrid systems integrate supervised machine learning and topology-based prediction. For example, Backstrom and Leskovec[120] applied supervised machine learning approaches to predict the strength of a connection (or a collaboration in a research network). The predicted weight is subsequently used to guide the random walk. The stationary probability of landing on a particular node is considered the chance of a connection from the starting node. Wang et al[127] modeled the local topological structure by Markov Random Field to infer the co-occurrence probability of two nodes, and subsequently use it together with other topological and semantic features for link prediction.

A significant amount of work has analyzed network structures and their evolution. For example, Adamic and Huberman[128] concluded that the growth of the Web follows a power law distribution. The probability of a new node connecting with a given node is proportional to the degree of the node. This phenomenon, also frequently called preferential attachment, has also been observed in co-authorship networks, as well as other types of social networks [51, 54]. Small world is another pattern in various networks [62, 129]. A social network, such as co-authorship network, consists of both structured (close neighbors) and random contacts and one can navigate from one node to another with very few steps. Newman[54] found that only five to six steps are needed to navigate from one randomly chosen scientist to another in a community. In addition, social networks appear assortative, meaning that nodes tend to connect to other nodes of similar characteristics, such as the degree[130].

In addition to the above general structure and dynamics of networks, domain-specific co-authorship networks have been studied. For example, Newman [54] compared the co-authorship network in biomedicine with that in physics and found differences. In the biomedical domain, they found it is less common that two researchers collaborate when they have a mutual collaborator than in physics. In addition, the study found that in the biomedical domain the network structure is dominated by many little people with few collaborators, instead of a few people with many collaborators, as in other domains, such as computer science. Newman[52] found that two researchers are more likely to collaborate if they have had more collaborations in the past. Newman [51] showed that biomedical research has the highest degree of collaboration, compared with physics and mathematics. Huang, et al.[126] found that the collaboration pattern and its evolution in the computer science domain is more similar to the mathematics domain than to biology. Ding[131] found that, in the information retrieval field, productive authors tend to collaborate with and cite researchers who have the same research interests.

Factors that lead to successful collaborations have also been studied, including various social and environmental factors such as leadership, geographical proximity, and the personalities of the team members. For example, one study concluded that a leader in a research field typically plays an important broker role to bridge people from different disciplines [111]. Physical proximity between first and last author was found to be positively related to the impact of collaboration which was measure by the citation received based on the data collected from Harvard Medical School publications and office locations [132]. They argued that close geographical distance is important for the outcome of the collaboration. International collaborations, however, are also a positive factor for the impact. As found in [133], the average number of citations increases with the number of affiliated countries. Certain characteristics of team members, such as openness and flexibility, also contribute to the success of the collaboration [111, 134]. Our work is most closely related to the work of Al Hasan, et al. [119]. However, unlike their approach which mainly explored topological features, we explore a wealth of semantic features derived from the author profile, including publication history similarity, citation similarity, and common co-authors, and we show that these semantic features significantly improve the research collaboration predictions.

4.3 Method

4.3.1 Features

We derived a novel, comprehensive feature set to model aspects of multiple collaborations. We explored five categories of features, as described below in detail: Research Interest Profile, Publication Productivity, Author Seniority, Network Connectivity. The formal definitions of each feature are given in Table 4.1.

4.3.1.1 Research Interest Profile

The Research Interest Profile similarity of two researchers may be indicative of the likelihood of a research collaboration: if two researchers have similar research profiles, they may be more likely to collaborate. In contrast, inter-disciplinary research requires collaborations between two researchers with distinct research profiles. This category of features will capture both types of research collaboration.

We build the Research Interest Profile of a researcher by aggregating all information about his/her publications, including abstracts, the assigned Medical Subject Headings (MeSH) terms, and the citations to that work and by that work. Taken together, these three characteristics comprehensively cover the author's research interests: the abstract is a summary of the article; MeSH terms represent the main topics of the article; and out-citing citations (other articles cited by the article) show the relevant background information of the article while the in-citing citations (other articles that cite the article in question) represent the recognition of the work by their peers. We assume the abstract is sufficient for representing the main content of the article and therefore did not include full-text as a data source for the research interest profile; it is also more generalizable to use the abstract instead of the full-text as the abstracts are more freely available in knowledge bases.

We use a term frequency-inverse document frequency (TF-IDF) term vector to represent the abstract, in-citing citation, and out-citing citation profiles. Assume the abstracts of all the publications of an author by a certain year are D. Thus, the aggregated TF-IDF for term t in D is

$$tfidf(t, D) = idf(t, D) * \sum_{d \in D} tf(t, d)$$

, where $idf(t, D)$ is the inverse document frequency of term t in the abstract collection D and $tf(t, d)$ is the term frequency of term t in abstract d. An author's in-citing and out-citing citation profiles are computed similarly by considering the document set D above as the set of in-citing citations and out-citing citations, respectively. We treat MeSH terms as keywords, and the MeSH profile is created by collecting all the MeSH terms as a set from the author's publications. We used this simplified representation for the MeSH profile instead of computing the MeSH TF-IDF vector as our preliminary analysis found the TF-IDF representation did not improve the performance significantly.

We then define the features simText, simOutcite, and simIncite as the cosine similarity of two researchers' abstract profiles, out-citing citation profiles, and in-citing citation profiles, respectively. Specifically,

$$simText(x, y) = \frac{abstract(x) \bullet abstract(y)}{|abstract(x)||abstract(y)|}$$

where abstract(.) is the TF-IDF term vector of the author's publication history. Similarly,

$$simOutcite(x, y) = \frac{outcite(x) \bullet outcite(y)}{|outcite(x)||outcite(y)|}$$

and

$$simIncite(x, y) = \frac{incite(x) \bullet incite(y)}{|incite(x)||incite(y)|}$$

where outcite(.) is the TF-IDF term vector of the author's out-citing citations from the publication history, while incite(.) is the TF-IDF term vector of the authorâĂŹs in-citing citations of the publication history. We also define simMeSH as the Jaccard coefficient of the two researchers' MeSH profiles. Concretely,

$$simMeSH = \frac{|MeSH(x) \cap MeSH(y)|}{|MeSH(x) \cup MeSH(y)|}$$

where MeSH(.) is the set of MeSH terms from the author's publication history. simMeSH serves as a baseline feature as it has been found to be a well-performing feature for collaboration prediction in[119]; please see the Baseline Model Comparison section below for more details about the baselines.

4.3.1.2 Publication Productivity

Our initial data exploration of the MEDLINE database showed that the number of publications and the number of co-authors increases yearly as shown in Figure 4.2, suggesting a positive relation between co-authorship and research publication productivity. We therefore hypothesize that a researcher's productivity is related to the likelihood of their research collaborations. We used the average number of publications per year to measure productivity and introduce the feature *sumPub* as the sum of two researchers' average publications. Formally, it is defined as

$$sumPub(x, y) = avgPub(x) + avgPub(y)$$

where avgPub(.) is average publication per year for an author.

In addition, we attempted to model how active a researcher has been in the near past and defined an author's recency as the sum of the inversed publication time distances to the present, which favors more recent activity because we hypothesize that recent activity is more related to future activity. The recency of author x is

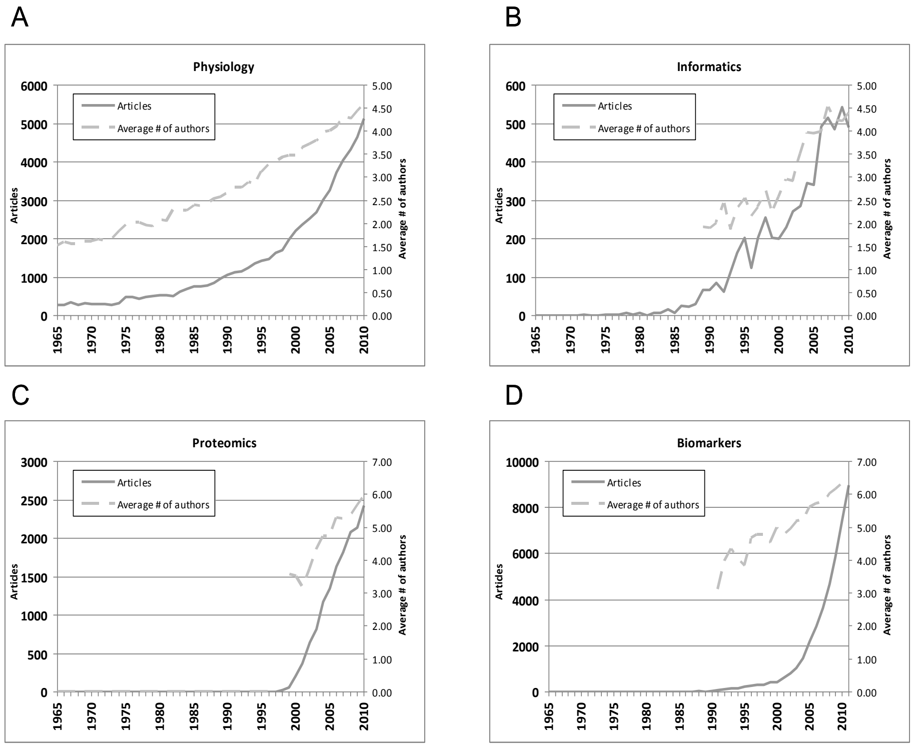


Figure 4.2: Articles and authors over years in four biomedical research fields. As shown in the figure, number of new articles (solid line) and average number of authors per article are generally increasing.

defined as

$$recency(x) = \sum_{i \in papers(x)} \frac{1}{t_i}$$

where papers(x) is the publications of the author x, and t_i is the difference between the publication year of paper i and the present year; we add 1 to account for a difference of 0. We use the inverse of the time difference to favor the more recent publishing activities as they are likely more related to future collaborations. The *sumRecency* is thus the sum of two researchers' recency scores.

4.3.1.3 Author Seniority

The possible hierarchical relationships between two researchers include persistent junior-senior relations (e.g., a student and advisor who always publish in the same order) and more reciprocative collegial relations, which may be important for predicting future research collaborations. We define an author's seniority as the average number of times that author has been a senior author, which is defined as being the corresponding author on the paper. The Seniority of author x is defined as

$$seniority(x) = \frac{1}{|papers(x)|} \sum I$$

where I(.) is the indicator function and equals to 1 if x is the corresponding author of the particular publication i, and is 0 otherwise. We normalize this sum by the number of papers published by x. For example, assume an author had 5 publications and was the corresponding author on 2 of them, then the author's seniority is 2/5. The feature diffSeniority is thus defined as the seniority difference between two researchers.

4.3.1.4 Network Connectivity

We hypothesize that two researchers are more likely to collaborate with each other if their network connectivity is high. We identified multiple features from this network topology: we create the novel *sumCoauthor* feature, which is the sum of each researcher's average number of unique co-authors per year and represents how active a researcher is in collaborating with others. It is defined as

$$sumCoauthor(x, y) = avgCoauthor(x) + avgCoauthor(y)$$

where avgCoauthor(.) is the average number of unique co-authors per year for an author. The feature *numCommonCoauthor* is the number of co-authors two researchers have in common; it has been found [51] that two scientists with a common collaborator are approximately 45 times more likely to co-author a paper than two scientists who have no common collaborator. Therefore, we also use *numCommonCoauthor* as a baseline feature along with simMeSH. *numCommonCoauthor* of authors x and y is defined as

$$numCommonCoauthor(x, y) = |\Gamma(x) \cap \Gamma(y)|$$

where $\Gamma(.)$ is the set of co-authors. The feature *sumClusteringCoef* is the sum of each researcher's clustering coefficient [52], which is a measure of the probability that a researcher's collaborators have collaborations among themselves; *coauthorJaccard* [135] is the number of co-authors two researchers have in common normalized by the total number of their unique co-authors; and Adamic [136, 66] measures the similarity of two nodes by their common neighbors. For authors x and y,

$$Adamic(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log|\Gamma(z)|}$$

where z is the common neighbor of x and y and $\Gamma(z)$ is the set of co-authors of z. For example, assuming researcher z is the only common neighbor of x and y, z has no connections other than x and y, the Adamic value of z is 1/log(2). On the other hand if z has 3 connections in addition to x and y, the value becomes 1/log(5). These features are also formally summarized in Table 4.1.

4.3.2 Models

We formulate research collaboration prediction as a classification task. We explore five supervised machine learning models: naive Bayes, naive Bayes multinomial, Support Vector Machines (SVMs), logistic regression and K-nearest neighbors (KNN), all commonly used for classification tasks. A naive Bayes classifier is a probabilistic classifier based on Bayes theorem with the independence (naive) assumption that the

Table 4.1: Feature Definition.

Category	Feature	Definition
Research Profile Similarity	simText	$\frac{absract(x).abstract(y)}{\lvert abstract(x)\rvert\lvert abstract(y)\rvert}$
	simMeSH	$\frac{\lvert MeSH(x)\cap MeSH(y)\rvert}{\lvert MeSH(x)\cup MeSH(y)\rvert}$
	simIncite	$\frac{incite(x).incite(y)}{\lvert incite(x)\rvert\lvert incite(y)\rvert}$
	simOutcite	$\frac{outcite(x).outcite(y)}{\lvert outcite(x)\rvert\lvert outcite(y)\rvert}$
Publication Productivity	sumPub	$avgPub(x) + avgPub(y)$
	sumRecency	$recency(x) + recency(y)$
Seniority	diffSeniority	$\lvert seniority(x) - seniority(y)\rvert$
Connectivity	sumCoauthor	$avgCoauthor(x) + avgCoauthor(y)$
	numCommonCoauthor	$\lvert\Gamma(x) \cap \Gamma(y)\rvert$
	sumClusteringCoef	$clusteringCoef(x) + clusteringCoef(y)$
	coauthorJaccard	$\frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\lvert\Gamma(x)\cup\Gamma(y)\rvert}$
	Adamic	$\sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{log\lvert\Gamma(z)\rvert}$

features are generated independently from each other, given the instance label[137]. The naive Bayes multinomial model assumes the conditional probability of the feature, given a class, follows a multinomial distribution [138][139]. SVMs are based on the concept of maximum margin decision planes that define generalizable decision boundaries for classification and regression. An SVM constructs a hyperplane to maximize the margin between the data points and the hyperplane, often after mapping the data points to a higher-dimensional space in which they are linearly separable or close to it[140]. We explore an SVM model with the widely-used linear kernel for its efficiency. Logistic regression estimates discrete or continuous value parameters to predict discrete category values. The probabilities that describe the possible class of a single instance are trained as a function of explanatory variables, using a logistic function[137]. These four classifiers are not only the well-studied models in a variety of classification tasks[141], but also widely available in open source software communities. In addition we used K-nearest neighbor model as it particularly learns non-linear decision boundaries and is easy to interpret[141][142]. We use data mining software Weka[143] to build and evaluate naive bayes, naive bayes multinomial and logistic regression models, LIBSVM[144] for SVM, and the Python machine learning package Scikit [145] for KNN.

4.3.3 Baselines

We compared our approaches with three competitive baseline systems. The first baseline model, *RandomBaseline*, predicts positive instances based on the distributional probability in the training dataset. Specifically we predict an instance as positive with the probability that equals to the percentage of positive instances in the training set. It evaluates how random guess based on prior knowledge perform on the prediction.

The second baseline model, called *PreferentialAttachment*, is based on network topology. As described in the background section, preferential attachment is a well-studied network growth pattern. The more existing links a node has, the higher the chance a new node will link to it. We implemented this baseline based on Liben-Nowell and Kleinberg's description [66]. Specifically $score(x, y) = |\Gamma(x)||\Gamma(y)|$, where $\Gamma(.)$ represents the neighboring nodes as defined in 4.3.1. For each pair of nodes in a testing set, we computed the corresponding $score(x, y)$. The higher the score, the larger the chance that the two nodes and will connect (or collaborate).

We also used *JaccardBaseline*, which describes the importance of the common coauthor in the author pair, as the third baseline model as it has demonstrated strong performance in previous research [66]. Its definition is the same as for the feature $coauthorJaccard$.

4.3.4 The CiteGraph Dataset

We used the citation/co-authorship network database CiteGraph[68] as our data source. The CiteGraph database is comprised of 1.6 million full-text articles, a joint set of the Elsevier database (1899-2011) and the MEDLINE database. We created the CiteGraph Dataset by only selecting those articles that were at the intersection

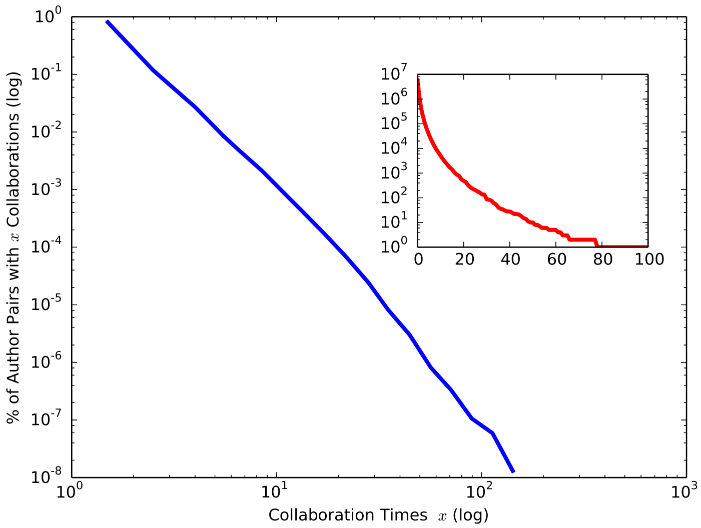


Figure 4.3: Collaboration frequency distribution for the CiteGraph Dataset. The distribution is highly skewed and most of the author pairs (80.5%) collaborated only once in our dataset. The larger, outer plot shows the log-log scale while the smaller, inset plot shows the skewed log-linear scale.

of both the Elsevier and the MEDLINE databases as each database had distinct attributes and we needed the overlap of all the attributes for feature extraction. Using this combined dataset, we were able to gather the title, author(s), abstract, full text, year of publication, and MeSH terms, the in-cites, and the out-cites for each article. In addition, we disambiguated the author names and created a co-authorship network for all the authors. Figure 4.3 shows the collaboration frequency distribution in our CiteGraph Dataset. As shown, a majority of researcher pairs (80.5%) collaborate only once, while less than 20% collaborate two or more times. The highest number of collaborations for the same researcher pair is 159, spanning 12 years.

4.3.5 Training Dataset

We used CiteGraph for both the training and testing data. We selected equal numbers of positive and negative instances for training and testing. The positive training instances are author pairs whose first collaborations took place in 2007 or 2008. The negative training instances are author pairs who did not collaborate before 2009. We

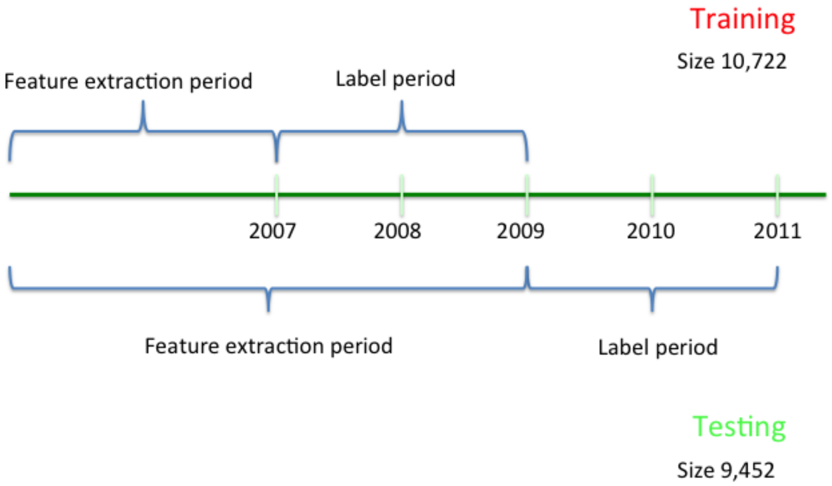


Figure 4.4: Feature extraction. We define positive training examples are the author pairs who collaborated in 2007 and 2008 (Label period), and the features are extracted based on the network status of 2007 (feature extraction period). The label period for testing data is 2009 and 2010.

randomly selected 10,000 positive and 10,000 negative author pairs and extracted each pair's features, and the sampling method is similar with the static graph sampling algorithm proposed in[146]. For some author pairs the article information, including the abstract, was not available. Therefore, we filtered out these pairs as well as tailored them to equal size for a final total of 5361 positive instances and 5361 negative instances. The combined group of 10,722 author pairs was used as the training set. The illustration of feature extraction is shown in Figure 4.4.

4.3.6 Testing Dataset

We created two testing datasets. The first testing dataset, *RandomPairCategory*, was created from a random selection of publications from 2009 and 2010. The positive instances were those in which the author pair first collaborated in 2009 or 2010 (Figure 4.4), while the negative instances were author pairs who never collaborated before 2011. We randomly identified a total of 10,000 positive and 10,000 negative author pairs. Of these, we found that 4726 positive and 4726 negative author pairs had

Table 4.2: *IndividualAuthorCategory* testing sets

Author	Number of Publications	Sub-graphSize	Positves	Negatives Sampled
Jeroen Bax	28	69,487	31	200
Mathew Farrer	10	66,876	13	200
Filippo Marte	59	418	11	200
Christodoulos Stefanadis	30	33,869	16	200

complete features. These 9452 author pairs were used as the testing set. Note that the selection method of *RandomPairCategory* also used the sampling method [146] that was utilized for the training data; therefore, the two datasets represent the same distribution.

The second testing dataset, *IndividualAuthorCategory*, was selected based on the collaboration network topology. We randomly selected four authors (target authors) with numerous publications (we set a minimum of 10) in 2009 and 2010. For each author, we built a sub-graph comprising three hops of a breadth-first traversal of the collaboration network established prior to 2011. We thus not only built a sub-graph, but also created the testing set with authors who are close topologically. The positive instances are collaborations established by authors (in the sub-graph) who collaborated with the target author during 2009 and 2010 and the negative instances are those (in the sub-graph) who did not collaborate with the target author before the end of 2010. The statistics of each sub-graph are shown in Table 4.2. When constructing the testing set for each author, we used all the positive instances and randomly sampled 200 negative instances for authors Jeroen Bax, Mathew Farrer and Christodoulos Stefanadis and 200 for author Filippo Marte. The *IndividualAuthorCategory* evaluation dataset complements the *RandomPairCategory* dataset because the former consists of author pairs who are more similar in research while the latter represents a broader selection of potential collaborators.

We calculate precision (TP/(TP + FP)), recall (TP/(TP + FN)), the receiver operating characteristic or ROC, the area under the curve of the true positive rate

(TPR) over the false positive rate (FPR), sensitivity (the same as recall), specificity (TN/(FP+TN)), and accuracy ((TP+TN)/ALL), where TP, FP, TN, FN and ALL stand for number of true positives, false positives, true negatives, false negatives, and number of total instances respectively. F1 score is defined as the harmonic mean of recall and precision, specifically 2*recall*precision/(recall+precision). In addition we use log loss [147] to measure the prediction cost of logistic regression model. It is defined as

$$J = \sum_{i=1}^{m} y_t log y_p + (1 - y_t) log(1 - y_p)$$

, where $y_t \in \{0, 1\}$ is class label, and $y_p = P(y_t = 1)$ is the predicted probability of being positive.

4.4 Results

4.4.1 10 Fold Cross Validation on Training Set

Table 4.3 shows the 10-fold cross-validation results on the training dataset. The logistic regression and SVM demonstrate the best performance, with a 0.878 ROC and 0.797 F1 for logistic regression and 0.878 ROC and 0.780 F1 for SVM. The naive Bayes model performs the second best, with an ROC of 0.838. The naive Bayes multinomial performed the worst among the models. Logistic regression as well as SVM outperformed the naive Bayes and naive Bayes multinomial models with statistical significance ($p < 0.05$, t-test). Table 10-fold cross-validation on the training set.

4.4.2 Testing Set 1

Table 4.4 shows the results of models that were trained on the entire training dataset and then tested on the *RandomPairCategory* testing set, which was created by ran-

Table 4.3: 10-Fold cross-validation on the training set.

Model	ROC	Precision	Recall	F1	Accuracy
NaÃŕve Bayes	0.838	0.798	0.708	0.684	0.708
NaÃŕve Bayes Multinomial	0.659	0.795	0.655	0.609	0.655
Logistic Regression	0.878	0.803	0.797	0.796	0.797
SVM	0.878	0.855	0.718	0.780	0.798
KNN (N =51)	0.858	0.868	0.636	0.734	0.769

Table 4.4: *RandomPairCategory* evaluation results.

Model	ROC	Precision	Recall	F1	Accuracy
Naive Bayes	0.819	0.786	0.694	0.667	0.694
NaÃŕve Bayes Multinomial	0.626	0.790	0.644	0.592	0.644
Logistic Regression	0.871	0.794	0.789	0.788	0.789
SVM	0.871	0.842	0.708	0.769	0.787
KNN (n=51)	0.850	0.854	0.632	0.726	0.762
RandomBaseline	-*	0.504	0.500	0.502	0.504
PreferentialAttachment	0.584	0.574	0.567	0.556	0.567
JaccardBaseline	0.639	0.789	0.639	0.585	0.639

*The value is unavailable due to the nature of the evaluation metric.

domly selecting author pairs published during 2009 and 2010. Consistent with the cross-validation results, the logistic regression and SVM outperformed the other models, yielding an ROC of 0.871 and an F1 of 0.789 for logistic regression and 0.871 ROC and 0.769 F1 for SVM. The *RandomBaseline* model had an F1 of 0.502 and 0.504 accuracy, while the topology baseline models *PreferentialAttachment* and *JaccardBaseline* yielded ROC values of 0.4583 and 0.278, respectively. All the supervised machine learning models outperformed the baseline systems. Since *RandomBaseline* was based on a prior distribution of classes, it has a single value and therefore no ROC is reported. Logistic regression outperformed the naive Bayes and naive Bayes multinomial models with statistical significance ($p < 0.05$, t-test).

4.4.3 Testing Set 2

We evaluated the best supervised machine learning model, logistic regression, on the *IndividualAuthorCategory* testing set, and the results are shown in Table 4.5. Our

Table 4.5: *IndividualAuthorCategory* evaluation results.

Author	ARCR ROC	Pref.Attach.* ROC	JaccardBaseline ROC
Jeroen Bax	0.917	0.634	0.620
Mathew Farrer	0.980	0.537	0.917
Filippo Marte	0.800	0.302	0.455
Christodoulos Stefanadis	0.766	0.548	0.313
Macro Average	0.866	0.505	0.576

*Pref.Attach stands for PreferentialAttachment.

Table 4.6: True positive predictions for author Filippo Marte.

Author Name	Collaborated Publication(PMID)
Carmelo Anfuso	18006092
Scipione Carerj	18199503
Fabio Minutoli	18006092
Concetta Zito	19324436
Sebastiano Coglitore	18006092
Ignazio Salamone	19395074
Roberto Gaeta	18178269
Marco Cerrito	18280595
Giuseppe Turiano	18579228

model has a ROC ranging from 0.766 to 0.980, while the best ROC for the baseline models was 0.634 for the prediction for author Jeroen Bax for the *PreferentialAttachment* model; the *JaccardBaseline* model performed best for predicting collaborators of Mathew Farrer, with an ROC of 0.917. The performance differences between ARCR and the baseline systems are both statistically significant ($p < 0.05$, t-test). A list of true positive predictions for author Filippo Marte is listed in Table 4.6.

4.4.4 Inter- vs. Intra-diciplinary Collaboration

We further broke down our datasets to examine the inter- and intra-disciplinary collaboration predictions separately. We split the training data using different values of *simText* as our threshold in order to approximate inter-discipline and intra-discipline collaboration. *simMeSH* can also be chosen as the discipline measure but we assume the abstract has more detailed information than keywords. The training set and the *RandomPairCategory* testing set were divided into inter-/intra-disciplinary training/testing sets using the threshold. We varied the threshold from *simText* values of 0.01 to 0.30 with 15 evenly distributed data points. For example, when the *simText* threshold was set to 0.01, author-pair instances with simText values less than 0.01 were categorized as inter-disciplinary whereas the author-pair instances with *simText* values greater than 0.01 were categorized as intra-disciplinary. For each threshold value, we trained inter- and intra-disciplinary learning models using their respective training sets and then tested them on the corresponding inter-/intra-disciplinary testing sets. There is very little data when *simText* is larger than 0.3 so we did not explore larger thresholds. A small threshold indicates restricting inter-disciplinary samples and relaxing intra-disciplinary samples, while a large threshold indicates the opposite. As shown in Figure 4.5, when simText < 0.19, the intra-disciplinary model has better performance according to F1. When *simText* > 0.19, the inter-disciplinary model outperformed the intra-disciplinary model.

4.4.5 Feature Ranking

To identify the features' contributions, we ranked them using information gain[148]. As shown in Table 4.7, the research interest features *simOutcite* and *simText* are the top-ranked features, both with information gain greater than 0.2. The features *coauthorJaccard*, *Adamic*, and *numCommonCoauthor* are the next top ranked, based on the common co-author count. The next features are *simMeSH* and *simIncite*, which also represent research interest. In contrast, the contributions of *sumClusteringCoef* and *sumPub* are considerably smaller and *diffSeniority* shows no contribution.

Table Training set feature ranking, by information gain.

In addition we calculated sensitivity and specificity for the logistic regression model evaluation on all the testing sets to help us understand error patterns in different datasets. As shown in Table 4.8, the *RandomPairCategory* testing set has lower

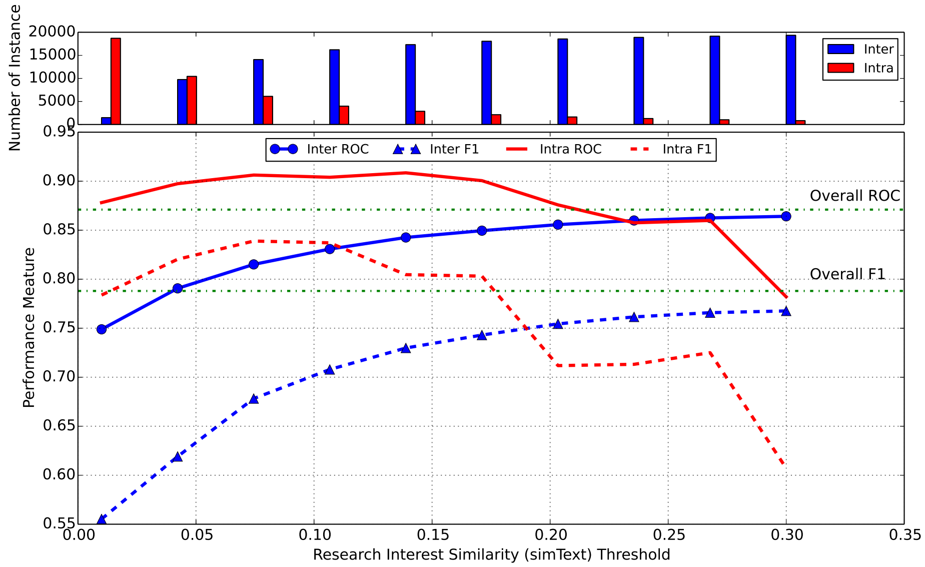


Figure 4.5: Intra- and inter-disciplinary collaboration prediction performance by ROC and F1 measurement. Training set and *RandomPairCateory* test set were divided by the threshold into inter- (<threshold) and intra-disciplinary (>threshold) training/test sets. For each threshold, we trained inter- and intra-disciplinary models and tested them on the corresponding inter-/intra-disciplinary testing sets. The histogram on the top is the number of instances (training+testing) of inter- and intra-disciplinary subset according to the threshold cutoff.

Table 4.7: Training set feature ranking, by information gain.

Rank	Feature	Infomation Gain
1	*simOutcite*	0.265
2	*simText*	0.202
3	*coauthorJaccard*	0.173
4	*Adamic*	0.173
5	*numCommonCoauthor*	0.173
6	*simMeSH*	0.145
7	*simIncite*	0.101
8	*sumCoauthor*	0.055
9	*sumRecency*	0.024
10	*sumPub*	0.022
11	*sumClusteringCoef*	0.002
12	*diffSeniority*	0

Table 4.8: Sensitivity and specificity for all testing sets by logistic regression model.

Testing Set		Sensitivity	Specificity
RandomPairCategory		0.718	0.859
IndividualAuthor Category	Jeroen Bax	0.968	0.345
	Mathew Farrer	1.000	0.125
	Filippo Marte	0.818	0.460
	Christodoulos Stefanadis	0.813	0.352
	Macro Average	0.900	0.321

sensitivity than the *IndividualAuthorCategory* testing set, while the former has higher specificity than the latter.

In our study we found research interest features to be an important feature category (Table 4.7) while they have not been well studied in other work. We therefore further analyzed their characteristics, such as their relation with author seniority. As shown in Figure 4.6, the research profile similarity features *simText*, *simIncite*, and *simOutcite* all increase as author seniority increases. This might indicate that, in the early stages of a researcher'ÂŹs career, collaborators with less research similarity are found but collaboration between two experienced researchers shows greater research interest similarity. In addition, the number of author pairs decrease as the authors get more senior, indicating fewer collaborations as the researchers become more senior. We ploted the careers within 15 years as the data becomes very sparse beyond 15 years.

4.5 Discussion

4.5.1 Models

The evaluation results for both the 10-fold cross-validation and the testing data (on the *RandomPairCategory* testing set) show that the logistic regression and SVM models perform best (logistic regression yielded an F1 score of 0.796 for 10-fold cross-validation and a score of 0.788 for testing while SVM produced 0.780 and 0.769

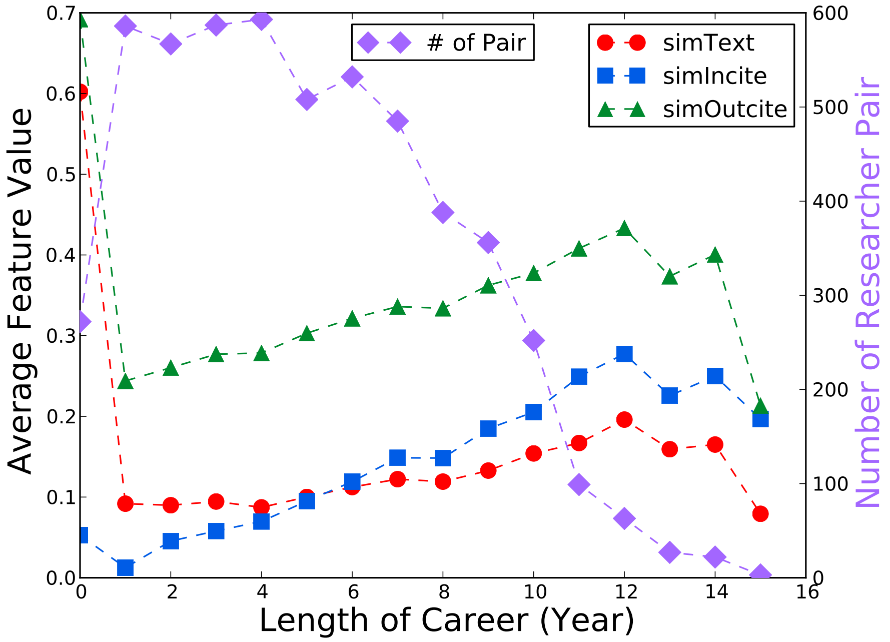


Figure 4.6: Research interest similarities over collaborator career length. In the early stages of a researcher's career, collaborators with less research similarity are found and collaboration between two experienced researchers shows greater research interest similarity.

respectively, as shown in Tables 4.3 and 4.4). *IndividualAuthorCategory* is a more challenging evaluation data set as we tried to predict collaborators for individual researchers from the candidates that were collected from their close neighbors in the network. ARCR outperformed all the baseline system with statistical significance (Table 5), which further shows that our model has the ability to recommend collaborators for the researcher. KNN model, which learns a non-linear decision boundary, did not perform as well as SVM and Logistic Regression as it only had an F1 score of 0.734 for the cross-validation on the training dataset and 0.726 on the RandomPairCategory testing set. This suggests that a linear decision boundary might be preferred.

Neither the naive Bayes nor the naive Bayes multinomial model perform well (an F1 score of 0.684 for 10-fold cross-validation and a score of 0.667 for the RandomPairCategory test set with the naive Bayes model and an F1 score of 0.609 for 10-fold cross validation and 0.592 for the *RandomPairCategory* test set with the naive Bayes multinomial model). The performance differences between logistic regression and the naive Bayes and naive Bayes multinomial models are both statistically significant ($p < 0.05$, t-test). A possible reason for this under-performance is that both models assume conditional independence, which might not hold in our study. For example, *coauthorJaccard* and *numCommonCoauthor* are related, since they both depend on the number of common co-authors.

4.5.2 Error Analysis

In order to determine if our data size or features are sufficient, we analyzed the learning curve for the logistic regression model. The training set was split into two sets: training (66% of total instances) and validation set, and log loss is used for the error metric for the curve. As shown in Figure 4.7 the training error and validation error converges by the time the dataset reaches the size of 1000 author pairs. Therefore

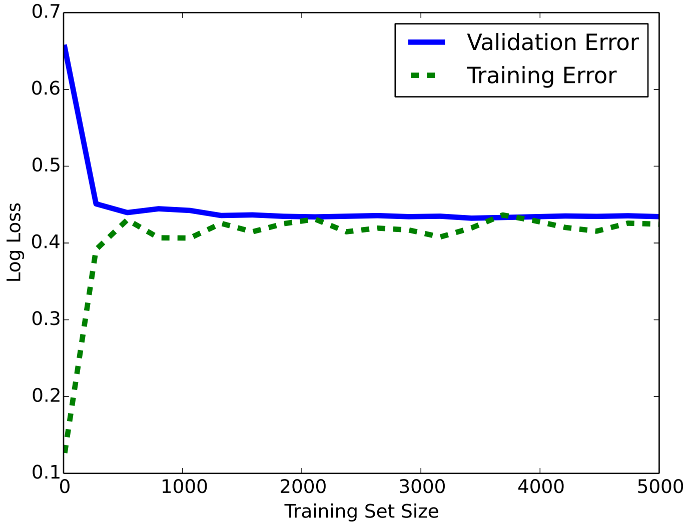


Figure 4.7: Learning curve for logistic regression with log loss metric. The training error and validation error converges by the time the dataset reaches the size of 1000 author pairs. Therefore the training set size (5361 positive instances and 5361 negative instances) is sufficient for the task

the training set size (5361 positive instances and 5361 negative instances) is sufficient for the task.

Figure Learning curve for logistic regression with log loss metric.

We also manually analyzed the prediction errors. The data sparseness is the one of the most important reasons for false negatives. Specifically, if the author has few publications and few co-authors in the past, there is little information we can derive for features such as research interest, network topology, and activity level. The network that we used in this study is a sub-graph of MEDLINE publications only and therefore provides an incomplete picture of the publication history of certain authors. For example, Flaumenhaft R (author of PMID 12837380) has only one publication prior to 2009 with only one co-author. His/her pairing with Laurence RG (author of PMID 18715793) has a *simText* value of 0.010 and a *simOutcite* value of 0.143 (the average value for each feature in the positive training data was 0.134 and 0.325, respectively), although Laurence RG is more prolific in our network with 12 publications and 35 co-authors. In fact, by searching PubMed we found that Flaumenhaft R has been publishing almost every year from 2003 to present and has many common co-authors with Laurence RG; this information was missing entirely in our network, which was built using the MEDLINE data only. As a result, our models predicted Flaumenhaft R was unlikely to collaborate with Laurence RG, which is therefore a false negative. On the other hand, false positive errors can arise due to the fact that these author pairs have features very much like those in the positive training data. These authors, however, might never have had a chance to actually know each other, leading to a false positive.

Furthermore, author name disambiguation contributes errors for both false positives and false negatives. Data sparseness arises when one author is mapped to two unique IDs by the author name disambiguation database we used. For example Guida M (author of PMID 17113552) has two IDs. There are only five publications assigned to the ID that we happened to use in our network, while there are 94 publications under the other ID. We are also aware that it is possible for an author to share the same ID with another, unrelated author; this can also cause a disambiguation error and the information from the unrelated author will be wrongly attributed to the original author. However, we did not actually find any such cases in our test sets.

Recall we have built two different testing data sets, and our analyses of the true positive, false positive, true negative and false negative of the three testing data sets show interesting results (Table 4.8). In the *RandomPairCategory* dataset (i.e, positive and negative author pair data were randomly selected) our classification has low sensitivity (0.718) and high specificity (0.859) while the *IndividualAuthorCategory* yielded the opposite (0.900 sensitivity and 0.321 specificity). The high specificity of *RandomPairCategory* evaluation is due to the fact that the negative instances are very negative as they were constructed by the random combination of two authors; therefore, they tend to share few research interests and even fewer common friends.

In contrast, the negative instances of *IndividualAuthorCategory* testing set were from the sub-graph of the author, so they do have similar research interests and have a higher chance of sharing a common collaborator. The sensitivity advantage of *IndividualAuthorCategory* can be understood in a similar way, as the positive instances, which were sampled from the sub-graph of the author, are ÂIJvery positiveÂİ and share research interests and common collaborators, which increases the likelihood of the classifiers to classify them as positive.

4.5.3 Feature Analysis

Table 4.7 shows that the most important feature for collaboration prediction, as measured by information gain, is the similarity of out-citing citations, which represents an author's knowledge background. True positive instances tend to have a larger *simOutcite* value than negative instances (mean simOutcite is 0.305 for *RandomPairCategory* and 0.462 *IndividualAuthorCategory* positive instances while it is 0.159 and 0.264 for negative instances in the two categories respectively), suggesting that common background knowledge increases the chance for collaboration. As for the feature *simOutcite*, collaborating pairs have a higher *simText* score than non-collaborating pairs do. An author'ÂŹs publication history represents the author's research area and *simText* shows the similarity of two researchers' fields. As for the feature *simOutcite*, collaborating pairs have a higher *simText* score than non-collaborating pairs do. Our results also show that research field overlap is positively related to potential collaborations.

MeSH terms can be considered the topics of a biomedical article, with the feature *simMeSH* a measure of research interest similarity. Therefore it is not surprising that *simMeSH* contributes to the classification. Keyword overlap was explored in a previous study[119] and was a top-ranking feature. In contrast to that study and others [119][120] that did not explore text as features, we found that the feature *simMeSH* ranks below *simText* in information gain. We speculate that although MeSH terms represent an article'ÂŹs semantic content, they are not as robust as the bag of words formulation of *simText* for the task of author collaboration classification, because MeSH terms may not be considered as fine grained as word features in the abstract.

Our results also show that neighborhood structure plays an important role in predicting collaboration. The features *numCommonCoauthor*, *coauthorJaccard*, and *Adamic* all have large information gain. Positive instances tend to have larger number of common co-authors than the negative instances, as for the *IndividualAuthorCategory* testing dataset, where researchers are topologically close (mean *numCommonCoauthor* is 1.0) but negative pairs still tend not to have common collaborators (the mean is closed to 0). Our results suggest that the strength of social ties is important for establishing collaboration. This conclusion is consistent with our hypothesis and previous findings, which show that common neighbors are a very effective predictor in social networks [66].

Features that are related to researcher productivity, such as *sumCoauthor*, *sumRecency*, and *sumPub*, are ranked lower than *simOutcite*, *simText*, *coauthorJaccard*, *Adamic*, *numCommonCoauthor*, *simMeSH*, and *simIncite*, as measured by information gain, suggesting that two researchers' specific activities do not have to be closely related to establish a new collaboration. In contrast, the sum of co-authors was found to be the most important feature in work[119], but it is not clear if this was influenced by the normalization by year, as carried out in our study. Consistent with previous findings, the clustering coefficient, which describes the transitivity of a collaboration, is not an effective feature [119]. It is also interesting to note that difference in seniority between collaborators, described by *diffSeniority*, has no impact on establishing a new collaboration in our approach.

Furthermore, we trained classifiers using every single feature individually and analyzed the performance as shown in Figure 4.5.3. Research interest features *simText*, *simMesh* and *simOutcite* (Figure 4.5.3 panels a, b and d) have large ROC areas, showing that they are informative for the classification. *simIncite* (Figure 4.5.3c) however is not as large as other features in this category with 0.56 ROC only. 2) Common co-author based features (Figure 4.5.3 panels f, k and l) exhibit distinct patterns and are essentially equivalent features as they have large correlation coefficients among each other. For example *Adamic* and *numCommonCoauthor* have a correlation coefficient of 0.96. This suggests that we can use *numCommonCoauthor* as a feature and remove *Adamic*. The ROC curve for them is a straight line due to the fact that most of the author pairs (especially negative instances) don't have any common coauthors, and only 1/3 of the positive instances have non-zero common co-authors. 3) Other features such as *sumCoauthor* (Figure 4.5.3f) is also effective for classification with 0.67 ROC. Activity feature *sumRecency* (Figure 4.5.3i) has 0.60 ROC, and so does *sumPub* (Figure 4.5.3e). *sumClusteringCoef* and *diffSeniority* (Figure 4.5.3 panels h, j) show only 0.53 and 0.54 ROC respectively The individual ROC is consistent with information gain analysis, which also shows that the research interest features are most informative, followed by common neighbor based features.

In summary, previous work in author collaboration prediction mainly explored topological features. Our results, in contrast, show that research interest is important for establishing a new collaboration. Specifically, research profile similarity features such as *simOutcite* and *simText*, as shown in Table 4.7, are the most important features for the classification. Tables 4.4 and 4.5 show that the supervised machine learning models that incorporate research similarity features significantly outperformed the baseline systems, which were built upon widely used topological features (*PreferentialAttachment*, *JaccardBaseline*). We speculate that knowing the other's work is a form of shared experience and the foundation of trust between two researchers. Their common knowledge, represented by the research similarity features, plays an important role for building collaboration.

As discussed earlier, although seniority plays a limited role in a collaboration, Figure shows that when researchers are in their early career stages, they are more likely to collaborate with those whose research interests differ from theirs, suggesting that junior faculty are more open to interdisciplinary collaborations. In contrast, collaborations between two senior researchers exhibit a higher degree of research interest similarity, suggesting that established researchers are more comfortable in their own fields and are less likely to initiate an interdisciplinary collaboration.

4.5.4 Limitations

There are several limitations to this study. First, we did not explore learning features of broad social factors, including institutional policies like the status of an IRB application or institution-specific restrictions, because it is difficult to generate these data computationally. Second, our data were incomplete and contained missing information. We used a sub-graph of the MEDLINE co-author network and therefore the author publication histories may not be complete, as we described in the error analysis. The missing publications could represent different research interests. It also takes time for an article to accumulate citations, so simIncite may be biased to have more citations for older works than for recent ones. Third, our training and testing period time cutoff is arbitrary, and we define a negative instance pair as authors who did not collaborate by the time of the training or testing period. However it is possible that they collaborated later. Finally, We are aware that in the real world, the network density is as very low (7.6e-06 in CiteGraph as computed in Section 2.5.2.1). Yet in our study we used equal number of positive and negative examples for training and testing as our major purpose is to demonstrate the performance of our rich feature set. The evaluation on real world environment requires additional experiment design such as including heuristics to prune negative instances.

4.5.5 Real World Application

There are several steps to integrate our application to real world application. The idea is similar to *IndividualAuthorCategory* evaluation, which samples candidate collaborators from the author's sub-network instead of whole author space. The intuition to eliminate the extremely unlikely collaborators who are very distant in the co-authorship network. The details are shown as following.

1 For author s, obtain the subnetwork G' by breadth–first traversal of k layers in the co-authorship network G starting from s.

2 For those researchers in the G' never collaborated with the author s, predict the probability of future collaboration with s.

3 Rank those authors by their probability and the ones that are above certain threshold are considered likely future collaborators.

Our system fully supports the above procedure as we used it for *IndividualAuthorCategory* candidate collaborator sampling. The feature extraction can be easily scaled-up by using parallel computing framework Hadoop [76] as each author pair's features extraction are completely independent from other author pairs.

In social network applications users are often interested to know the shortest distance between them and their potential collaborators in the co-authorship network as it suggests the way to reach (or get introduced) this researcher through one's network. The centrality of a particular researcher, which is often viewed as an indication of authority in the community, is also valuable information. Our co-authorship network database supports the calculation of these metrics as the database has stored the co-authorship relations in CiteGraph articles, and the shortest path and centrality algorithms are well studied and easy to implement.

4.6 Conclusion

In this study we developed supervised machine learning approaches for automatic research collaboration prediction. We derived novel features to model research interest while incorporating various co-authorship network topology characteristics. We demonstrated the efficacy of our models for research collaboration prediction. Moreover, we found logistic regression to be the best model for this approach, with the highest ROC to be 0.980. We also identified the key factors for establishing collaboration: first, individuals tend to collaborate with those who share certain degrees of research interest; second, common collaborators (common friends) are also important; and third, researcher activities, such as the average number of collaborations and the average number of publications, are also influential but less so than degree of research interest overlap and the number of common collaborators. Our approach takes advantage of the big-data literature resources but is efficient to implement. It has thus demonstrated its effectiveness as a research collaboration recommendation application. The datasets of this study can be downloaded from https://github.com/qingzhanggithub/medline-collaboration-datasets.

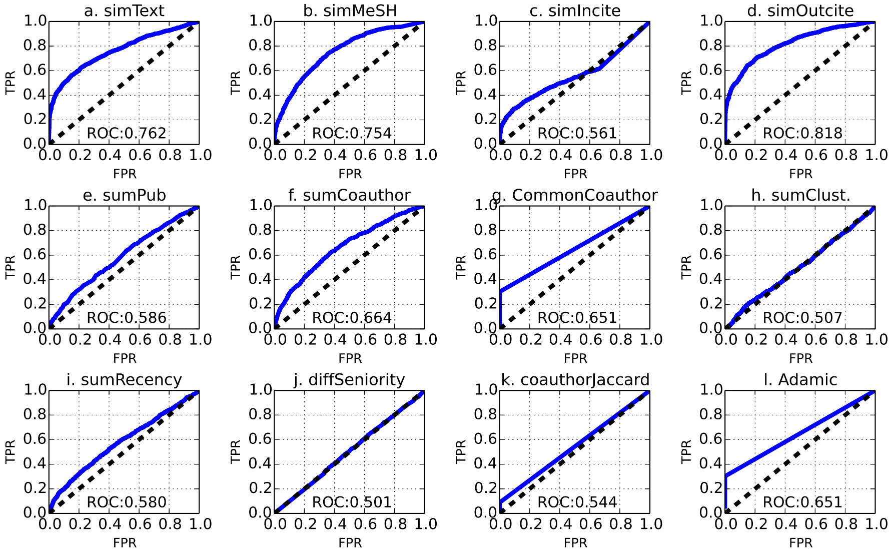


Figure 4.8: ROC for classifiers trained by single feature. 1) Research interest features simText, simMesh and simOutcite (panels a, b and d) have large ROC areas, showing that they are informative for the classification. simIncite (panel c) however is not as large as other features in this category with 0.56 ROC only. 2) The ROC curves for common co-author based features (panels f, k and l) are a straight lines due to the fact that most of the author pairs (especially negative instances) don't have any common coauthors, and only 1/3 of the positive instances have non-zero common co-authors. 3) Other features such as sumCoauthor (panel f) is also effective for classification with 0.67 ROC. Activity feature sumRecency (panel i) has 0.60 ROC, and so does sumPub (panel e). sumClusteringCoef and diffSeniority (panels h, j) show only 0.53 and 0.54 ROC respectively The individual ROC is consistent with information gain analysis, which also shows that the research interest features are most informative, followed by common neighbor based features.

Chapter 5

Recurrent Collaboration Prediction

5.1 Introduction

Research collaboration is distinct from many of other social networks. When represented as a time series, a particular co-author collaboration is seen as a transient, discrete event at a given point in time; as such, two researchers may choose to collaborate multiple times and each of these collaborations would be represented separately in time. This characteristic of a time series is distinctive from link establishment in many other social networks, where the users establish a connection once and it persists in time unless this connection is no longer desirable. Efforts have been made to model such connections as a time series[149, 150, 151]; however, the predictors used in these previous works are mainly derived from network topology and did not utilize other useful information, such as the publication history and the number of citations, which could be helpful features. To our knowledge, existing collaboration prediction methods have also not explored geographical information as a predictor even though the location of the researcher is often an important factor in collaborations[132]. In addition, no previous work has looked at recurring collaboration prediction, defined as the probability of two researchers who have collaborated once to collaborate again.

We are therefore motivated to study recurrent collaboration prediction. Since joint publication is one of the most effective representations of collaboration, we formulate recurrent research collaboration prediction as a link prediction problem in the co-authorship network where co-authorship indicates a collaborative relation. Our goal is to predict whether a node pair (two researchers) with an existing link (collaboration) will connect again. As illustrated in Figure 5.1, the collaborations between researchers A, B, ... , F are shown with solid links. The number on the edge indicates the collaboration frequency. Our goal is to use supervised machine learning models to predict, given the history of the collaboration network, whether two researchers (e.g., A and B) will collaborate again. In this study, we collected 112,733 author pairs from MEDLINE publications. We derived novel features from five different categories (Research Interest Profile, Publication Productivity, Author Seniority, Network Connectivity, and Geographical Location) and built supervised models using this data. We analyzed the factors for recurring collaborations and found characteristics that were very different from one-time only collaborations.

Our main contributions are: deriving novel, comprehensive features to model aspects of recurrent collaborations; shedding light on understanding long-term collaborations with multiple collaborations; and developing features and models that are computationally efficient and hold the promise of being applied to real-world research collaboration recommendation applications.

5.2 Background

Here's a revisit of the background of co-authorship network and research collaboration analysis. Factors that lead to successful collaborations have also been studied, including various social and environmental factors such as leadership, geographical proximity, and the personalities of the team members. For example, one study con-

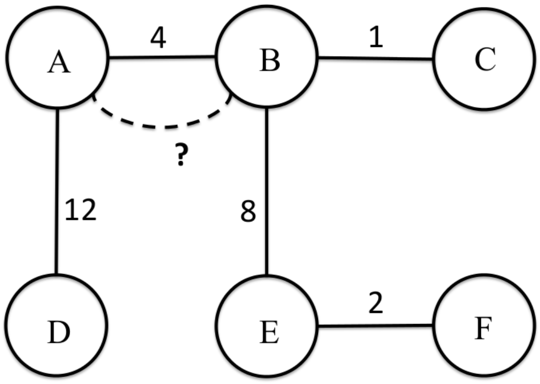


Figure 5.1: An illusion of re-collaboration prediction. A solid line shows existing collaborations and a dashed line represents the link that we would like to predict. The number on the edge shows the collaboration frequency.

cluded that a leader in a research field typically plays an important "broker" role to bridge people from different disciplines[111]. Physical proximity between the first and last author was found to be positively related to the impact of collaboration as measured by the number of citations the paper received; this study was based on the data collected from Harvard Medical School publications and the office locations of the researchers[132]. They argued that close geographical distance between the first and last author is important for the outcome of the collaboration. International collaborations, however, are also a positive factor in determining the impact of the publication. As found in [133], the average number of citations for a paper increases with the number of countries affiliated with the authors of that paper. Certain characteristics of the team members, such as openness and flexibility, also contribute to the success of the collaboration[111, 134].

Research collaboration recommendation is often cast as a link prediction problem in social networks. Liben-Nowell and Kleinberg [66] comprehensively evaluated a collection of topological predictors for co-authorship prediction in Physics; this work has largely served as one of the foundations for many later studies. In addition, link prediction can be formulated as a probabilistic distribution problem (e.g., random walk on the co-authorship network [66]) or a binary classification problem (e.g., classifying an author pair as collaborating or non-collaborating) [119, 122] or the blending of both [120], where the features (or predictors) are extracted based on the node attributes (e.g., overlap of keywords) as well as distance-based measures (e.g., shortest distance between two nodes) [119, 122, 120].

5.3 Methods

5.3.1 Features

In this chapter we extend the feature set that we have developed in Chapter 4.

5.3.1.1 Geographical Location

Furthermore, we extracted features based on an author's location, defined by their affiliation, history. For each publication in the feature extraction period, we calculated the number of co-authors that were at a different institution from author x. We then calculated the percentage of outside institution collaborators in the author's entire publication history during that period, which we defined as an author's openness,

$$openness(x) = \frac{|inter - institution\, collaborators|}{|collaborators|}$$

Subsequently, we defined the feature *sumOpenness* by summing the openness score of two authors. We also define a binary feature *sameLocation*, which is computed by comparing the affiliation addresses of the two authors from their initial collaboration. Specifically,

$$sameLocation = I(location(x) = location(y))$$

where $location(.)$ is the initial collaboration location. Given the variability of affiliation formats, the most consistent granularity was at the level of state/province; thus,

Table 5.1: Recurrent collaboration feature definition. It is an extension of the feature set in Chapter 4.

Category	Feature	Definition
Research Profile Similarity	simText	$\frac{absract(x).abstract(y)}{\lvert abstract(x)\rvert\lvert abstract(y)\rvert}$
	simMeSH*	$\frac{\lvert MeSH(x)\cap MeSH(y)\rvert}{\lvert MeSH(x)\cup MeSH(y)\rvert}$
	simIncite	$\frac{incite(x).incite(y)}{\lvert incite(x)\rvert\lvert incite(y)\rvert}$
	simOutcite	$\frac{outcite(x).outcite(y)}{\lvert outcite(x)\rvert\lvert outcite(y)\rvert}$
Publication Productivity	sumPub	$avgPub(x) + avgPub(y)$
	sumRecency	$recency(x) + recency(y)$
Seniority	diffSeniority	$\lvert seniority(x) - seniority(y)\rvert$
Connectivity	sumCoauthor	$avgCoauthor(x) + avgCoauthor(y)$
	numCommonCoauthor*	$\lvert\Gamma(x)\cap\Gamma(y)\rvert$
	sumClusteringCoef	$clusteringCoef(x) + clusteringCoef(y)$
	coauthorJaccard	$\frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\lvert\Gamma(x)\cup\Gamma(y)\rvert}$
	Adamic	$\sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{log\lvert\Gamma(z)\rvert}$
	PreferencialAttachment*	$\lvert\Gamma(x)\rvert\lvert\Gamma(y)\rvert$
	sumOpenness	$avgOpenness(x) + avgOpenness(y)$
	sameLocation	$I(location(x) = location(y))$

*Baseline feature (please see Section 5.3.3 Baseline Model Comparison for more information).

if two authors are from the same state/province, the *sameLocation* feature value is set to 1; otherwise, it is set to be 0. In order to get the organization and location information we parsed the address of the author using the named entity recognition models (rganization model and location model) provided by Apache OpenNLP[152].

5.3.2 Supervised Machine Learning Models

We formulate the recurrent collaboration prediction problem as a machine classification task. We explore four supervised machine-learning models: Naive Bayes (NB), Support Vector Machines (SVMs), Logistic Regression (LR) and Random Forest (RF), all of which are commonly used for classification tasks [141]. A naive Bayes classifier is a probabilistic classifier based on Bayes theorem with the independence (naive) assumption that the features are generated independently from each other, given the instance label[137]. SVMs are based on the concept of maximum margin decision planes that define generalizable decision boundaries for classification and regression. An SVM constructs a hyperplane to maximize the margin between the data points and the hyperplane, often after mapping the data points to a higher-dimensional space in which they are linearly separable or close to it [140]. In particular, we use an RBF kernel for its ability to model a nonlinear decision boundary for the data. Logistic Regression estimates discrete or continuous value parameters in order to predict discrete category values. The probabilities that describe the possible class of a single instance are trained as a function of explanatory variables within a logistic function [137]. The Random Forest [147] algorithm is a bagging approach applied on a collection of independent decision trees and has been widely used for classification [153, 154, 155]. These four classifiers are not only well-studied models in a variety of classification tasks[141] but are also widely available in open source software communities.

We use the python machine learning package, Scikit-learn[145], for model training and testing. It has emerged as one of the most widely used machine learning libraries in Python and includes commonly used algorithms for supervised and unsupervised machine learning as well as regression. Note that it uses the popular SVM package LIBSVM [144] as its underlying SVM implementation. Scikit-learn is under rapid development and is extensively documented.

5.3.3 Baseline Model Comparison

We use *simMeSH*, *numCommonCoauthor* and *PreferencialAttachment* as baseline features as previous studies have shown they are effective for link predictions. Keyword overlap has been found as top features for collaboration prediction[119] and *simMeSH* as mentioned before is serving a similar purpose. Number of common coauthor is strong predictor for link establishment in co-authorship networks[51, 66].

As described in Chapter 4, preferential attachment is a well-studied network growth pattern. Concretely, the more existing links a node has, the higher the chance a new node will link to it. We used the same supervised machine learning models (LR, NB, SVM, and RF) on the individual and combined baseline features *simMeSH*, *numCommonCoauthor* and *PreferentialAttachment*.

5.3.4 Data

In this study we also use CiteGraph as our data source. The details of the CiteGraph database can be found in Section 4.3.4. The description of how we created CiteGraph is in Chapter 2.

5.3.5 Training Dataset

We used the CiteGraph Dataset for both the training and testing data. The positive training instances are author pairs who have collaborated more than one time. In particular, their first time collaboration has to occur between the years 2000 and 2004. This time-window based sampling method follows the standard approach for link prediction problems such as [66, 119, 150]. We further filter the authors who have at least 3 publications in our database to overcome the feature sparseness. In total we get 432,855 valid positive instances (30.9% of multiple-time collaboration author pairs in our dataset) and we randomly sampled 20,000 author pairs from these for the positive instances in our training set. From a total of 740,588 valid negative candidates (12.8% of one-time collaboration author pairs in our database) we randomly selected 20,000 author pairs for the negative instances in our training set. The negative training instances are author pairs who collaborated only once and that collaboration occurred between years 2000 and 2004. We created a separate testing set also by sampling from the 432,855 valid positive and 740,588 valid negative instances. In total, we collected 72,773 testing instances with 35,931 (49.4%) of them being positive instances above. The illustration of sampling positive and negative examples are shown in Figure 5.2.

We decomposed our training and testing set to understand the details of dataset. As shown in Figure 5.3A, the Collaboration Frequency pie chart (left panel) shows the percentage of author pairs that have certain collaboration frequencies. Two-time collaboration pairs occupy 41.44% of the dataset, and the percentage decreases as the number of collaborations increase, which means high frequency collaborations (i.e., $>$ 5 collaborative publications for that co-author pair between 2000 - 2009) are rare. The right panel shows the percentage of author pairs that have a specific collaboration year span. 21.02% of the collaborations have a two-year span, and the percentage of instances with a certain year span decreases as the year span itself gets larger. The Pearson correlation coefficient for both collaboration frequency distribution and collaboration year span comparing with overall data are over 0.99, ensuring that our sampled 20,000 positive instances are representative. In the testing set data overview in Figure 5.3B, the collaboration frequency also has a skewed distribution with 40.98% two-time collaborations and 27.19% three-time collaborations. On the other hand, the percentage of collaborations with a certain year span monotonically decreases as the time span itself increases.

5.3.6 Evaluation Metrics

We calculate precision (TP/(TP + FP)), recall (TP/(TP + FN)), and ROC AUC, the area under the curve of the true positive rate (TPR) over the false positive rate (FPR) - and accuracy ((TP+TN)/ALL), where TP, FP, TN, FN and ALL stand for number of true positives, false positives, true negatives, false negatives, and number of total instances respectively. The F1 score is defined as the harmonic mean of recall and precision, specifically 2*recall*precision/(recall+precision).

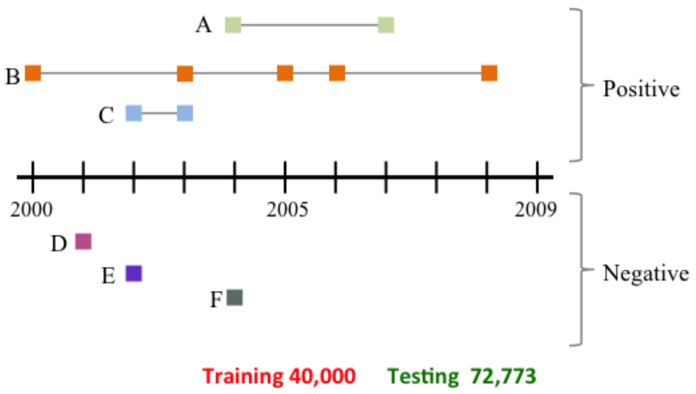


Figure 5.2: Sampling positive / negative examples. Every row represents an author pair, and every dot denotes a collaboration event. Positive examples are defined as whose who initally collaborated between 2000 and 2004 and later recollaborated. On the other hand the negative examples are defined as the author pairs who collaborated only once and it happened between 2000 and 2004.

5.4 Results

We first performed five-fold cross validation on the training set with both the full feature set and only the baseline features from the Baseline Model Comparison (Section 5.3.3). As shown in Table 5.2, the random forest model on all features (RF: All features) yielded the best AUC (0.720) and F1 (0.648) among all the models. The best baseline AUC was seen for the SVM model (with RBF kernel) trained on all three baseline features (SVM: All baseline features), yielding 0.613 AUC, while the best baseline F1 resulted from the logistic regression model trained by the PreferentialAttachment feature only (LR:Pref.Attach.). Subsequently, we evaluated our systems on the testing set. As shown in Table 5.3, the results are consistent with the cross validation. The random forest model trained by all features (RF: All features) outperformed the others with 0.732 AUC and 0.653 F1. Amongst the baseline models, the SVM trained by all the baseline features (SVM: All baseline features) showed the best AUC (0.617) and logistic regression with the single *PreferentialAttachment* feature had the best baseline F1 (0.633). A list of top ten true positive predictions (according to the probability) is shown in Table 5.4.

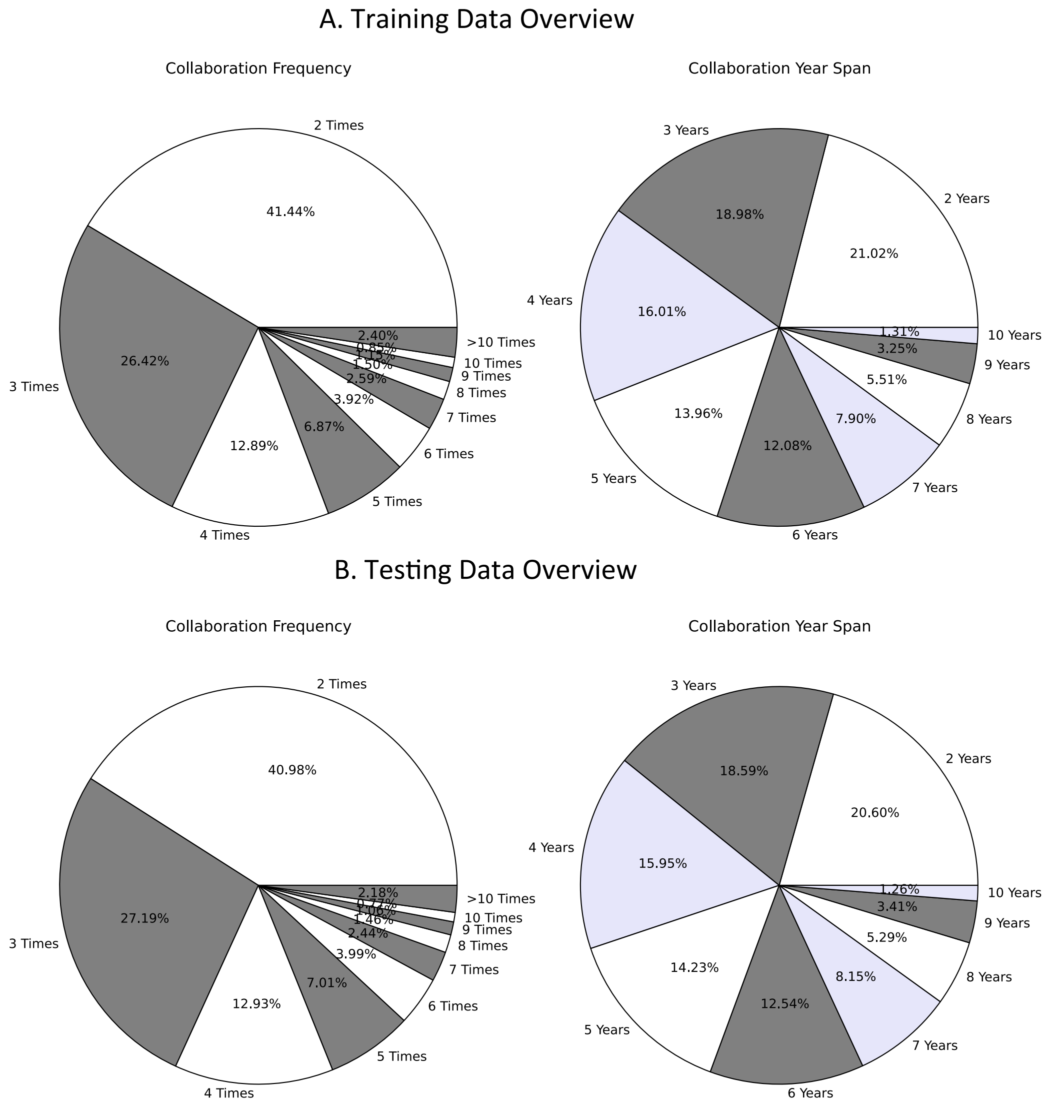


Figure 5.3: Overview of the training (A) and testing (B) sets. A) The Collaboration Frequency pie chart shows the percentage of author pairs that have specific collaboration frequencies. 2-time collaboration pairs (author pairs that have published together twice only) occupy 41.44% of the training set and the percentage decreases along with increasing number of collaborations, which means high frequency collaborations are rare. The right panel shows the percentage of author pairs that have a specific collaboration year span. 21.02% of the collaborations have a two year span (i.e., all their collaborations occur over two years), and the number of collaborations that occurred over a certain time span decreases as the time span gets larger. B) The testing set distribution is consistent with the training set distribution: 40.98% of the author pairs are 2-time collaborations and the percentage continues to decrease when the collaboration frequency gets larger. 20.60% of collaborations have a two year span, and the proportion shrinks as the collaborations occur over a longer time span.

Table 5.2: Recurrent collaboration prediction results for the five-fold cross validation on the training set. The bold-faced values are the highest values in the column.

Task	Models	ROC AUC	Precision	Recall	F1
Our system	LR:All features	0.710	0.656	0.639	**0.618**
	SVM: All features	0.713	**0.684**	0.604	0.642
	RF: All features	**0.720**	0.677	0.622	**0.648**
	NB: All features	0.671	0.657	0.503	0.570
Individual Baseline Features	LR: Pref. Attach.	0.596	0.533	**0.794**	0.638
	LR: MeSH	0.611	0.619	0.394	0.481
	LR: Common Coauthor	0.529	0.510	0.673	0.580
Combined Baseline Featrures	LR: All baseline features	0.612	0.610	0.443	0.513
	SVM: All baseline features	0.613	0.600	0.509	0.550
	RF: All baseline features	0.565	0.549	0.536	0.541
	NB: All baseline features	0.588	0.577	0.523	0.537

We performed a grid search using Scikit-learn in order to tune all hyper-parameters such as the number of estimators of the random forest model and the penalty of misclassification for SVMs (the coefficient C, kernels). Therefore, the results reported here are all for the models with the best performing hyper-parameters.

To further study the performance of our model, we broke down the testing set by collaboration frequency (i.e., the number of collaborations between two authors). We divided the testing set based on the collaboration frequencies, ranging from 2 joint publications from 2000 - 2009 (called 2-time collaborations) to 10 joint publications from 2000 - 2009 (called 10-time collaborations). For each of these collaboration frequencies, we tested each subset with the logistic regression model trained above as it yielded the best F1 score and is computationally efficient. As shown in Figure 5.4, the 2-time collaboration testing subset (i.e., two authors publishing together twice between 2000 and 2009) had 0.627 AUC and 0.556 F1, and the performance increased

Table 5.3: Recurrent collaboration prediction results on the testing set. The boldfaced values are the highest values in the column.

Task	Models	ROC AUC	Precision	Recall	F1
Our System	LR: All features	0.712	0.654	0.644	0.649
	SVM: All features	0.714	**0.678**	0.609	0.642
	RF: All features	**0.732**	**0.678**	0.629	0.653
	NB: All features	0.673	0.653	0.505	0.570
Individual Baseline	LR: Pref. Attach.	0.600	0.527	**0.791**	0.633
	LR: MeSH	0.612	0.616	0.396	0.482
	LR: Common Coauthor	0.529	0.503	0.670	0.575
Combined Baseline Features	LR: All baseline features	0.614	0.607	0.445	0.513
	SVM: All baseline features	0.617	0.597	0.511	0.551
	RF: All baseline features	0.583	0.554	0.541	0.547
	NB: All baseline features	0.587	0.559	0.533	0.545

Table 5.4: List of true positive predictions. First Collab. lists the publication (PubMed ID) of the first time collaboration, and Second Collab. is the publication of the second time collaboration.

Author A	Author B	First Collab.	Second Collab.	Probability
Elliott Antman	Robert Califf	15063425	16084152	0.994
Zhong-Wen, Lin	Han Dong sun	10925014	11077166	0.990
Anne Marie Api	Charlene Letizia	12804650	12804651	0.990
Linda Boyken	Ronald Jones	14972378	14972378	0.988
Seong Ho Choi	Kwang Woong Lee	15561200	15561210	0.986
Robert Califf	Matthew Roe	11153779	11230834	0.982
David Apple	Luis Vargas	11821220	12106732	0.979
Robert Califf	Monica Shah	11275915	11376303	0.978
Kyoichi Saito	Kazuyuki Sugita	10949471	11519816	0.978
Hwan Hyo Lee	J-H Park	15561216	15561222	0.973

to 0.752 and 0.693 for the 3-time collaboration subset. Both AUC and F1 generally increase as the collaboration frequency gets larger, yielding 0.824 and 0.755 for 9-time collaboration as the best-performing subset.

5.4.1 Feature Analysis

To better understand the contribution of each of the features, we divided the training set by the collaboration frequency, as well. We calculated the 10-fold cross validation AUC (Figure 5.5A) and information gain (Figure 5.5B) for each feature in each collaboration frequency training data subset. In addition, as we see in Figures 5.4 and 5.5, the 2-time collaboration is a distinct outlier from rest of the collaboration frequency subsets. Therefore, we separate the collaboration frequency into three sections: 2-time collaborations, 3-5 time collaborations (low-frequency collaboration), and 6-10 time collaborations (high-frequency collaboration). In Figure 5.6, we calculated the average AUC for low-frequency collaborations, high-frequency collaborations, and the combined 3-10-time collaborations for each feature (from Figure 5.5A) in order to summarize the contribution of each of feature and highlight the salient characteristics of more frequent recurrent collaborations.

Research Interest Profile features (*simText* and *simOutcite*) showed a strong contribution across different collaboration frequencies as seen in both Figure 5.5A and 5.5B. The two demonstrated an average AUC of 0.662 and 0.642, respectively, on the Overall 3-10 time collaboration subset as seen in Figure 5.6. Geographical Location features are also at the top as shown in Figure 5.5A (they were excluded from Figure 5.5B as the high information gain skewed the data and made the heat map almost white for the rest of the features. The information gain of *sameLocation* ranges from 0.015 to 0.115 and for *sumOpenness* it is from 0.000 to 0.646). The Network Connectivity feature *sumCluteringCoef* also showed an advantage over all the collaboration frequencies. In addition, *sumCoauthor* (average AUC for low-/high-frequency was

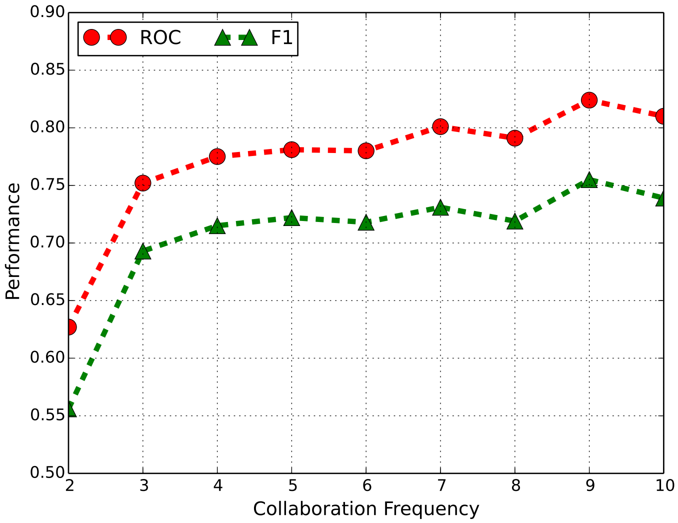


Figure 5.4: Recurrent collaboration prediction testing result on subsets with a particular collaboration frequency. Collaboration frequency is defined as the number of collaborations between two authors between 2000 and 2009. 2-time collaborations had 0.627 AUC and 0.556 F1, and the performance increased to 0.752 and 0.693 for the 3-time collaboration subset. Both AUC and F1 generally increase as the collaboration frequency gets larger, yielding 0.824 and 0.755 for 9-time collaborations as the best-performing subset.

0.523/0.576) and *Adamic* (average AUC for low-/high-frequency was 0.522/0.582) also show a relative advantage on high-frequency collaboration. *PreferentialAttachment*, on the other hand, has a better performance on low frequency collaborations (average AUC 0.599) than on high frequency ones (average AUC 0.506). Productivity feature *sumPub* has better performance on high-frequency collaborations (average AUC 0.592) than low-frequency ones (average AUC 0.649).

Average co-authors per year (*sumCoauthor*), the author pair seniority difference (*diffSeniority*) and their recency (*sumRecency*) in general did not show strong performance, with average AUC for 3-10 time collaboration 0.543, 0.556 and 0.547 respectively.

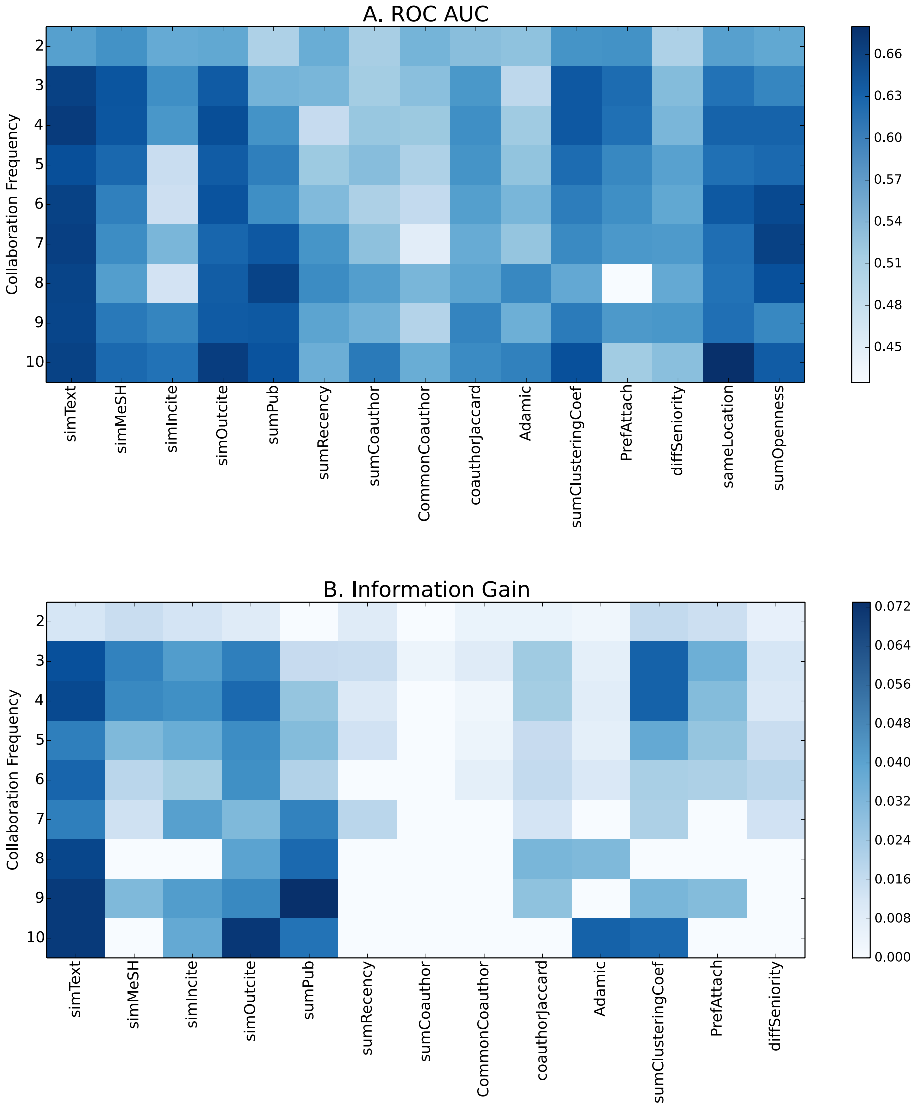


Figure 5.5: Recurrent collaboration prediction feature analysis with AUC and information gain. A) AUC of each feature across different collaboration frequencies (darker colors are better). B) Information gain for each feature across different collaboration frequencies (darker colors are better). The Geographical Location features, sameLocation and sumOpenness, are excluded from the heat map as their high values skewed the color map and overshadowed the color distribution of the other features (The information gain of sameLocation ranges from 0.015 to 0.115 and for sumOpenness it is from 0.000 to 0.646). As can be seen from the 2-time collaboration frequency row, all features have very low information gain for the 2-time collaboration frequency.

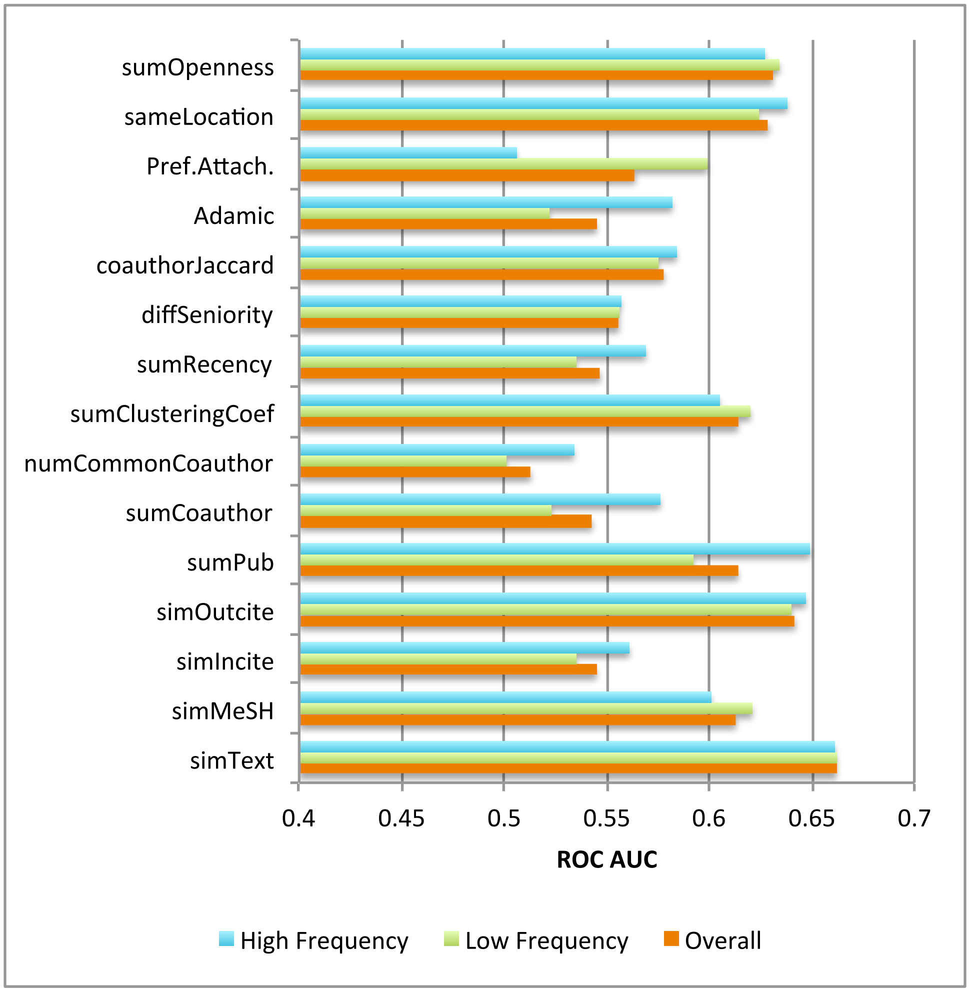


Figure 5.6: Average AUC for Low Frequency (3-5-time collaborations), High Frequency (over 5-time collaborations), and Overall (3-10-time collaborations) data. 2-time collaborations were excluded because of the large number of 2-time collaborations, as seen in Figure 4, and the performance difference between the 2-time and 3-10-time collaborations, as seen in Figure 5, would skew the results; here, we want to highlight the salient characteristics of more frequent recurrent collaborations only.

5.5 Discussion

5.5.1 Features

The rich feature set we introduced in this paper provides useful information for recurrent collaboration prediction. As shown in Table 5.2 and 5.3, the models trained by our full feature set show a significant performance advantage over the models trained by just the baseline features: the best baseline supervised machine learning model is the SVM trained by all three baseline features with an AUC of 0.617 while the random forest model trained by our full feature set yielded 0.732 AUC. The performance advantage of our full feature set over all the baseline permutations suggests that our feature set is informative for predicting recurrent collaboration.

In particular, Research Interest Profile features were seen to be strong predictors of recurrent collaboration. Specifically, *simText* and *simOutcite*, which indicate publication history similarity and knowledge background similarity respectively, are top features, as also observed in the first-time collaboration prediction task in Chapter 4 (Section 4.4.5). Positive instances (co-author pairs who have published together more than twice) in the training set tend to have larger research interest similarity, as shown by their average *simText* and *simOutcite* AUC values of 0.508 and 0.647; negative instances (co-author pairs who have not published together at least twice), on the other hand, have average AUC values of only 0.398 and 0.568. This suggests that knowing each other's work is important for long-term collaborations. However, the similarity of in-citing citations (*simIncite*) is not as informative as the rest of the Research Interest Profile features for three possible reasons: 1) in-citing citations to a paper indicate how much impact a paper has on the research community but is not directly related to the ability of the authors to continue to collaborate with each other specifically; and 2) it takes time to accumulate citations so there is a temporal bias towards older papers (i.e., older work is more likely to be cited more extensively thus having more in-citing documents and allowing for a richer profile calculation).

Our study found that Geographical Location is an important predictor for recurrent collaboration, as well. The average value of the *sameLocation* feature for positive instances in the training data is 0.784 (SD 0.412), while it is 0.585 (SD 0.493) for the negative instances. This suggests that co-location is positively related to recurring collaborations. On the other hand, the positive instances on average have smaller *sumOpenness* scores than the negative instances (0.364 and 0.440 respectively), which suggests that people who collaborate extensively with others in the same institution tend to re-collaborate and is consistent with the trend suggested by the *sameLocation* feature. As confirmed by other studies [156], geographical proximity helps to reduce communication costs and allows collaborators to have face-to-face meetings more easily. The colocation-collaboration research by Harvard Medical School [132] also showed that physical proximity of the first and last author was positively related to the impact of the paper.

Interestingly, among the Network Connectivity features, we found the number of common co-authors did not contribute as much as the other features, as shown in Figures 5.5 and 5.6, which is different from what we found for first-time collaborators in Chapter 4. This could be due to the fact that two authors who do not know each other might rely on a mutual collaborator to introduce them and facilitate their collaboration; however, after two researchers have already collaborated once, the number of mutual collaborators they had in the past is no longer a strong influence on how well they conduct future collaborations as have already established their initial connection. Instead, the high performance of the *sumCluteringCoef* feature shows that the sub-networks of each author in a co-author pair does have an influence. This might be due to each co-author's sub-network (i.e., the collaborators of each co-author who are not mutual collaborators with the other co-author) representing complementary skill-sets with the co-authors then serving as intermediaries who bridge the differing expertise of the collaborators in the two sub-networks.

PreferentialAttachment is a modest feature that is more informative than common co-author based features for collaborations more than two times as it yielded compelling F1 score in both cross validation (0.638) and testing (0.633). It preforms better with low-frequency subsets, with average ROC AUC of 0.599 for low-frequency and 0.506 for high-frequency collaborations. It is noticeable that the LR:Pref.Attach. baseline showed a strong performance in both validation and testing. In particular, it has the best recall among all models. In order to validate this performance, we further analyzed the prediction results and found that the recall of LR:Pref.Attach. was actually an artifact of the decision threshold. Concretely, when we fixed the recall at 0.791, the precision of the LR: All Baseline model is 0.536 and the F1 is 0.639, which is better than the LR:Pref.Attach. precision of 0.527 and F1 of 0.633. In addition, the information gain analysis provides further evidence that PreferentialAttachment is not as informative as the other top-ranked features such as the Research Interest Similarity features (*simText*, *simMeSH*, *simOutcite*) and the Geographical Location features (*sameLocation* and *sumOpenness*).

5.5.2 Error Analysis

We further analyzed the incorrect predictions. First, as shown in the feature analysis section above, the research interest features and geographical features are major players. Therefore, many of the error cases are also due to their mis-classification. In order to explore the source of error, we used the probability from the LR classification to rank the false positives and false negatives. Using these ranked error cases, we found that 90% of the top 100 author pairs with false positive predictions were co-located (i.e., the *sameLocation* value of that co-author pair was 1), while 100% of the top 100 false negative author pairs were in different locations (i.e., the *sameLocation* value of that co-author pair was 0); Second, we found that the 2-time collaboration frequency was not as informative for prediction (all the features proved non-informative for that frequency) and 55.7% of all the false negatives were 2-time collaborators, whose feature values are statistically similar to the feature values exhibited by the negative instances (i.e., one-time collaborations). For example, the mean of *simText* for 2-time collaborations was 0.459 while the mean for 3-time or higher frequency collaborations was 0.543 ±0.007, showing a distinct change in going from 2-time collaborations to 3-time or higher collaborations. The 2-time collaborations, in fact, were very close to the negative cases (i.e., 1-time collaborations), which had a mean of 0.399 for *simText*.

A subtle error source is the incomplete publication data. First, it makes the collaboration history incomplete. For example for the author pair of Cichon S and Kelsoe JR, their initial collaboration is article PMID 12802785 in year 2003. However by searching PubMed we found the two actually collaborated in 1997 with the publication PMID 9433543, which is missing in our database. Such types of instances have been found from both training and testing set and thus some noise has been introduced. Second, the incomplete publication history brought distortion to the author research interest profile. For instance Keijzer R has no publication prior to 2000 in our database but the author actually has publications in the 1990s (PMIDs 1360499, 7678025, 8630280 etc). The missing publications included studies on DNA, Lung, Rats, which are also the area of Post M. The author pair was predicted negative by our models due to the low research interest profile similarity (e.g. *simText* was only 0.044) but they did re-collaborate.

5.6 Conclusion

In this study, we modeled recurrent collaboration as a link prediction problem and used supervised machine learning models to predict whether two authors will collaborate again or not based on the features that are extracted by the time of their initial collaboration. We developed novel features to model an author's research interest profile, publication productivity, geographical location, author seniority, and the co-authorship network connectivity. Machine learning models, including logistic regression, support vector machines, naive Bayes, and random forest, were trained by 40,000 author pairs from the co-authorship network that were created using MEDLINE publications. The best performing model, random forest yielded a 0.720 ROC, which outperformed the best baselines by 11%. Our analysis shows that authors with high frequency collaboration have more distinctive characteristics compared with one-time-only collaborations. The model has the best performance on predicting the subset with 9-time collaboration instances as positive instances, for which it achieved a 0.824 ROC, while the prediction on 2-time collaboration instances yielded a 0.627 ROC. Our feature analysis further shows that research interest profile similarity is the most informative feature category and an author's location proximity is also a top predictor. In addition, local clustering of the author's neighbors contributes to the prediction, as well. An author's publication productivity, the average number of publications per year, is more related to high frequency collaboration. Our study sheds light on the characteristics of recurrent collaborations from a comprehensive yet simple analysis. Our prediction approach is highly salable and has a promising potential to be applied to research collaboration recommendation applications.

Chapter 6

Protein–Protein Interaction Prediction for Hypothesis Generation

6.1 Introduction

More and more biological knowledge databases have been made publicly available. Genome and sequence databases, as well as interacting protein databases, have been created and constantly updated since the early 1990s. Notable databases include the database of the complete sequencing of the yeast genome [112], the Biomolecular Interaction Network Database (BIND) [8], the Protein Information Resource (PIR) [9], the Munich Information Center for Protein Sequences (MIPS) [10], and the Database of Interacting Proteins (DIP) [157]. There is also a growing trend for single knowledge sources integrating with, or cross-referencing, other databases to make its data more comprehensive. For example, DIP cross-references three major sequence databases: Swiss-Prot, GeneBank, and PIR. In addition, results from high-throughput computational efforts increasingly contribute to these knowledge bases.

Protein–protein interaction (PPI) networks are important for understanding disease mechanisms [158] and determining drug targets. Interaction networks derived known interactions have unveiled PPI network structures within cells. Strikingly, PPI networks exhibit properties that have also been discovered in other networks, such as social networks and even the Internet; in particular, they demonstrate properties that dictate network growth and connectivity [159], such as being scale free (with some highly connected hubs of activity) and small world (where the nodes are highly clustered but the minimum distance between any two random nodes is short).

While not typically hubs, disease genes and proteins cluster in the same network neighborhoods, as shown by Goh and colleagues, who reported a 10-fold increase in the number of physical interactions observed between gene products associated with the same disease than would be expected by chance [160]. Chen et al. applied link analysis approaches to PPI networks to identify disease candidate genes [161]. In addition, genes linked to diseases with similar pathophenotypes have a higher likelihood of interacting with each other than those not linked to these pathophenotypes [162][163]. Taken together, these observations support the notion that the disease-related components of a network are likely to comprise a sub-network, or disease module.

Advances in supervised machine learning, an important subset of artificial intelligence, have made network interaction predictions increasingly reliable. Supervised machine learning models, such as logistic regressions and support vector machines (SVMs), as well as unsupervised approaches such as K-means clustering, are widely used for prediction tasks. Such analyses are supported by computational advances in conducting resource-intensive “big data” experiments.

In this chapter, we utilize supervised machine learning approaches to generate new hypothesis for PPIs. We define the task as a link prediction problem. Specifically, we formalize the problem as a classification task. Given two proteins whose interaction is unknown, we would like to determine if they can interact or not. We utilize existing knowledge bases to extract the proteins’ features to train the supervised machine learning models. We extract network features about the proteins, such as common neighbors, the number of neighbors, and clustering coefficients, by looking at protein connections within a network of existing PPIs obtained from the BioGRID database, a curated biological database of PPIs and genetic interactions. Using MEDLINE, we also extract text features from research articles about each of these proteins, including literature features, such as the similarity of MeSH (Medical Subject Headings) terms, article co-occurrence, and the number of publications. We use the network and literature features to train four supervised machine learning models: the naive Bayes (NB), naive Bayes multinomial (NBM), SVM, and logistic regression (LR) models. The best-performing model was the LR model, which achieved 0.95 AUC. Our approach is independent of the details of protein functions, and analyzes the protein network from a novel perspective. It is also efficient and highly generalizable.

6.2 Background

6.2.1 Biological Knowledge Bases

Rich resources such as genome and sequence databases, as well as interacting protein databases, have been created and constantly updated. For example, BIND is a database that stores biomolecular interactions, complexes, and pathway information [8]. The PIR provides a protein sequence database [9]. The MIPS in Germany hosts repositories for genome and sequence data [10, 11, 12, 13, 14] in collaboration with PIR. UniProt [15] integrates sequence and gene function information with information from both literature curation and automatic classification from Swiss-Prot [16, 17], PIR, and other resources. GeneOntology [18] provides a structured and controlled vocabulary that describes the roles of genes and gene products. DIP is a repository of experimentally determined interacting proteins [157][164].

6.2.2 PPI Networks

Since the late 1990s, high-throughput computational approaches have been heavily used to create protein interaction networks. A yeast protein network was built with over 2,000 interactions between more than 1,000 proteins curated from the literature [56]. STRING is a database of functional associations between proteins [57]. BioGRID is a database that incorporates physical and genetic protein interactions that have been manually curated from primary literature [58]: As of 2013, it has archived 500,000 manually annotated interactions from more than 30 model organisms [59]. There is, in fact, a trade-off between coverage (the number of protein–protein pairs classified as interacting) and accuracy (whether the protein–protein pair in fact interacts) due to high-throughput, large-scale experiments [165]. Rual et al. [166] used a high-throughput yeast two-hybrid (Y2H) system to create a human PPI map. Rhodes et al. [167] employed trained probabilistic models (decision tree, naive Bayes) on existing databases, including interaction networks, expression, and gene ontology, to predict new interactions for human genes. Yu et al. [168] performed a quality assessment of a current Y2H network system and conducted analysis that reveals new characteristics of the Y2H network. Lim et al. [158] used Y2H screening to construct a human PPI network for a particular disease (human inherited ataxias). Jones et al. [169] used residue patches on the surface of protein structures to predict the location of PPI sites. Zhong et al. [170] integrated interactive data, gene expression data, phenotype data, and functional annotation data to model the interactions of three organisms. Utilizing multiple data sources and a naive Bayes model, the yeast functional gene network was used to construct a functional gene network [171, 172]. In our approach, we use naive Bayes as well as more complex machine learning models and features to predict PPI.

6.2.3 PPI Prediction

Probabilistic predictions have been applied to proteome data to discover new relations between proteins or their functions. Jansen et al. [173] used a Bayesian network model to predict PPIs in yeast using both experimental and genomic data. Albert et al. [174] used topological patterns in PPI networks to train supervised machine learning models. Specifically, they identified the number of occurrences of a particular connection pattern for a pair and used it as their feature. Tong et al. [175] used both computational and experimental approaches to derive two separate networks and used their intersection for high-confidence prediction. Enright et al. [176] predicted PPI by gene fusion events identified by gene sequence comparisons. Bader et al. [177] explored graph-based predictors in a Y2H model and co-immunoprecipitated (Co-IP) protein complex data with logistic regression to show that the protein complexes to which the two proteins belong are the most important predictor. Tsuda et al. [178] used kernel methods to predict the weight for protein–protein associations in the context of five different networks, such as physical interactions and genetic interactions. The authors subsequently combined the weights of a particular pair across the networks for classification. Homologous is also used for inferring interactions [179]. Bock et al. [180] used SVMs with features extracted from protein structures. Samanta et al. [181] used common interactors between two proteins as predictors of functional association. If two proteins shared a significantly large number of interactors compared to the value in the random graph, they were deemed more likely to have functional associations.

Qi et al. [182] evaluated classification methods such as SVM, naive Bayes, and logistic regression on different data sources for PPI prediction. They separated the prediction task into physical interactions, co-complex relationships, and pathway co-memberships and extracted features from multiple data sources, including DIP and Gene Ontology. Their results show that gene expression data are the most important feature. Using a similar data source, Qi et al. [183] explored a random forest based approach to compute the similarities between two proteins and applied K-nearest neighbors for prediction. We use the physical interaction prediction approach of these authors as our main state-of-the-art comparison and find that our approach achieves significantly higher prediction, as detailed in the results section.

Microarray data are also used for training machine learning models to predict the function class of a gene [184]. Work on PPI networks varies by scale and species and by whether it is an original network or a refinement. In addition, considerable work has been conducted in comparing and assessing different knowledge bases. The Y2H system is an example of such an intensively studied dataset [185, 186].

Network-based approaches have been widely used for functional predictions [187]. The second-degree neighbor (the neighbor of a neighbor) feature has been found to be an important predictor of protein functions in PPI networks [188]. In addition, PPI networks created by predictions can be used to detect functional modules [189]. Clustering algorithms have been proposed to discover protein complexes from the analysis of PPI networks [190]. In addition, PPI networks are used to predict the cellular functions of proteins, utilizing, for example, distance-based predictors [191].

6.3 Methods

6.3.1 Problem Formulation

We propose using supervised machine learning methods to predict potentially interacting pairs of proteins. Given a cutoff year t_0, we sample interacting pairs that were discovered before t_0. Assume an interacting pair of proteins, u and v, in the training set that is described by the tuple (u, v, t_i), where t_i is the time of the interaction and $t_i < t_0$. The features of this interaction are extracted from the state of the network prior to t_i. In other words, we learn from the history of the two interacting proteins.

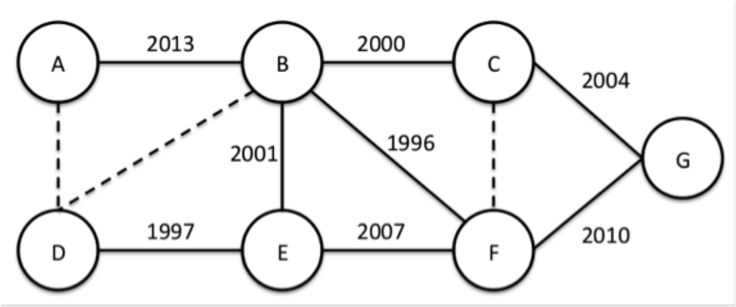


Figure 6.1: An illustration of a PPI network. The solid lines represent existing interactions between two proteins, marked by the year the interaction was discovered. The dashed lines are negative interactions, which indicate no interactions discovered between those two proteins up to that point in time.

Negative instances are always random protein pairs that never interacted before t_0. An illustration of a network model is shown in Figure 6.1. Therefore, given a new pair of proteins that have never interacted before t_0, we use the model to predict the probability of interaction after t_0.

6.3.2 Supervised Machine Learning Models

We predict potential protein pair interactions by ranking protein pairs according to the probability of their interaction as determined by a classifier. In the link prediction task, there is no absolute negative instance, since each pair for which we have not seen an interaction so far may interact in the future. Our work is based on the assumption that the network is sparse and most of the pairs are not going to interact at all; those pairs with no recorded interaction up to the cutoff time are what the model will use to learn negative instances. Given a new, possibly interacting pair of proteins, we use the model to predict the probability of the interaction. If a pair of likely interacting proteins in the database has a high probability (rank) of doing so, that pair is then considered a possible interaction prediction.

We explore four widely used supervised machine learning models for this PPI prediction: NB, NBM, SVM, and LR. A naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with the strong (naive) independence assumption that the features are generated independently from each other, given the instance label [137]. The naive Bayes multinomial model assumes the conditional probability of the feature, given that a class (the likelihood) follows a multinomial distribution [138, 139]. SVMs are based on the concept of maximum margin decision planes that define generalizable decision boundaries for classification and regression. An SVM constructs a hyperplane to maximize the margin between the data points and the hyperplane, often after mapping the data points to a higher-dimensional space in which they are linearly separable or close to it [140]. In particular, we explore the linear kernel for its efficiency. Logistic regression estimates the parameters from discrete or continuous values to predict discrete category values. The probabilities that describe the possible class of a single instance are trained as a function of explanatory variables, using a logistic function [137]. The four aforementioned classifiers are not only the best-performing models demonstrated in a variety of classification tasks, but also robust, fast, and easy to implement. We use the data mining software Weka [143] for model training and testing.

6.3.3 Features

6.3.4 Literature Features

We hypothesize that there is the evidence in the literature for the interaction between two proteins. The measure *simMeSH* is the similarity of the MeSH terms in the publication history of the two proteins. Our intuition is that if the two proteins have a certain degree of similarity according to MeSH, they are likely to be functionallly related. The measure *JaccardArticleCoOccurence* is based on the assumption that the mention of both proteins in the same article suggests the two proteins are related. The measure *sumPub* is an indicator of the research effort focused on the two proteins. This category of features can be directly extracted from our indexed MEDLINE records. With a protein's name and its synonyms, we can construct queries to obtain all the abstracts that mention it. We subsequently use the MeSH terms of these articles to calculate *simMeSH* and use the articles to calculate *JaccardArticleCoOccurence* and *sumPub*. In other words,

$$simMeSH(x,y) = \frac{|M(x) \cap M(y)|}{|M(x) \cup M(y)|}$$

where M(.) is the set of MeSH terms of all articles that mention the protein and

$$sumPub(x,y) = |Pub(x)| + |Pub(y)|$$

where Pub(.) is the set of publications that mention the protein. In addition,

$$jaccardArticleCoOccurence(x,y) = \frac{|Pub(x) \cap Pub(y)|}{|Pub(x) \cup Pub(y)|}$$

6.3.5 Network Features

The hypothesis of this category of features is that the connectivity of a protein node in a PPI network implies its likelihood of connecting to a new protein. We define *numCommonNeighbor* as the number of common neighbors of two proteins,

$$numCommonNeighbor(x,y) = |\Gamma(x) \cap \Gamma(y)|$$

where Γ(.) is the neighbors of the node and *sumNeighbor* is simply the total number of neighbors of the two proteins,

$$sumNeighbor(x,y) = |\Gamma(x)| + |\Gamma(y)|$$

As described by Small (1999) [69] and Watts(1998) [62], the clustering coefficient of the vertex v in a graph is the proportion of v's neighbors that connect among themselves. For example, assume v has four neighbors (i.e., six possible connections among them, since the number of "4 choose 2" combinations) and there are three pre-existing links. Therefore, the clustering coefficient is 0.5 (3/6). This coefficient has been found to be an useful predictor of PPI [174]. The feature *sumClusteringCoef* is the sum of the two nodes' clustering coefficients to obtain information about the edge between the two nodes,

$$sumClusteringCoef(x,y) = cluterCoef(x) + cluterCoef(y)$$

where the clustering coefficient $clusterCoef(x) = \frac{|E'|}{n(n-1)/2}$ and E' is the edges among the neighbors of x. The Adamic, introduced by Adamic and Adar (2003) [136], is a measurement of the similarity between two Web pages. The rationale is that two Web pages are more similar if they share more unique items (links, text). Liben-Nowell and Kleinberg [66] later applied the Adamic to measure node similarity in social networks and showed it to be an effective predictor for establishing new links. In this chapter, we apply it to measure the topological similarity of two proteins,

$$Adamic(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log|\Gamma(z)|}$$

The *Jaccard* measure is the number of common neighbors divided by the number of total unique neighbors of the two proteins,

$$Jaccard(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$

Both the Adamic and the Jaccard coefficient are topological measures that are effective for link prediction tasks and we therefore incorporate them here as well. The

Table 6.1: Feature definitions.

Category	Feature Name	Definition	Source
Literature Features	simMeSH	$\frac{\lvert M(x) \cap M(y)\rvert}{\lvert M(x) \cup M(y)\rvert}$	MEDLINE
	sumPub	$\lvert Pub(x)\rvert + \lvert Pub(y)\rvert$	MEDLINE
	jaccardArticleCoOccurrence	$\frac{\lvert Pub(x) \cap Pub(y)\rvert}{\lvert Pub(x) \cup Pub(y)\rvert}$	MEDLINE
Network Features	Adamic	$\sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log\lvert\Gamma(z)\rvert}$	BioGRID
	numCommonNeighbor	$\lvert\Gamma(x) \cap \Gamma(y)\rvert$	BioGRID
	Jaccard	$\frac{\lvert\Gamma(x) \cap \Gamma(y)\rvert}{\lvert\Gamma(x) \cup \Gamma(y)\rvert}$	BioGRID
	sumNeighbor	$\lvert\Gamma(x)\rvert + \Gamma(y)$	BioGRID
	sumClusteringCoef	$clustCoef(x) + clustCoef(y)$	BioGRID

network features described here are constructed solely from the interaction network, which is built from the BioGRID database. Formal definitions of all our features are shown in Table 6.1.

6.4 Data

6.4.1 BioGRID

We constructed the PPI network using the BioGRID database, a comprehensive interaction repository with over 648,000 interactions from over 38,000 PubMed articles. Our network contains 48,438 unique interactors from 45 different organisms and 26.2% of them have only one interaction. The mean degree is 21 and the median is four, where a node's degree is the number of edges at that vertex; the minimum and maximum degrees are one and 10,095, respectively. In addition, the standard deviation of the number of degrees of the vertices is 75. The distribution of the degree of the nodes can be fit by either a power-law or a lognormal distribution. The β in the $logy = \alpha - \beta logx$ of the power-law fit is 3.76 for the range in which the node degree (horizontal axis) is larger than 530. The node degree distribution of BioGRID proteins is shown in Figure 6.2. x is the protein node degree and y is the frequency of the proteins who have degree x.

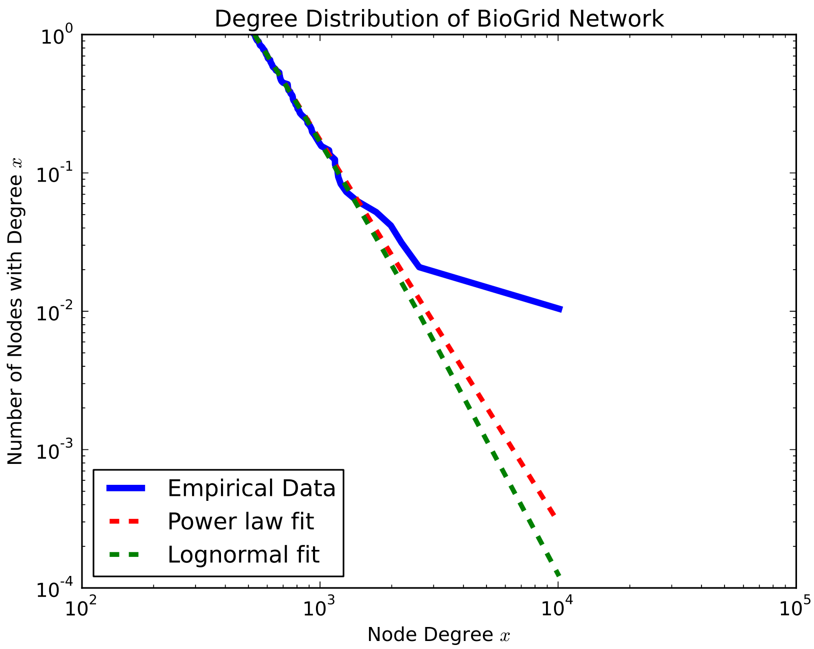


Figure 6.2: Node degree distribution of BioGRID proteins (double logarithmic scale). The results for the dashed line are obtained by the linear regression.

6.4.2 MEDLINE

As described in Chapters 1 and 2, Medline is a database of biomedical literature with over 20 million article records provided by PubMed. Each article record contains the title, author(s), journal, year, and MeSH terms. Every article is also assigned a unique PubMed ID.

6.4.3 Training and Testing Sets

We construct a training set named BIG by randomly selecting 16,278 (48.9%) positive pairs and 17,011 negative pairs before the year 2011. The testing set is constructed from the 2013 data and consists of 3,153 positive protein pairs showing interactions before 2013 and 3,153 negative protein pairs with no known interaction before 2013. The negative instances were randomly chosen to match the number of positive protein pairs; 3,153 positive protein pairs were chosen because that was the number of positive protein pairs in the dataset.

Furthermore, we evaluate our approach with longitudinal datasets. Specifically, we construct a training set by using data before year y and use it to predict the new pairs in year y. We apply this approach from year 1995 to year 2012. For each year, we randomly select 1,000 positive and 1,000 negative training instances from the data before the year y and train the models. We then sample up to 1,000 new pairs that were reported for the first time in y as positive test instances and select random pairs as negatives. In total, there are 18 training sets and 18 testing sets.

Note that it is difficult to find a real negative instance from knowledge bases, since findings about a non-interacting protein pair are much less common. However, the density of the protein interaction network is as small as 0.00056. Therefore, given two randomly selected proteins, the chance that they interact is small and we can use the pair as a negative instance.

6.4.4 Evaluation Metrics

We calculate precision (TP/(TP + FP)), recall (TP/(TP + FN)), the receiver operating characteristic (ROC) curve—or ROC AUC, the area under the curve of the true positive rate (TPR) over the false positive rate (FPR)—and accuracy ((TP+TN)/ALL), where TP, FP, TN, FN, and ALL stand for the numbers of true positives, false positives, true negatives, false negatives, and total instances, respectively. The F1 score is defined as the harmonic mean of recall and precision, specifically, 2*recall*precision/(recall+precision).

6.5 Results

6.5.1 10-Fold Cross-Validation

For the BIG training set evaluation, we use 10-fold cross-validation and the results are shown in Table 6.2. Logistic regression achieves the best performance, with 0.856

Table 6.2: 10-fold cross-validation on the training set.

	Model	ROC AUC	Precision	Recall	F1
All features	NB	0.832	0.787	0.697	0.668
	NBM	0.738	0.639	0.618	0.607
	SVM	0.765	0.771	0.766	0.765
	LR	0.856	0.781	0.772	0.769
Topological Features Only	NB	0.763	0.788	0.691	0.659
	NBM	0.690	0.811	0.727	0.705
	SVM	0.740	0.799	0.745	0.731
	LR	0.766	0.815	0.757	0.744
Literature Features Only	NB	0.770	0.690	0.667	0.654
	NBM	0.763	0.711	0.710	0.710
	SVM	0.708	0.708	0.708	0.708
	LR	0.783	0.712	0.712	0.712
Common Neighbor Baseline	LR	0.754	0.824	0.760	0.746
Co-occurrence Baseline	LR	0.555	0.605	0.538	0.432

AUC and 0.769 F1. The naive Bayes model is the second best, with 0.832 AUC and 0.668 F1. The naive Bayes multinomial and SVM models are less effective than the previous two, with 0.738 AUC and 0.765 AUC, respectively.

Furthermore, we experiment on topological and literature feature subsets separately. The best AUC for the topological subset is 0.766 and the best for the literature subset is 0.783, both achieved with the logistic regression model. It is noticeable that the naive Bayes multinomial has better performance for the literature subset than for the overall feature set.

We calculate information gain to better understand the contribution of the features. As shown in Table 6.3, the network features' Adamic has the highest information gain, 0.306, followed by *numCommonNeighbor* (0.291), *Jaccard*(0.287), and *sumNeighbor* (0.176), which are all in this category. The simMeSH measure, with information gain 0.122, is ranked the highest among the literature features. Next highest is *sumClusteringCoef* (0.105) and at the bottom is the article co-occurrence

Table 6.3: Information gain of the training set BIG. Neighbor-based measures *Adamic*, *numCommonNeighbor*, *Jaccard*, and *sumNeighbor* are the four top-ranked features, with information gain ranging from 0.178 to 0.306, which shows they are the most informative features.

Rank	Feature	Information Gain
1	*Adamic*	0.306
2	*numCommonNeighbor*	0.291
3	*Jaccard*	0.287
4	*sumNeighbor*	0.176
5	*sumPub*	0.135
6	*simMeSH*	0.122
7	*sumClusteringCoef*	0.105
8	*jaccardArticleCoOccur*	0.038

feature *jaccardArticleCoOccur* (0.038).

6.5.2 Evaluation Using the DIP Database

Since yeast is a widely studied species, we particularly evaluate our approach to protein interactions within this species. We use protein pairs from the data of Qi et al. and compare our results with theirs. As described in the background, one of the tasks is to use machine learning to predict physical interaction and the authors developed comprehensive biological features.

We downloaded their data from the project website: They include 2,865 positives and 237,384 negatives. During preprocessing, we mapped protein symbols to our network, resulting in 2,704 valid positives and 122,161 negatives. We also double-checked the protein pairs against our training set and there was no overlap between the training and testing sets.

As shown in Tables 6.4 and 6.5, the models trained by our features outperform the data of Qi et al. In 10-fold cross-validation, the LR model trained by all features yields 0.887 F1, while that trained by Qi et al. results in 0.560. Similar results are shown in the test (the LR for our feature yields 0.892 F1 and 0.547 F1 for the feature of Qi et al.).

Table 6.4: 10-fold cross-validation using proteins from the dataset of Qi et al.

Model	ROC AUC	Precision	Recall	F1
All Features	0.990	0.996	0.887	0.887
Literature Features	0.856	0.899	0.303	0.453
Network Feature	0.990	0.959	0.816	0.882
Qi et al. Detailed Feature	0.922	0.689	0.471	0.560

Table 6.5: Testing on protein pairs from the dataset of Qi et al.

Model	ROC AUC	Precision	Recall	F1
All Features	0.982	0.922	0.865	0.892
Literature Features	0.868	0.934	0.364	0.524
Network Features	0.989	0.959	0.821	0.885
Qi Detailed Feature	0.919	0.720	0.441	0.547

6.5.3 Predicting the Next Year

We also test our approach by predicting the next year, using the model trained by existing interactions. As shown in Figure 6.4, the performance of the overall data is generally consistently above 0.7 ROC, except for a drop in the year 2003 and 2004. The predictions for the human species are also generally above 0.7 ROC, except for a drop in 2012. In contrast, all the models trained and tested on yeast subsets have remained below 0.7 ROC since 2002.

We further analyze the predictions for the two feature categories literature features and network features. Specifically, we train LR models by using these two subcategories and evaluate them by predicting the next year. As shown in Figure 6.4, network features outperform literature feature for each year's predictions. Both the literature and network models consistently show a significant performance drop in the year 2003.

Furthermore, we predict the feature subsets for the human and yeast systems in the next year. As shown in Figure 6.5, network features also consistently outperform literature features for all years except 2012, which is similar to the results for the overall data (Figure 6.4). However, for yeast, there is no clear winner between the

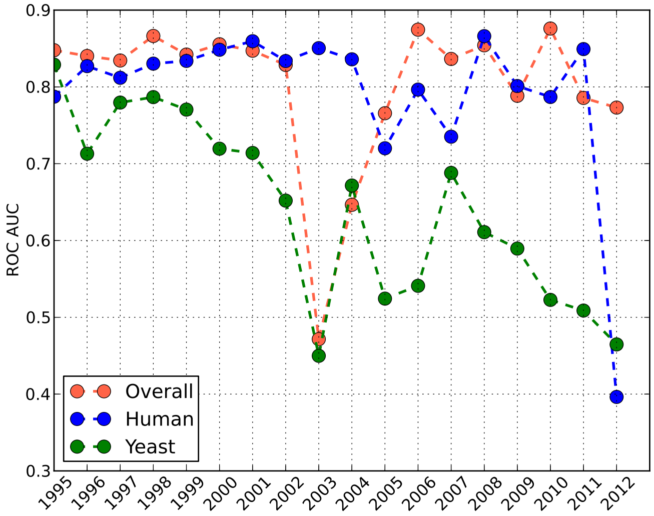


Figure 6.3: Evaluation of all the data.

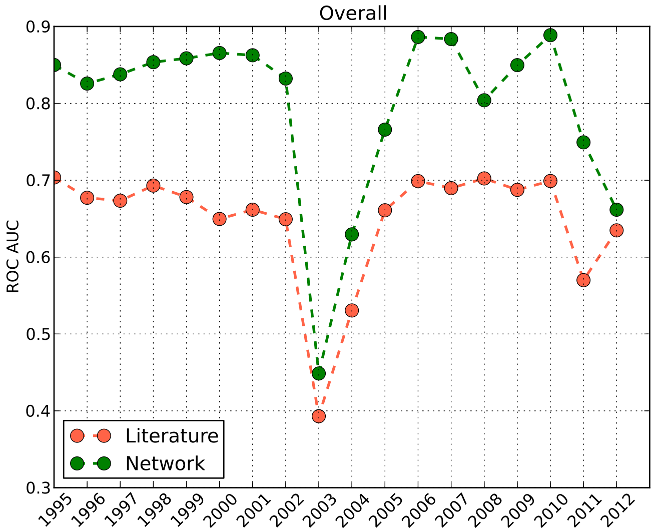


Figure 6.4: Predicting next year by subsets of features (literature features and network features). Network features outperform literature features for each year's predictions.

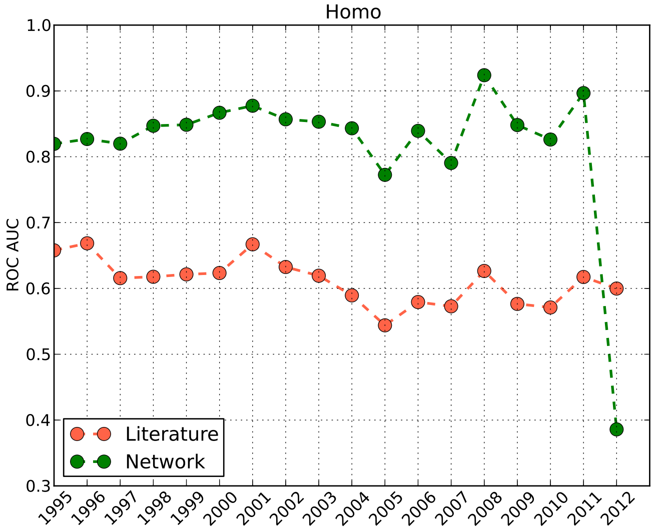


Figure 6.5: Evaluation on the human system.

two feature categories. Network features have better performance in 1995–2000, but the literature features show advantages in all the years after 2000, except for 2002, 2004, and 2007.

Our approach demonstrates a performance advantage in predicting PPIs. Using the protein pairs sampled from the DIP database, the model trained by the features we propose outperforms the model trained by biological features. As shown in Table 6.5, our model has 0.982 ROC AUC and 0.892 F1, while the state-of-the-art results (Qi et al.) are 0.922 ROC AUC and 0.560 F1, respectively. By further exploring the feature set of Qi et al., we found that 37.9% of the feature values are unavailable, suggesting that biological features are very sparse. On the other hand, our feature set produces less sparse features. The upper bound of missing values in our feature set for the DIP evaluation data is 31.8% (31.8% of feature values are zero).

Overall, the network features are more informative in predicting new interactions than the literature features. As shown in the information gain analysis (Table 6.3), the network features are all ranked higher than the literature features.

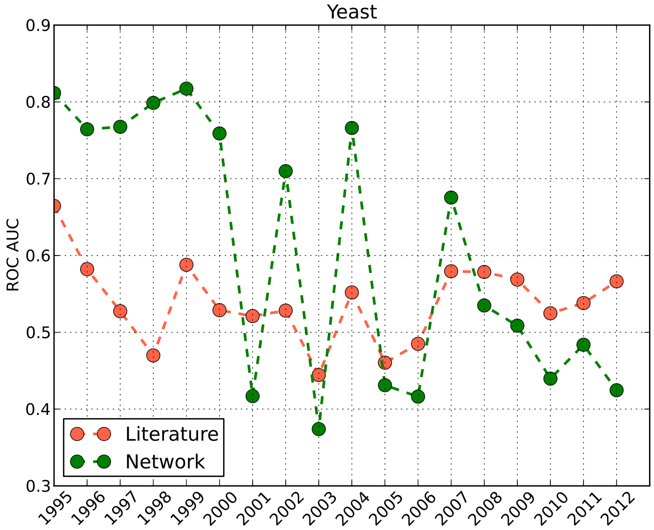


Figure 6.6: Evaluation on the yeast system.

6.6 IBioNet.org

We implemented our system and it is available from ibionet.org. We generated 2 million new hypothesis for 415 proteins researched by the University of Massachusetts Medical School so that we provide the researchers one more resource for identifying potential interactors.

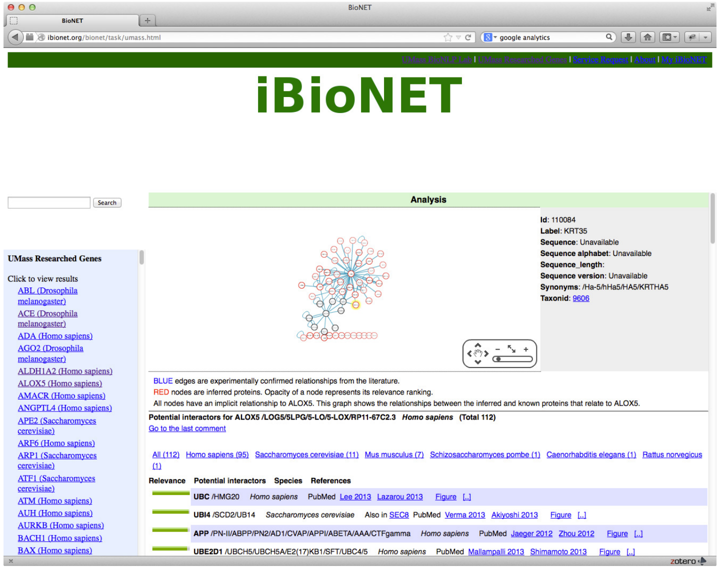
The system has four components: a BFS candidate collector, a feature extractor, an interaction classifier, and a result generator.

The BFS candidate collector generates candidate interactors from the PPI network. For instance, assume we want to predict new interactors for the protein retinitis pigmentosa 2 (RP2), then we use the collector to obtain the subgraph of RP2, where each protein is within six hops of another. We consider those proteins that have never interacted with RP2 as candidates.

The feature extractor extracts the features of potential interacting pairs. With the RP2 example, the extractor extracts features for each pair (for RP2 and one of the candidates). The component is based on Hadoop framework [76].

The pairs with extracted features are then fed into the interaction classifier for

Figure 6.7: IBioNet page providing predictions for genes researched by the University of Massachusetts Medical School.


BioNET
BioNET
ibionet.org/bionet/task/umass.html
google analytics
UMass BioNLP Lab | UMass Researched Genes | Service Request | About | My iBioNET
Search
zotero

iBioNET

Analysis

Id: 110084
Label: KRT35
Sequence: Unavailable
Sequence alphabet: Unavailable
Sequence_length:
Sequence version: Unavailable
Synonyms: /Ha-5/hHa5/HA5/KRTHA5
Taxonid: 9606

UMass Researched Genes

Click to view results
- ABL (Drosophila melanogaster)
- ACE (Drosophila melanogaster)
- ADA (Homo sapiens)
- AGO2 (Drosophila melanogaster)
- ALDH1A2 (Homo sapiens)
- ALOX5 (Homo sapiens)
- AMACR (Homo sapiens)
- ANGPTL4 (Homo sapiens)
- APE2 (Saccharomyces cerevisiae)
- ARF6 (Homo sapiens)
- ARP1 (Saccharomyces cerevisiae)
- ATF1 (Saccharomyces cerevisiae)
- ATM (Homo sapiens)
- AUH (Homo sapiens)
- AURKB (Homo sapiens)
- BACH1 (Homo sapiens)
- BAX (Homo sapiens)

BLUE edges are experimentally confirmed relationships from the literature.
RED nodes are inferred proteins. Opacity of a node represents its relevance ranking.
All nodes have an implicit relationship to ALOX5. This graph shows the relationships between the inferred and known proteins that relate to ALOX5.

Potential interactors for ALOX5 /LOG5/5LPG/5-LO/5-LOX/RP11-67C2.3 ***Homo sapiens*** **(Total 112)**
Go to the last comment

All (112) Homo sapiens (95) Saccharomyces cerevisiae (11) Mus musculus (7) Schizosaccharomyces pombe (1) Caenorhabditis elegans (1) Rattus norvegicus (1)

Relevance	Potential interactors	Species	References
🟩	**UBC** /HMG20	*Homo sapiens*	PubMed Lee 2013 Lazarou 2013 Figure [..]
🟩	**UBI4** /SCD2/UB14	*Saccharomyces cerevisiae*	Also in SEC8 PubMed Verma 2013 Akiyoshi 2013 Figure [..]
🟩	**APP** /PN-II/ABPP/PN2/AD1/CVAP/APPI/ABETA/AAA/CTFgamma	*Homo sapiens*	PubMed Jaeger 2012 Zhou 2012 Figure [..]
🟩	**UBE2D1** /UBCH5/UBCH5A/E2(17)KB1/SFT/UBC4/5	*Homo sapiens*	PubMed Mallampalli 2013 Shimamoto 2013 Figure [..]

prediction. The probability of interaction is calculated for each input pair. This component is also implemented as a Hadoop application.

Post-processing is carried out by the result generator. It imports the results into a MySQL database and associates it with other protein-related information, such as official names, synonyms, species, and related publications.

6.7 Conclusion

In this study, we propose using a machine learning approach to generate a new PPI hypothesis. We model PPI prediction as a link prediction problem. Existing interactions are considered positive training data and the features are extracted from the protein pairs by the time of the discovery of their interaction. We use the PPI database BioGRID, as well as MEDLINE, as our data source for creating training data. We conduct comprehensive evaluations on BioGRID data, as well as the DIP database, on overall species, as well as on human and yeast species only. In addition, we conduct a longitudinal evaluation by creating 18 training sets and 18 test sets across 18 years to evaluate our approach on each year's new discoveries. Logistic regression shows the best performance, with ROC AUC 0.856 on 10-fold cross-validation. Our approach is generalizable, since the model trained by protein pairs from the DIP database achieves 0.982 ROC AUC in predicting physical interactions, which significantly outperformed the state-of-the-art system of Qi et al. We find that network features, such as the Adamic and the number of common friends, are the most important predictors for PPIs. We also build the hypothesis generation system ibionet.org, which provides University of Massachusetts Medical School researchers a new tool for identifying possible PPIs.

Chapter 7

Conclusions

In this thesis we have shown that network topological structure as well as semantic information of the node can be leveraged to predict new links in a network when we put network evolution into perspective. The network status, both network structure and semantic information, by the time of the link creation can be used to learn the patterns. Concretely we model link prediction as a classification problem, and we use network status by the time of establishment of existing links as features and the link as label. Therefore the classification models can be trained, and used for predicting the probability of being connected given a new pair of nodes and their related network status.

In research collaboration network, we derived features such as number of common collaborators, research interest similarity, research productivity. We found that, research interest similarity, such as publication history similarity, citation similarity, as well as number of common researchers, are most informative predictors for first time collaboration. Future more, geographical location as well as research interest similarity are top-ranked features for predicting recurrent research collaboration.

On the other hand, network structure features show advantages in protein-protein interaction prediction. We also introduced literature features such as the protein pair's number of co-occurrence in the articles. Yet, the prediction performance varies dramatically across different settings, such as species. Literature features tend to perform stable compared to the network features.

We also created the citation and co-authorship network for biomedical literature. Our analysis showed that the networks pertained the characteristics, such as scale-free and small world, that have observed in other complex networks. Our network system provides an unique resource for studying the citation and co-authorship in biomedical field.

7.1 Future Directions

There are several future directions. First, the group-group collaboration recommendation will be very useful. In real world collaborations often initiated between two labs in stead of two individuals, and the collaboration is often based on the purpose of integrating the expertise from different disciplines. The collaboration prediction task is more challenging yet extremely useful. Second, protein interactions tend to have restrictions, for example they have to be in the same type of cell (e.g. T-cell) or same tissue (brain). The future prediction should take these factors into account.

Bibliography

[1] Web of Knowledge Thomson Reuters;. Available from: `http://wokinfo.com/`.

[2] Google Scholar;. Available from: `http://scholar.google.com/`.

[3] Elsevier Scopus;. Available from: `https://www.elsevier.com/online-tools/scopus/features`.

[4] CiteSeerX;. Available from: `http://citeseerx.ist.psu.edu/`.

[5] PubMed - NCBI;. Available from: `http://www.ncbi.nlm.nih.gov/pubmed`.

[6] OMIM - NCBI;. Available from: `http://www.ncbi.nlm.nih.gov/omim`.

[7] DrugBank;. Available from: `http://www.drugbank.ca/`.

[8] Bader GD, Betel D, Hogue CW. BIND: the biomolecular interaction network database. Nucleic acids research. 2003;31(1):248–250. 00908. Available from: `http://nar.oxfordjournals.org/content/31/1/248.short`.

[9] Barker WC, Garavelli JS, Huang H, McGarvey PB, Orcutt BC, Srinivasarao GY, et al. The protein information resource (PIR). Nucleic acids research. 2000;28(1):41–44. 00160. Available from: `http://nar.oxfordjournals.org/content/28/1/41.short`.

[10] Mewes HW, Albermann K, Heumann K, Liebl S, Pfeiffer F. MIPS: a database for protein sequences, homology data and yeast genome information. Nu-

cleic Acids Research. 1997;25(1):28–30. 00238. Available from: http://nar.oxfordjournals.org/content/25/1/28.short.

[11] Mewes HW, Hani J, Pfeiffer F, Frishman D. MIPS: a database for protein sequences and complete genomes. Nucleic Acids Research. 1998;26(1):33–37. 00133. Available from: http://nar.oxfordjournals.org/content/26/1/33.short.

[12] Mewes HW, Frishman D, Gruber C, Geier B, Haase D, Kaps A, et al. MIPS: a database for genomes and protein sequences. Nucleic acids research. 2000;28(1):37–40. 00319. Available from: http://nar.oxfordjournals.org/content/28/1/37.short.

[13] Mewes HW, Frishman D, Güldener U, Mannhaupt G, Mayer K, Mokrejs M, et al. MIPS: a database for genomes and protein sequences. Nucleic acids research. 2002;30(1):31–34. 00875. Available from: http://nar.oxfordjournals.org/content/30/1/31.short.

[14] Mewes HW, Amid C, Arnold R, Frishman D, Güldener U, Mannhaupt G, et al. MIPS: analysis and annotation of proteins from whole genomes. Nucleic acids research. 2004;32(suppl 1):D41–D44. 00489. Available from: http://nar.oxfordjournals.org/content/32/suppl_1/D41.short.

[15] Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. The universal protein resource (UniProt). Nucleic acids research. 2005;33(suppl 1):D154–D159. 01317. Available from: http://nar.oxfordjournals.org/content/33/suppl_1/D154.short.

[16] Bairoch A, Boeckmann B. The SWISS-PROT protein sequence data bank. Nucleic Acids Research. 1991;19(Suppl):2247. 00368. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC331359/.

[17] Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic acids research. 2003;31(1):365–370. 02580. Available from: http://nar.oxfordjournals.org/content/31/1/365.short.

[18] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nature genetics. 2000;25(1):25–29. 11892. Available from: http://www.nature.com/ng/journal/v25/n1/abs/ng0500_25.html.

[19] Garfield E. Citation analysis as a tool in journal evaluation; 1972. .

[20] Garfield E. The history and meaning of the journal impact factor. Jama. 2006;295(1):90–93. Available from: http://jama.jamanetwork.com/article.aspx?articleid=202114.

[21] Garfield E. Journal impact factor: a brief review. Canadian Medical Association Journal. 1999;161(8):979–980. Available from: http://www.cmaj.ca/content/161/8/979.short.

[22] Garfield E. The meaning of the impact factor. Revista Internacional de Psicología clínica y de la Salud. 2003;3:363–369. Available from: http://core.kmi.open.ac.uk/download/pdf/5330795.pdf.

[23] Hirsch JE. An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences of the United states of America. 2005;102(46):16569. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1283832/.

[24] Bordons M, Fernández MT, Gómez I. Advantages and limitations in the use of impact factor measures for the assessment of research performance. Scien-

tometrics. 2002;53(2):195–206. Available from: http://www.akademiai.com/index/E63NN06QD6H05P2T.pdf.

[25] Caramoy A, Korwitz U, Eppelin A, Kirchhof B, Fauser S. Analysis of Aggregate Impact Factor Inflation in Ophthalmology. Ophthalmologica. 2012;Available from: http://content.karger.com/produktedb/produkte.asp?doi=343081.

[26] Opthof T. Sense and nonsense about the impact factor. Cardiovascular research. 1997;33(1):1–7. Available from: http://cardiovascres.oxfordjournals.org/content/33/1/1.short.

[27] Seglen PO. Why the impact factor of journals should not be used for evaluating research. BMJ: British Medical Journal. 1997;314(7079):498. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2126010/pdf/9056804.pdf.

[28] Van Raan AF. Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. scientometrics. 2006;67(3):491–502. Available from: http://www.akademiai.com/index/JX04N72413RK03U5.pdf.

[29] Bornmann L, Daniel HD. Does the h-index for ranking of scientists really work? Scientometrics. 2005;65(3):391–392. Available from: http://www.akademiai.com/index/K105L61U11G22127.pdf.

[30] Nakov PI, Schwartz AS, Hearst M. Citances: Citation sentences for semantic analysis of bioscience text. In: Proceedings of the SIGIR'04 workshop on Search and Discovery in Bioinformatics; 2004. p. 81–88. Available from: http://biotext.berkeley.edu/papers/citances-nlpbio04.pdf.

[31] Han H, Giles L, Zha H, Li C, Tsioutsiouliklis K. Two supervised learning approaches for name disambiguation in author citations. In: Digital Libraries, 2004. Proceedings of the 2004 Joint ACM/IEEE Conference on. IEEE; 2004. p. 296–305. Available from: `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1336139`.

[32] Han H, Xu W, Zha H, Giles CL. A hierarchical naive Bayes mixture model for name disambiguation in author citations. In: Proceedings of the 2005 ACM symposium on Applied computing. ACM; 2005. p. 1065–1069. Available from: `http://dl.acm.org/citation.cfm?id=1066920`.

[33] Kajikawa Y, Ohno J, Takeda Y, Matsushima K, Komiyama H. Creating an academic landscape of sustainability science: an analysis of the citation network. Sustainability Science. 2007;2(2):221–231. Available from: `http://link.springer.com/article/10.1007/s11625-007-0027-8`.

[34] Kajikawa Y, Takeda Y. Citation network analysis of organic LEDs. Technological Forecasting and Social Change. 2009;76(8):1115âĂŞ1123.

[35] Chen P, Redner S. Community structure of the physical review citation network. Journal of Informetrics. 2010;4(3):278âĂŞ290.

[36] Hummon NP, Dereian P. Connectivity in a citation network: The development of DNA theory. Social Networks. 1989;11(1):39–63. Available from: `http://www.sciencedirect.com/science/article/pii/0378873389900178`.

[37] Calero-Medina C, Noyons E. Combining mapping and citation network analysis for a better understanding of the scientific development: The case of the absorptive capacity field. Journal of Informetrics. 2008;2(4):272âĂŞ279.

[38] Zhang P, Koppaka L. Semantics-based legal citation network. In: Proceedings

of the 11th international conference on Artificial intelligence and law; 2007. p. 123–130.

[39] CsÃąrdi G, Strandburg K, ZalÃąnyi L, Tobochnik J, ÃĽrdi P. Modeling innovation by a kinetic description of the patent citation system. Physica A: Statistical Mechanics and its Applications. 2007;374(2):783âĂŞ793.

[40] Bilke S, Peterson C. Topological properties of citation and metabolic networks. Physical Review E. 2001;64(3):036106.

[41] Redner S. Citation statistics from more than a century of physical review. Arxiv preprint physics/0407137. 2004;.

[42] De Castro R, Grossman JW. Famous trails to Paul Erdős. The Mathematical Intelligencer. 1999;21(3):51–53.

[43] Nascimento MA, Sander J, Pound J. Analysis of SIGMOD's co-authorship graph. ACM SIGMOD Record. 2003;32(3):8–10.

[44] Liu X, Bollen J, Nelson ML, Van de Sompel H. Co-authorship networks in the digital library research community. Information Processing & Management. 2005;41(6):1462–1480.

[45] Shortliffe EH, Patel VL, Cimino JJ, Barnett GO, Greenes RA. A study of collaboration among medical informatics research laboratories. Artificial intelligence in medicine. 1998;12(2):97–123. Available from: http://bmir.stanford.edu/file_asset/index.php/270/BMIR-1997-0685.pdf.

[46] Heinze T, Kuhlmann S. Across institutional boundaries?: Research collaboration in German public sector nanoscience. Research Policy. 2008;37(5):888–899. Available from: http://www.sciencedirect.com/science/article/pii/S0048733308000255.

[47] Gomez I, Fernandez MT, Sebastian J. Analysis of the structure of international scientific cooperation networks through bibliometric indicators. Scientometrics. 1999;44(3):441–457. Available from: `http://link.springer.com/article/10.1007/BF02458489`.

[48] Börner K, Dall'Asta L, Ke W, Vespignani A. Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams. Complexity. 2005;10(4):57–67. Available from: `http://onlinelibrary.wiley.com/doi/10.1002/cplx.20078/full`.

[49] Wagner CS, Leydesdorff L. Network structure, self-organization, and the growth of international collaboration in science. Research policy. 2005;34(10):1608–1618. Available from: `http://www.sciencedirect.com/science/article/pii/S0048733305001745`.

[50] Lundberg J, Tomson G, Lundkvist I, Sk? r J, Brommels M. Collaboration uncovered: Exploring the adequacy of measuring university-industry collaboration through co-authorship and funding. Scientometrics. 2006;69(3):575–589. Available from: `http://www.akademiai.com/index/71251183574hjq55.pdf`.

[51] Newman MEJ. Coauthorship networks and patterns of scientific collaboration. Proceedings of the National Academy of Sciences of the United States of America. 2004;101(Suppl 1):5200.

[52] Newman ME. Clustering and preferential attachment in growing networks. Physical Review E. 2001;64(2):025102. Available from: `http://pre.aps.org/abstract/PRE/v64/i2/e025102`.

[53] Newman ME. Assortative mixing in networks. Physical review letters. 2002;89(20):208701. Available from: `http://prl.aps.org/abstract/PRL/v89/i20/e208701`.

[54] Newman MEJ. The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences. 2001;98(2):404–409. Available from: http://www.pnas.org/content/98/2/404.short.

[55] Palla G, Barabási AL, Vicsek T. Quantifying social group evolution. Nature. 2007;446(7136):664–667. Available from: http://www.nature.com/nature/journal/v446/n7136/abs/nature05670.html.

[56] Schwikowski B, Uetz P, Fields S. A network of protein–protein interactions in yeast. Nature biotechnology. 2000;18(12):1257–1261. 01082. Available from: http://www.nature.com/nbt/journal/v18/n12/abs/nbt1200_1257.html.

[57] Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic acids research. 2011;39(suppl 1):D561–D568. 00908. Available from: http://nar.oxfordjournals.org/content/39/suppl_1/D561.short.

[58] Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic acids research. 2006;34(suppl 1):D535–D539. 01220. Available from: http://nar.oxfordjournals.org/content/34/suppl_1/D535.short.

[59] Chatr-aryamontri A, Breitkreutz BJ, Heinicke S, Boucher L, Winter A, Stark C, et al. The BioGRID interaction database: 2013 update. Nucleic acids research. 2013;41(D1):D816–D823. 00066. Available from: http://nar.oxfordjournals.org/content/41/D1/D816.short.

[60] Jeong H, Mason SP, Barabási AL, Oltvai ZN. Lethality and centrality in protein networks. Nature. 2001;411(6833):41–42. Available from: http://www.nature.com/nature/journal/v411/n6833/abs/411041a0.html.

[61] Barabási AL, Albert R. Emergence of scaling in random networks. science. 1999;286(5439):509–512. Available from: http://www.sciencemag.org/content/286/5439/509.short.

[62] Watts DJ, Strogatz SH. Collective dynamics of 'small-world'networks. nature. 1998;393(6684):440–442. Available from: http://www.nature.com/nature/journal/v393/n6684/abs/393440a0.html.

[63] Kleinberg J. The small-world phenomenon: an algorithm perspective. In: Proceedings of the thirty-second annual ACM symposium on Theory of computing; 2000. p. 163–170. Available from: http://dl.acm.org/citation.cfm?id=335325.

[64] Page L, Brin S, Motwani R, Winograd T. The pagerank citation ranking: Bringing order to the web. Technical report,Stanford University. 1998;.

[65] Kleinberg JM. Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM). 1999;46(5):604 – 632.

[66] Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. Journal of the American society for information science and technology. 2007;58(7):1019–1031. Available from: http://onlinelibrary.wiley.com/doi/10.1002/asi.20591/full.

[67] Agarwal S, Choubey L, Yu H. Automatically Classifying the Role of Citations in Biomedical Articles. In: AMIA Annual Symposium Proceedings. vol. 2010; 2010. p. 11.

[68] Zhang Q, Yu H. CiteGraph: A Citation Network System for MEDLINE Articles and Analysis. In: Studies in Health Technology and Informatics. vol. 192; 2013. p. 832–836.

[69] Small H. Visualizing science by citation mapping. Journal of the American society for Information Science. 1999;50(9):799–813.

[70] Greenberg S. How citation distortions create unfounded authority: analysis of a citation network. BMJ: British Medical Journal. 2009;339.

[71] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine* 1. Computer networks and ISDN systems. 1998;30(1-7):107–117.

[72] Lempel R, Moran S. The stochastic approach for link-structure analysis (SALSA) and the TKC effect1. Computer Networks. 2000;33(1-6):387–401.

[73] Ng AY, Zheng AX, Jordan MI. Link analysis, eigenvectors and stability. In: International Joint Conference on Artificial Intelligence. vol. 17; 2001. p. 903–910.

[74] Ng AY, Zheng AX, Jordan MI. Stable algorithms for link analysis. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval; 2001. p. 258–266.

[75] Google Images;. Available from: `http://images.google.com/`.

[76] Apache Hadoop (R);. Available from: `http://hadoop.apache.org/`.

[77] Torvik VI, Smalheiser NR. Author name disambiguation in MEDLINE. ACM Transactions on Knowledge Discovery from Data (TKDD). 2009;3(3):11. Available from: `http://dl.acm.org/citation.cfm?id=1552304`.

[78] Aiello W, Chung F, Lu L. A random graph model for massive graphs. In: Proceedings of the thirty-second annual ACM symposium on Theory of computing; 2000. p. 171–180. Available from: `http://dl.acm.org/citation.cfm?id=335326`.

[79] Kleinberg J, Kumar R, Raghavan P, Rajagopalan S, Tomkins A. The web as a graph: Measurements, models, and methods. Computing and Combinatorics. 1999;p. 1–17. Available from: http://www.springerlink.com/index/9nh2gmpxq7kx1bd8.pdf.

[80] White HD, Griffith BC. Author cocitation: A literature measure of intellectual structure. Journal of the American Society for Information Science. 1981;32(3):163–171.

[81] Hersh WR, Cohen AM, Roberts PM, Rekapalli HK. TREC 2006 Genomics Track Overview. In: TREC; 2006. 00376. Available from: http://skynet.ohsu.edu/~hersh/trec-06-genomics.pdf.

[82] Garfield E, Merton RK. Citation indexing: Its theory and application in science, technology, and humanities. vol. 8. Wiley New York; 1979. 01901. Available from: http://www.garfield.library.upenn.edu/cifwd.html.

[83] Tbahriti I, Chichester C, Lisacek F, Ruch P. Using argumentation to retrieve articles with similar citations: an inquiry into improving related articles search in the MEDLINE digital library. International Journal of Medical Informatics. 2006;75(6):488–495. 00039. Available from: http://www.sciencedirect.com/science/article/pii/S1386505605000894.

[84] Voos H, Dagaev KS. Are All Citations Equal? Or, Did We Op. Cit. Your Idem?. Journal of Academic Librarianship. 1976;1(6):19–21. 00039. Available from: http://www.eric.ed.gov/ERICWebPortal/recordDetail?accno=EJ131338.

[85] Herlach G. Can retrieval of information from citation indexes be simplified? Multiple mention of a reference as a characteristic of the link between cited and citing article. Journal of the American Society for Information Science.

1978;29(6):308–310. 00032. Available from: http://onlinelibrary.wiley.com/doi/10.1002/asi.4630290608/abstract.

[86] Garfield E. Can citation indexing be automated. In: Statistical association methods for mechanized documentation, symposium proceedings; 1965. p. 189–192. Available from: http://books.google.com/books?hl=en&lr=&id=r56ZrbfTdkYC&oi=fnd&pg=PA189&dq=Can+citation+indexing+be+automated&ots=2uYvwcXb7z&sig=2uFKuIEauU0pqYPLekHhLS_MCp4.

[87] Schwartz AS, Hearst M. Summarizing key concepts using citation sentences. In: Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis; 2006. p. 134–135. Available from: http://dl.acm.org/citation.cfm?id=1567650.

[88] Börner K, Maru JT, Goldstone RL. The simultaneous evolution of author and paper networks. Proceedings of the National Academy of Sciences of the United States of America. 2004;101(Suppl 1):5266–5273. 00194. Available from: http://www.pnas.org/content/101/suppl.1/5266.short.

[89] Barabási AL, Jeong H, Néda Z, Ravasz E, Schubert A, Vicsek T. Evolution of the social network of scientific collaborations. Physica A: Statistical Mechanics and its Applications. 2002;311(3):590–614. Available from: http://www.sciencedirect.com/science/article/pii/S0378437102007367.

[90] Zhang Q, Cao YG, Yu H. Parsing citations in biomedical articles using conditional random fields. Computers in biology and medicine. 2011;41(4):190–194. 00002. Available from: http://www.sciencedirect.com/science/article/pii/S0010482511000291.

[91] HighWire;. Available from: http://highwire.stanford.edu/.

[92] Giles CL, Bollacker KD, Lawrence S. CiteSeer: An automatic citation indexing system. In: Proceedings of the third ACM conference on Digital libraries; 1998. p. 89–98.

[93] Lawrence S, Lee Giles C, Bollacker K. Digital libraries and autonomous citation indexing. Computer. 1999;32(6):67–71. 00795. Available from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=769447.

[94] Wei W, King I, Lee J. Bibliographic attributes extraction with layer-upon-layer tagging. In: Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on. vol. 2; 2007. p. 804–808. 00008. Available from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4377026.

[95] Besagni D, Belaïd A. Citation recognition for scientific publications in digital libraries. In: Document Image Analysis for Libraries, 2004. Proceedings. First International Workshop on; 2004. p. 244–252. 00021. Available from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1263253.

[96] Powley B, Dale R. Evidence-based information extraction for high accuracy citation and author name identification. In: Large Scale Semantic Access to Content (Text, Image, Video, and Sound); 2007. p. 618–632. 00021. Available from: http://dl.acm.org/citation.cfm?id=1931448.

[97] Pasula H, Marthi B, Milch B, Russell S, Shpitser I. Identity uncertainty and citation matching. In: Advances in neural information processing systems; 2002. p. 1401–1408. 00281. Available from: http://machinelearning.wustl.edu/mlpapers/paper_files/AP01.pdf.

[98] Takasu A. Bibliographic attribute extraction from erroneous references based on a statistical model. In: Digital Libraries, 2003. Proceedings. 2003 Joint

Conference on; 2003. p. 49–60. 00083. Available from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1204843.

[99] Chen CC, Yang KH, Kao HY, Ho JM. BibPro: A citation parser based on sequence alignment techniques. In: Advanced Information Networking and Applications-Workshops, 2008. AINAW 2008. 22nd International Conference on; 2008. p. 1175–1180. 00016. Available from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4483078.

[100] Kramer M, Kaprykowsky H, Keysers D, Breuel T. Bibliographic meta-data extraction using probabilistic finite state transducers. In: Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on. vol. 2; 2007. p. 609–613. 00014. Available from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4376987.

[101] Peng F, McCallum A. Information extraction from research papers using conditional random fields. Information Processing & Management. 2006;42(4):963–979. 00368. Available from: http://www.sciencedirect.com/science/article/pii/S0306457305001172.

[102] Councill IG, Giles CL, Kan MY. ParsCit: an Open-source CRF Reference String Parsing Package. In: LREC; 2008. 00107. Available from: http://svn.tribler.org/abc/branches/leo/p2p-search/lib/parscit-080917/doc/lrec08/lrec08.pdf.

[103] Lafferty J, McCallum A, Pereira FC. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001;06247. Available from: http://repository.upenn.edu/cis_papers/159/.

[104] Wellner B, McCallum A, Peng F, Hay M. An integrated, conditional model of information extraction and coreference with application to citation matching.

In: Proceedings of the 20th conference on Uncertainty in artificial intelligence; 2004. p. 593–601. 00127. Available from: http://dl.acm.org/citation.cfm?id=1036915.

[105] Settles B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics. 2005;21(14):3191–3192. 00271. Available from: http://bioinformatics.oxfordjournals.org/content/21/14/3191.short.

[106] MALLET: A Machine Learning for Language Toolkit;. 00930. Available from: http://mallet.cs.umass.edu.

[107] Language-Independent Named Entity Recognition (I);. Available from: http://www.cnts.ua.ac.be/conll2002/ner/.

[108] Luo J, Flynn JM, Solnick RE, Ecklund EH, Matthews KR. International stem cell collaboration: How disparate policies between the United States and the United Kingdom impact research. PLoS One. 2011;6(3):e17684. Available from: http://dx.plos.org/10.1371/journal.pone.0017684.

[109] Jones BF, Wuchty S, Uzzi B. Multi-university research teams: shifting impact, geography, and stratification in science. science. 2008;322(5905):1259–1262. Available from: http://www.sciencemag.org/content/322/5905/1259.short.

[110] Okeke IN, Wain J. Post-genomic challenges for collaborative research in infectious diseases. Nature Reviews Microbiology. 2008;6(11):858–864. Available from: http://www.nature.com/nrmicro/journal/vaop/ncurrent/full/nrmicro1989.html.

[111] Gray B. Enhancing transdisciplinary research through collaborative leader-

ship. American journal of preventive medicine. 2008;35(2 Suppl):S124. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2542584/.

[112] Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, et al. Life with 6000 genes. Science. 1996;274(5287):546–567. 03228. Available from: http://www.sciencemag.org/content/274/5287/546.short.

[113] Harvard Catalyst;. Available from: http://catalyst.harvard.edu/.

[114] SciVal Experts;. Available from: http://info.scival.com/experts.

[115] COS Pivot: the next generation of funding and research expertise—connected.;. Available from: http://www.refworks-cos.com/pivot/.

[116] VIVO;. Available from: http://vivoweb.org/.

[117] BiomedExperts;. Available from: http://www.biomedexperts.com/.

[118] Yu H, Agarwal S, Johnston M, Cohen A. Are figure legends sufficient? Evaluating the contribution of associated text to biomedical figure comprehension. Journal of Biomedical Discovery and Collaboration. 2009;4(1):1. Available from: http://archive.biomedcentral.com/1747-5333/4/1/abstract.

[119] Al Hasan M, Chaoji V, Salem S, Zaki M. Link prediction using supervised learning. In: SDM'06: Workshop on Link Analysis, Counter-terrorism and Security; 2006. .

[120] Backstrom L, Leskovec J. Supervised random walks: predicting and recommending links in social networks. In: Proceedings of the fourth ACM international conference on Web search and data mining; 2011. p. 635–644.

[121] Zhang Q, Yu H. Computational Approaches for Predicting Biomedical Research Collaborations. PLoS ONE. 2014;Acccepted.

[122] Sun Y, Barber R, Gupta M, Aggarwal CC, Han J. Co-author relationship prediction in heterogeneous bibliographic networks. In: Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on; 2011. p. 121–128. 00044. Available from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5992571.

[123] Chellappa R, Jain A. Markov random fields. Theory and application. Boston: Academic Press, 1993, edited by Chellappa, Rama; Jain, Anil. 1993;1. Available from: http://adsabs.harvard.edu/abs/1993mrft.book.....C.

[124] Richardson M, Domingos P. Markov logic networks. Machine learning. 2006;62(1):107–136. Available from: http://www.springerlink.com/index/W55P98P426L6405Q.pdf.

[125] Getoor L. Link mining: a new data mining challenge. ACM SIGKDD Explorations Newsletter. 2003;5(1):84–89. Available from: http://dl.acm.org/citation.cfm?id=959242.959253.

[126] Huang J, Zhuang Z, Li J, Giles CL. Collaboration over time: characterizing and modeling network evolution. In: Proceedings of the international conference on Web search and web data mining; 2008. p. 107–116. Available from: http://dl.acm.org/citation.cfm?id=1341548.

[127] Wang C, Satuluri V, Parthasarathy S. Local probabilistic models for link prediction. In: Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on; 2007. p. 322–331. 00097. Available from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4470256.

[128] Adamic LA, Huberman BA. Power-law distribution of the world wide web. Science. 2000;287(5461):2115–2115. Available from: http://www.sciencemag.org/content/287/5461/2115.short.

[129] Kleinberg JM, et al. Navigation in a small world. Nature. 2000;406(6798):845–845. Available from: http://web.thu.edu.tw/wang/www/Small_World/Smallworld02.pdf.

[130] Newman ME. Mixing patterns in networks. Physical Review E. 2003;67(2):026126. Available from: http://pre.aps.org/abstract/PRE/v67/i2/e026126.

[131] Ding Y. Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. Journal of informetrics. 2011;5(1):187–203. Available from: http://www.sciencedirect.com/science/article/pii/S1751157710000957.

[132] Lee K, Brownstein JS, Mills RG, Kohane IS. Does collocation inform the impact of collaboration? PloS one. 2010;5(12):e14279. Available from: http://dx.plos.org/10.1371/journal.pone.0014279.

[133] Pan RK, Kaski K, Fortunato S. World citation and collaboration networks: uncovering the role of geography in science. Scientific reports. 2012;2. Available from: http://www.nature.com/srep/2012/121129/srep00902/full/srep00902.html.

[134] Choi BC, Pak AW. Multidisciplinarity, interdisciplinarity, and transdisciplinarity in health research, services, education and policy: 2. Promotors, barriers, and strategies of enhancement. Clinical & Investigative Medicine. 2007;30(6):E224–E232. Available from: http://cimonline.ca/index.php/cim/article/viewArticle/2950.

[135] Gerard S, Michael JM. Introduction to modern information retrieval. McGraw-Hill; 1983.

[136] Adamic LA, Adar E. Friends and neighbors on the web. Social networks. 2003;25(3):211–230. Available from: http://www.sciencedirect.com/science/article/pii/S0378873303000091.

[137] Han J, Kamber M. Data mining: concepts and techniques. Morgan Kaufmann; 2006. Available from: http://books.google.com/books?hl=en&lr=&id=AfL0t-YzOrEC&oi=fnd&pg=PP2&dq=Data+Mining:+Concepts+and+Techniques,+Second+Edition&ots=UvYSwQ6nE3&sig=eWhqRaPt5Es96vFKixlbCQT6A2M.

[138] Witten IH, Frank E. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann; 2005. Available from: http://books.google.com/books?hl=en&lr=&id=QTnOcZJzlUoC&oi=fnd&pg=PR17&dq=data+mining+in+action&ots=3gmz9mVeMg&sig=bxwM1vDzZUKQJWzsfnMoMAUl8OY.

[139] McCallum A, Nigam K. A comparison of event models for naive bayes text classification. In: AAAI-98 workshop on learning for text categorization. vol. 752; 1998. p. 41–48. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.65.9324&rep=rep1&type=pdf.

[140] Vapnik V. The nature of statistical learning theory. springer; 1999. Available from: http://books.google.com/books?hl=en&lr=&id=sna9BaxVbj8C&oi=fnd&pg=PR7&dq=the+nature+of+statistical+machine+learning&ots=onM8MWhmff&sig=NwZyE1zPVekRBxXy-aY2Zdc4pA0.

[141] Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd international conference on Machine learning; 2006. p. 161–168. Available from: http://dl.acm.org/citation.cfm?id=1143865.

[142] Huang J, Lu J, Ling CX. Comparing naive Bayes, decision trees, and SVM

with AUC and accuracy. In: Data Mining, 2003. ICDM 2003. Third IEEE International Conference on; 2003. p. 553–556. 00081. Available from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1250975.

[143] Weka3: Data Mining Software In Java;. Available from: http://www.cs.waikato.ac.nz/ml/weka/.

[144] LIBSVM;. 15323. Available from: http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[145] Scikit;. 00333. Available from: http://scikit-learn.org/stable/.

[146] Ahmed NK, Neville J, Kompella R. Network sampling: from static to streaming graphs. arXiv preprint arXiv:12113412. 2012;00002. Available from: http://arxiv.org/abs/1211.3412.

[147] Hastie T, Tibshirani R, Friedman JJH. The elements of statistical learning. vol. 1. Springer New York; 2001. 00415. Available from: http://www-stat.stanford.edu/~tibs/book/preface.ps.

[148] Yang Y, Pedersen JO. A comparative study on feature selection in text categorization. In: ICML. vol. 97; 1997. p. 412–420. 04189. Available from: http://faculty.cs.byu.edu/~ringger/Winter2007-CS601R-2/papers/yang97comparative.pdf.

[149] Tylenda T, Angelova R, Bedathur S. Towards time-aware link prediction in evolving social networks. In: Proceedings of the 3rd Workshop on Social Network Mining and Analysis; 2009. p. 9. 00049. Available from: http://dl.acm.org/citation.cfm?id=1731020.

[150] O'Madadhain J, Hutchins J, Smyth P. Prediction and ranking algorithms for event-based network data. ACM SIGKDD Explorations Newsletter.

2005;7(2):23–30. 00113. Available from: http://dl.acm.org/citation.cfm?id=1117458.

[151] Huang Z, Lin DK. The time-series link prediction problem with applications in communication surveillance. INFORMS Journal on Computing. 2009;21(2):286–303. 00051. Available from: http://pubsonline.informs.org/doi/abs/10.1287/ijoc.1080.0292.

[152] Apache Lucene - Apache Solr;. Available from: http://lucene.apache.org/solr/.

[153] Díaz-Uriarte R, De Andres SA. Gene selection and classification of microarray data using random forest. BMC bioinformatics. 2006;7(1):3. Available from: http://www.biomedcentral.com/1471-2105/7/3.

[154] Jiang P, Wu H, Wang W, Ma W, Sun X, Lu Z. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic acids research. 2007;35(suppl 2):W339–W344. Available from: http://nar.oxfordjournals.org/content/35/suppl_2/W339.short.

[155] Liaw A, Wiener M. Classification and Regression by randomForest. R news. 2002;2(3):18–22. Available from: ftp://131.252.97.79/Transfer/Treg/WFRE_Articles/Liaw_02_Classification%20and%20regression%20by%20randomForest.pdf.

[156] Jarvenpaa SL, Leidner DE. Communication and trust in global virtual teams. Journal of Computer-Mediated Communication. 1998;3(4):0–0. Available from: http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.1998.tb00080.x/full.

[157] Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D. DIP: the database of interacting proteins. Nucleic acids research. 2000;28(1):289–291.

00617. Available from: http://nar.oxfordjournals.org/content/28/1/289.short.

[158] Lim J, Hao T, Shaw C, Patel AJ, Szabó G, Rual JF, et al. A protein–protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. Cell. 2006;125(4):801–814. 00484. Available from: http://www.sciencedirect.com/science/article/pii/S0092867406004399.

[159] Barabási AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nature Reviews Genetics. 2004;5(2):101–113. 03779. Available from: http://www.nature.com/nrg/journal/v5/n2/abs/nrg1272.html.

[160] Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL. The human disease network. Proceedings of the National Academy of Sciences. 2007;104(21):8685–8690. Available from: http://www.pnas.org/content/104/21/8685.short.

[161] Chen J, Aronow BJ, Jegga AG. Disease candidate gene identification and prioritization using protein interaction networks. BMC bioinformatics. 2009;10(1):73. Available from: http://www.biomedcentral.com/1471-2105/10/73.

[162] Gandhi TKB, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, et al. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. Nature genetics. 2006;38(3):285–293. Available from: http://www.nature.com/ng/journal/v38/n3/abs/ng1747.html.

[163] Xu J, Li Y. Discovering disease-genes by topological features in human protein–protein interaction network. Bioinformatics. 2006;22(22):2800–2805. Available from: http://bioinformatics.oxfordjournals.org/content/22/22/2800.short.

[164] Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic acids research. 2002;30(1):303–305. 01121. Available from: http://nar.oxfordjournals.org/content/30/1/303.short.

[165] Von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, et al. Comparative assessment of large-scale data sets of protein–protein interactions. Nature. 2002;417(6887):399–403. 01909. Available from: http://www.nature.com/nature/journal/v417/n6887/abs/nature750.html.

[166] Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, et al. Towards a proteome-scale map of the human protein–protein interaction network. Nature. 2005;437(7062):1173–1178. 01758. Available from: http://www.nature.com/nature/journal/vaop/ncurrent/full/nature04209.html.

[167] Rhodes DR, Tomlins SA, Varambally S, Mahavisno V, Barrette T, Kalyana-Sundaram S, et al. Probabilistic model of the human protein-protein interaction network. Nature biotechnology. 2005;23(8):951–959. 00339. Available from: http://www.nature.com/nbt/journal/v23/n8/abs/nbt1103.html.

[168] Yu H, Braun P, Yıldırım MA, Lemmens I, Venkatesan K, Sahalie J, et al. High-quality binary protein interaction map of the yeast interactome network. Science. 2008;322(5898):104–110. 00689. Available from: http://www.sciencemag.org/content/322/5898/104.short.

[169] Jones S, Thornton JM. Prediction of protein-protein interaction sites using patch analysis. Journal of molecular biology. 1997;272(1):133–143. 00385. Available from: http://www.sciencedirect.com/science/article/pii/S002228369791233X.

[170] Zhong W, Sternberg PW. Genome-wide prediction of C. elegans genetic in-

teractions. Science. 2006;311(5766):1481–1484. 00194. Available from: http://www.sciencemag.org/content/311/5766/1481.short.

[171] Lee I, Li Z, Marcotte EM. An improved, bias-reduced probabilistic functional gene network of baker's yeast, Saccharomyces cerevisiae. PloS one. 2007;2(10):e988. 00135. Available from: http://dx.plos.org/10.1371/journal.pone.0000988.

[172] Lee I, Date SV, Adai AT, Marcotte EM. A probabilistic functional network of yeast genes. science. 2004;306(5701):1555–1558. 00590. Available from: http://www.sciencemag.org/content/306/5701/1555.short.

[173] Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science. 2003;302(5644):449–453. 00955. Available from: http://www.sciencemag.org/content/302/5644/449.short.

[174] Albert I, Albert R. Conserved network motifs allow protein–protein interaction prediction. Bioinformatics. 2004;20(18):3346–3352. 00109. Available from: http://bioinformatics.oxfordjournals.org/content/20/18/3346.short.

[175] Tong AHY, Drees B, Nardelli G, Bader GD, Brannetti B, Castagnoli L, et al. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science. 2002;295(5553):321–324. 00622. Available from: http://www.sciencemag.org/content/295/5553/321.short.

[176] Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. Nature.

1999;402(6757):86–90. 01049. Available from: http://www.nature.com/nature/journal/v402/n6757/abs/402086a0.html.

[177] Bader JS, Chaudhuri A, Rothberg JM, Chant J. Gaining confidence in high-throughput protein interaction networks. Nature biotechnology. 2003;22(1):78–85. 00420. Available from: http://www.nature.com/nbt/journal/v22/n1/abs/nbt924.html.

[178] Tsuda K, Shin H, Schölkopf B. Fast protein classification with multiple networks. Bioinformatics. 2005;21(suppl 2):ii59–ii65. 00125. Available from: http://bioinformatics.oxfordjournals.org/content/21/suppl_2/ii59.short.

[179] Bock JR, Gough DA. Predicting protein–protein interactions from primary structure. Bioinformatics. 2001;17(5):455–460. 00467. Available from: http://bioinformatics.oxfordjournals.org/content/17/5/455.short.

[180] Aloy P, Russell RB. InterPreTS: protein interaction prediction through tertiary structure. Bioinformatics. 2003;19(1):161–162. 00168. Available from: http://bioinformatics.oxfordjournals.org/content/19/1/161.short.

[181] Samanta MP, Liang S. Predicting protein functions from redundancies in large-scale protein interaction networks. Proceedings of the National Academy of Sciences. 2003;100(22):12579–12583. 00261. Available from: http://www.pnas.org/content/100/22/12579.short.

[182] Qi Y, Bar-Joseph Z, Klein-Seetharaman J. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. Proteins: Structure, Function, and Bioinformatics. 2006;63(3):490–500. 00221. Available from: http://onlinelibrary.wiley.com/doi/10.1002/prot.20865/full.

[183] Qi Y, Klein-Seetharaman J, Bar-Joseph Z, Qi Y, Bar-joseph Z. Random Forest Similarity for Protein-Protein Interaction Prediction. In: Pac Symp Biocomput; 2005. 00128. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.83.1779.

[184] Vert JP, Kanehisa M. Graph-driven feature extraction from microarray data using diffusion kernels and kernel CCA. In: Advances in neural information processing systems; 2002. p. 1425–1432. 00083. Available from: http://machinelearning.wustl.edu/mlpapers/paper_files/AP04.pdf.

[185] Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, et al. A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. Nature. 2000;403(6770):623–627. 04446. Available from: http://www.nature.com/nature/journal/v403/n6770/abs/403623a0.html.

[186] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proceedings of the National Academy of Sciences. 2001;98(8):4569–4574. 03210. Available from: http://www.pnas.org/content/98/8/4569.short.

[187] Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. Molecular systems biology. 2007;3(1). 00540. Available from: http://www.nature.com/msb/journal/v3/n1/full/msb4100129.html.

[188] Chua HN, Sung WK, Wong L. Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions. Bioinformatics. 2006;22(13):1623–1630. 00309. Available from: http://bioinformatics.oxfordjournals.org/content/22/13/1623.short.

[189] Huynen MA, Snel B, Mering Cv, Bork P. Function prediction and protein networks. Current opinion in cell biology. 2003;15(2):191–198.

00151. Available from: http://www.sciencedirect.com/science/article/pii/S0955067403000097.

[190] King AD, Pr\vzulj N, Jurisica I. Protein complex prediction via cost-based clustering. Bioinformatics. 2004;20(17):3013–3020. 00396. Available from: http://bioinformatics.oxfordjournals.org/content/20/17/3013.short.

[191] Brun C, Chevenet F, Martin D, Wojcik J, Guénoche A, Jacq B. Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. Genome biology. 2004;5(1):R6–R6. 00227. Available from: http://www.biomedcentral.com/content/pdf/gb-2003-5-1-r6.pdf.

Qing Zhang

Santa Clara, CA , qing@uwm.edu , July 2014

EDUCATION

PhD, Computer Science
University of Wisconsin Milwaukee, Milwaukee, WI
Minor: Mathematics
Expected to graduate in August 2014

MS, Computer Science, University of Wisconsin Milwaukee, Milwaukee, WI, 2011
MEng, Software Engineering, Beihang University, Beijing, China, 2007
BEng, Mechanical Engineering, Beihang University, Beijing, China, 2004

RESEARCH INTEREST

Text mining, social network analysis, machine learning, information retrieval, natural language processing, hypothesis generation

TECHNICAL SKILLS

- Languages: Java, Python, R, Matlab
- Tools: Weka, Scikit, Lucene, Hadoop, Mellet, Solr, Relational databases
- Machine learning models/algorithms such as Nave Bayes, Support Vector Machine, Logistic Regression and Conditional Random Fields
- Information retrieval techniques including indexing, searching, result ranking

EMPLOYMENT

Applied Researcher and Machine Learning Engineer, eBay Inc.
San Jose, CA — Mar 2014 - Present
Structured data team, Marketplaces. Classification and catalog.

Research Assistant, University of Massachusetts Medical School
Worcester, MA — Feb 2013 - Feb 2014
1) Use machine learning approaches to model scientific collaboration network in biomedical field and recommend potential collaborators for researchers. 2) Analyze large-scale protein-protein interaction networks (4.5 million predicted interactions) and use machine learning model to generate new hypothesis based on the network structure and existing literature.

Intern, DataXu
Boston MA — June 2012- Aug 2012
Optimization team. Apply machine learning for online advertising. Train classification models on extremely unbalanced data set to predict the likelihood of the online audience of buying a certain product.

Intern, NextBio
Santa Clara CA — Jul 2011 - Dec 2011
Member of text analysis team. 1) Named entity extraction from medical literature. Improve the disambiguation result. 2) Worked on different types of indexing, including medical abstract (20 million MEDLINE), full text, and auto-completion index. 3) Search quality improvement, facet search, filter search.

Research Assistant, University of Wisconsin Milwaukee
Milwaukee, WI — Sep 2008 - Jan 2014
Large scaled citation network analysis (4 million articles) for article/author impact

ranking, clinical document classification with various machine learning models, citation parsing with Conditional Random Fields

Software Engineer, QualityEasy Inc.
Beijing, China Sep 2007 - Jul 2008
Developed an Eclipse plugin for automatic web application testing.

Intern, IBM China Development Lab
Beijing, China Jun 2006-Nov 2006
Domain analysis team which focused on developing tools for Service Oriented Domain Analysis. Worked on user interface development and system testing.

PUBLICATIONS

Qing Zhang, Hong Yu. Computational Approaches for Predicting Biomedical Research Collaborations. *PLOS ONE.* 2014. Accepted.

Qing Zhang, Hong Yu. CiteGraph: A Citation Network System for PubMed Articles and Analysis. *Studies in Health Technology and Informatics.* 2013 192, 823-836.

Qing Zhang, Hong Yu, AACRank - Ranking Author and Article in a Citation Network. In *Proceedings of 3rd Annual International SciTS Conference.* 2012.

Qing Zhang,Yong-gang Cao,Hong Yu. Parsing citations in biomedical articles using conditional random fields. *Computers of Biology and Medicine.* 2011.

Yong-gang Cao, Zuofeng Li, Feifan Liu, Shashank Agarwal, **Qing Zhang**, Hong Yu, An IR-Aided Machine Learning Framework for the BioCreative II.5 Challenge. *IEEE/ACM Transactions on Computational Biology and Bioinformatics.* 2010 7(3): 454 -461.

SOFTWARE

Jude
A software package for extracting citation information from PDF publications, and parse citations into specific fields such as author and title using Conditional Random Fields model.
URL http://code.google.com/p/jude/

HONORS

Chancellor's Graduate Student Award, University of Wisconsin Milwaukee 2014, 2013 and 2009