UTILIZAÇÃO DE SEMÂNTICA DAS RELAÇÕES PARA RECOMENDAR COLABORAÇÕES EM REDES SOCIAIS ACADÊMICAS

MICHELE AMARAL BRANDÃO

UTILIZAÇÃO DE SEMÂNTICA DAS RELAÇÕES PARA RECOMENDAR COLABORAÇÕES EM REDES SOCIAIS ACADÊMICAS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais - Departamento de Ciência da Computação. como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Mirella Moura Moro

Belo Horizonte
Março de 2013

MICHELE AMARAL BRANDÃO

USING LINK SEMANTICS TO RECOMMEND COLLABORATIONS IN ACADEMIC SOCIAL NETWORKS

Dissertation presented to the Graduate Program in Ciência da Computação of the Universidade Federal de Minas Gerais - Departamento de Ciência da Computação. in partial fulfillment of the requirements for the degree of Master in Ciência da Computação.

Advisor: Mirella Moura Moro

Belo Horizonte
March 2013

Amaral Brandão, Michele
B817u Using link semantics to recommend collaborations in
academic social networks / Michele Amaral Brandão.
— Belo Horizonte, 2013
xxii, 58 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de
Minas Gerais - Departamento de Ciência da
Computação.
Orientadora: Mirella Moura Moro

1. Computação - Teses. 2. Redes de relações sociais -
Teses. 3. Sistemas de recomendação - Teses.
I. Orientador. II. Título.

519.6*73(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Utilização de semântica das relações para recomendar colaborações em redes sociais acadêmicas (Using link semantics to recommend collaborations in academic social networks

MICHELE AMARAL BRANDÃO

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROFA. MIRELLA MOURA MORO - Orientadora
Departamento de Ciência da Computação - UFMG

PROF. ALBERTO HENRIQUE FRADE LAENDER
Departamento de Ciência da Computação - UFMG

PROF. CLODOVEU AUGUSTO DAVIS JÚNIOR
Departamento de Ciência da Computação - UFMG

PROF. JOSÉ PALAZZO MOREIRA DE OLIVEIRA
Instituto de Informática - UFRGS

Belo Horizonte, 15 de março de 2013.

Acknowledgments

My warm thanks to everyone who encouraged me and contributed in some way to the development of this work. Specially, I would like to thank:

- My advisor Mirella M. Moro for her teachings, disposition and friendship. Her guidance and knowledge were very essential;
- Professor Clodoveu A. Davis Jr., Dr. Giseli R. Lopes (from PUC-Rio) and Professor José Palazzo M. de Oliveira (from UFRGS) for their willingness to clarify doubts and give suggestions;
- The faculty at UFMG and UESC;
- My colleagues and friends of LBD. In particular, Daniel Hasan, Michele Brito, Péterson Procópio and Eduardo Barbosa for their friendship and all the help in solving doubts and problems;
- My parents, Ivan and Margarete, for all their support (which was immense and if I describe here, probably, I will forget some), love and care;
- My fiancé Rafick for his company, affection, respect, friendship, love and all the support;
- My family and dear friends, specially, my brother Tauan, my grandmother Janete, Delza, Arielle, Jorge, Tânia, Lívia, Iolanda and Leonardo for the affection, encouragement and friendship;
- My dear friends of the Master degree, Mônica, Ana Paula and Simone for the company, friendship and discussions that helped me to clarify doubts (especially, in the first semester that was very difficult for me);
- To God, my fountain of faith, ligth and hope;
- CAPES, CNPq, Fapemig and InWeb (Brazil) for partially funding this work.

"The task is not so much to see what no one yet has seen, but to think what nobody yet has thought about that which everybody sees."

(Arthur Schopenhauer)

Resumo

A análise de redes sociais tem sido realizada em muitos contextos com diferentes objetivos. Neste trabalho, usamos conceitos desse tipo de análise para recomendar colaborações em redes sociais acadêmicas. Como um trabalho recente mostra que grupos de pesquisa com uma rede acadêmica bem conectada tendem a ser mais produtivos, recomendar colaborações é essencial para aumentar as conexões em um grupo, e como consequência, aumentar a produtividade do grupo de pesquisa. Assim, propomos duas métricas e verificamos como elas influenciam nas recomendações de novas colaborações ou intensificação das já existentes. Cada métrica considera um princípio social (homofilia e proximidade) que é relevante no contexto acadêmico. Outro problema relevante é como analisar a qualidade das recomendações resultantes. Dessa forma, também propomos novos algoritmos para avaliar as recomendações com base em conceitos sociais (novidade, diversidade e cobertura) que nunca foram utilizados para tal objetivo. Nossa avaliação experimental em dados reais mostra que nossas novas métricas melhoram a qualidade das recomendações quando comparadas ao estado-da-arte. Finalmente, analisamos as propriedades das redes sociais utilizadas na experimentação. Isso contribui para entender os resultados das métricas de recomendação.

Palavras-chave: redes sociais, sistemas de recomendação, predição de links.

Abstract

Social network analysis has been explored in many contexts with different goals. In this work, we use concepts from such analysis for recommending collaborations in academic networks. As a recent work shows that research groups with well connected academic networks tend to be more prolific, recommending collaborations is essential for increasing a group's connections, then boosting the group research as a collateral advantage. Therefore, we propose two metrics and verify how they influence in recommendation of new collaborations or intensification of existing ones. Each metric considers a social principle (homophily and proximity) that is relevant within the academic context. Another relevant problem is how to analyze the quality of the resulting recommendations. Hence, we also propose new algorithms for evaluating recommendations based on social concepts (novelty, diversity and coverage) that have never been used for such a goal. Overall, our experimental evaluation on real datasets shows that using our new metrics improves the quality of the recommendations when compared to the state-of-the-art. Finally, we analyze the properties of the academic networks used in the experimentation. The analysis contributes to understand the results of the recommendation metrics.

Keywords: social networks, recommender system, link prediction.

List of Figures

List of Tables

Contents

List of Acronyms

Acronym	*Description*
CC	*Correlation Coefficient*
CORALS	*Collaboration Recommendation on Academic Social Networks*
Cp	*Cooperation*
Cr	*Correlation*
DBLP	*Digital Bibliography & Library Project*
GLI	*Geographic Location Information*
Sc	*Social Closeness*
SN	*Social Network*
SNA	*Social Network Analysis*
VSM	*Vector Space Model*

Chapter 1

Introduction

The complex networks area (includes neural networks, information networks, **social networks**, among others) is very important for modern science, revealing fundamental and still unknown aspects about the world [Figueiredo, 2011]. Despite some studies about specific complex networks and random network models (for example, Erdös-Rényi model [Erdös and Rényi, 1959] and the Milgram experiment [Milgram, 1967]), it was only in the last decade that researchers began to further study complex networks, their structures and impacts.

A social network (SN) is a collection of individuals (or organizations) that have relationships in a certain context, for example, friendship, politics and co-authorship. Social networks have been studied for over two decades in order to analyze the interactions between people and detect patterns in such interactions [Barabasi, 2002]. Many researchers have realized that the network perspective allows new leverage for answering standard social and behavioral science questions, by giving precise formal definition to aspects of the social structural environment [Wasserman and Faust, 1994].

Indeed, Social Network Analysis (SNA) includes patterns and principles that are defined by social theories, such as homophily, proximity, contagion, exchange, etc. [Contractor et al., 2006]. These principles apply to different links and connections, including marriage, friendship, work, among many others. Usually, the social network of each individual is then homogeneous considering socio-demographic, behavioral and intra-personal characteristics. Nonetheless, patterns, principles and models provided by SNA can assist in exploring and predicting the individuals' behavior.

In this context, many methods have been proposed for various aspects of SNA, including community detection [Qi et al., 2012], viral marketing [Subramani and Rajagopalan, 2003] and link prediction [Liben-Nowell and Kleinberg, 2003]. If we consider the analysis of *online* social networks, security aspects such as privacy and trust may also be explored [Yuan et al., 2010]. In such an online context, link prediction may also be mapped to link recommendation; so, instead of *inferring* future connections, it also allows to *suggest* new ones [Symeonidis et al., 2010]. Note that recommendation systems are well known for their personalized suggestions of items to users based on profiles, previous behavior and collective information [Shani and Gunawardana, 2011].

Among all types of social network, our focus is on those where social links are given by academic ties. For example, an academic tie exists between advisor and advisee [Wang et al., 2010], people from the same research group [Lopes et al., 2011] and co-authors [Laender et al., 2011; Tang et al., 2008]. Within those, *co-authorship social networks* are formed by researchers and their connections given by publication and patent collaborations. These co-authorship networks are valid proxies for collaboration because authorship sharing reflects a tangible engagement [Adams, 2012].

In this research-oriented world, recommending or predicting new links may help a researcher to form new groups or teams, to search for collaborations when writing a grant proposal and to investigate different research communities. Also, a recent work shows that research groups with a well connected co-authorship social network tend to be more prolific [Lopes et al., 2010]. Moreover, collaboration is normally a good thing from a wider public perspective [Adams, 2012].

However, discovering new links in this scenario is not a trivial task. As pointed out by Lopes et al. [2010], when recommending new friendships in a traditional social network, the number of common friends can be used to estimate the social proximity between users. On the other hand, in the academic context, social proximity has different interpretations, in which the social connection between people and their academic background (e.g., institutional affiliation, geographic location and research area) must be considered. Specifically, we are interested in discovering how the institutional affiliation and the geographic location of the researchers (link semantics in the SN) increase the quality of the recommendations and influence in the collaboration. Each of these link semantics considers different interpretations of the relation between researchers and institutions, and can also be used alone or in combination with each other for recommending to initiate and intensify collaborations.

Having a list of recommendations, another problem is how to evaluate them. Although there are common metrics for evaluating recommendations (such as precision and recall), they practically do not explore any particular feature of the social network. Therefore, we employ SNA-based concepts for evaluating the recommendations from the social network perspective (which makes sense because the recommendations were defined from the social perspective as well).

1.1 Main Contributions

Overall, the contributions of this work on recommending collaborations based on academic social networks are summarized as follows:

- An analysis of different approaches for social network studies, link prediction and recommendation of collaboration, as well as discussion on how social theories influence such approaches;

- A formal definition for the recommendation of collaborations;

- The definition of a metric, called *Affin*, and a recommendation function that combines two measures (institutional affiliation and social closeness) to recommend new collaborations considering link semantics (this function is also used to recommend collaboration intensification);

- The definition of a second metric, called *GLI* (Geographic Location Information), and a recommendation function that uses the geographic location (a link semantics) to recommend new collaborations and collaborations that can be intensified;

- The utilization of three metrics (*novelty*, *diversity* and *coverage*) for analyzing the quality of the recommendations. This is the first time that those metrics are used for evaluating recommendation of collaborations. Therefore, we also introduce one new algorithm for each of them;

- An experimental evaluation using two real SN and a comparison to the state-of-the-art in recommending academic collaborations. We also combine our metrics to the state-of-the-art to increase the accuracy of its recommendations;

- An analysis of the academic social networks used in the experimentation and discussion on how the SN properties influence the new recommendation metrics.

1.2 Text Organization

This work is organized as follows. Chapter 2 overviews related work and different social network properties. Chapter 3 presentes our new two metrics (*Affin* and *GLI*) to recommend collaborations. Chapter 4 describes the evaluation measures of the recommendations and their algorithms. Chapter 5 goes over our experimental evaluation and our main results. Finally, Chapter 6 concludes this work, reviewing our main contributions and presenting future work.

Chapter 2

Literature Review

In this section, we discuss related work grouped in the following categories: social networks, complex networks theory, link semantics, social principles, academic social networks, link prediction and recommendation definition. We then combine those concepts and discuss recommendation in SN and social principles on link prediction. Finally, we emphasize the main contributions of our work in relation to state-of-the-art.

2.1 Individual Concepts

2.1.1 Social Networks

Any society can be mapped to a social network; such a network can be analyzed to find out all types of information, from how a disease has spread and major political views to who the new criminals are. A social network is formally defined as a $graph(N, r)$, where N is the set of nodes (or vertices) representing individuals (persons, organizations, countries, etc), and r is the set of edges (or links) representing their relationships, given by an $n \times n$ matrix in which $r_{i,j}$ is the (weighted or not, directed or not) relation between nodes i and j [Barabasi, 2002; Newman, 2003; Wasserman and Faust, 1994]. For example, Figure 2.1 shows a friendship social network represented by a graph, where the nodes are people and the links represent the friendship between these people.

Social networks is a very prolific research area. Indeed, looking for "social network" on DBLP[1] returned over 5,600 entries in January 2013. In order to exemplify what those publications are about, we have chosen three aspects of social networks. First, its usage: social networks have been successfully employed for viral marketing, which has moved to the technology era through emails [Subramani and Rajagopalan,

[1] DBLP: `http://www.dblp.org/db/`

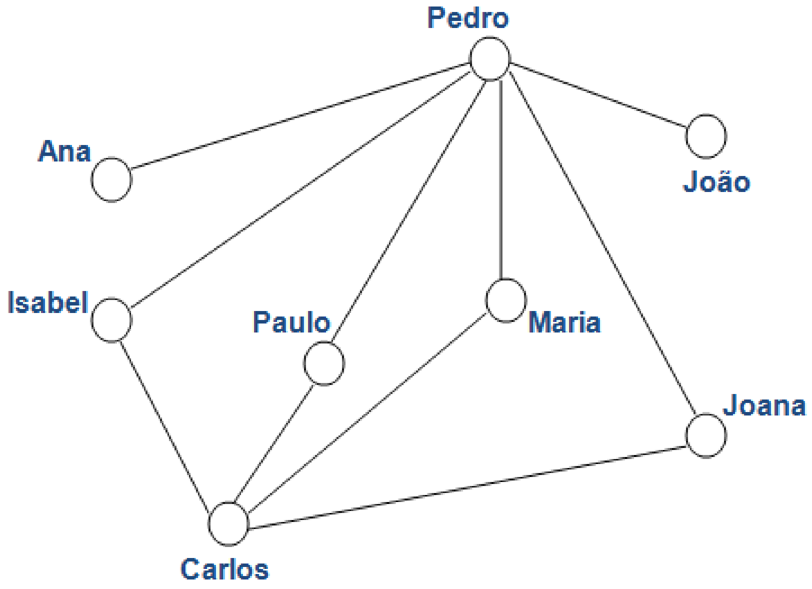


Figure 2.1. Example of a friendship social network.

2003] and has soon conquered the social media as well [Bigonha et al., 2011]. Second, its analysis: social networks have been created and analyzed for automatically discovering whole communities, which otherwise would be hidden [Qi et al., 2012]. Third, its "dark side": online social networks have been overly exploited for their priceless, private information; which in turn motivates a lot of research on privacy and security [Akcora et al, 2012; Yuan et al., 2010].

2.1.2 Complex Networks Theory

As presented in Chapter 1, social networks are a kind of complex network. Thus, concepts on complex networks can be applied in SN. Among those, we cite the following.

Density. It measures how close the network is to be complete. A complete network (like a complete graph) has all possible edges and density equal to one. For example, this measure has been applied to evaluate knowledge-sharing in social networks [Wiemken et al., 2012], to assess quality of graduate programs [Lopes et al., 2011] and to study communities structures [Newman, 2003].

Connected Components. It determines the number of subgroups not connected to any other subgroup in the network. The analysis of connected components has been used, for example, to detect communities [Levorato and Petermann, 2011], to evaluate knowledge-sharing in social networks [Wiemken et al., 2012] and to analyze online social networks [Mislove et al., 2007].

Average Clustering Coefficient. The clustering coefficient indicates how the nodes are inserted in their neighborhood. The average value provides a general indication of network clustering. This metric has been used to analyze sampled networks [Lee et al., 2006], to predict links in social networks [Huang, 2006; Liben-Nowell and Kleinberg, 2003] and to study collaboration networks [Newman, 2001].

Diameter. It is applied to measure the longest distance between any two nodes in the network, i.e, how separate the two farthest nodes are. Researchers have used the diameter to investigate different properties such as the structures of multiple online social networks [Mislove et al., 2007], topology structures of the Web [Albert et al., 1999] and design an optimal double-loop network [Qiong-Fang et al., 2010].

Average Degree. The degree of a node is the number of edges associated with it, and the average degree is the average of the degrees of all nodes in the network. This measure has been used to, for example, perform decentralized search in networks [Wu et al., 2011], resolve paths in complex networks [Li and Chen, 2009] and redistribute traffic load [Yiran and Wenwen, 2012].

Average Path length. It measures the average distance between all pairs of nodes in the network. In general, the average path length is used with other measures, such as the average clustering coefficient and degree [Lee et al., 2006], diameter and clustering coefficient [Mislove et al., 2007], diameter and average degree [M'Chirgui, 2010].

Small-world. A network that has small-world properties exhibits a small diameter and a high clustering coefficient. The small-world phenomenon was examined in offline social networks [Kleinberg, 2000] and online social networks [Mislove et al., 2007]. Studies have shown that small-world properties are everywhere: the Web [Albert et al., 1999], large-scale topology structure of the Internet [M'Chirgui, 2010], friendship [Backstrom et al., 2012; Ugander et al., 2011], the biggest Chinese language software technical forum [Yan and Assimakopoulos, 2007], scientific collaboration network of the biomedical, physics and computer science research [Newman, 2001], and general social networks [Adamic et al., 2003].

Random Networks. Complex data can often be represented in terms of random graphs or networks [Banyai et al., 2009]. These networks are usually constructed by randomly adding links to a static set of nodes. Studies about random networks started with Erdös and Réyni's work [Erdös and Rényi, 1959]. Researchers have shown that a general random graph is considered to be a small-world if its average path length and average degree are the same order as that of an Erdös-Réyni graph, but with a much larger clustering coefficient [Yagan and Makowski, 2009]. In general, random networks have been used as benchmarks. For example, in [Watts and Strogatz, 1998], the authors explored how to verify whether networks have small-world properties by comparing them to random networks.

2.1.3 Considering Link Semantics for Social Networks Analysis

In general, analyzing interactions or patterns in the SN requires only the topological characteristics [Huang, 2006; Liben-Nowell and Kleinberg, 2003; Pizzato and Silvestrini, 2011]. However, other authors show the importance of considering the link semantics [Lopes et al., 2010; Quercia and Capra, 2009; Aiello et al., 2012]. Specifically, two different characteristics that can be extracted from the link semantics in the SN are the affiliation and geographic proximity information.

Affiliation. Being part of a particular company, organization or neighborhood; frequenting a particular place; or pursuing a particular hobby or interest - these are all activities that, when shared between two people, tend to increase the likelihood that they will interact and hence form a link in the SN [David and Jon, 2010]. The association of a person with any such activity is called *affiliation*. The affiliation can be represented by a $graph(N, X, r)$: there is a node representing each person and each activity, and a person N_i is connected to activity X_i by an edge r_i if N_i participates in X_i [David and Jon, 2010]. Such a graph is referred to as an *affiliation network*.

The affiliation networks have been used to identify groups in their real state and virtual groups within an organization [Kitahara and Yoshikai, 2010], to model the dynamic behavior of an actor in the SN considering the concept of loyalty [Sharara et al., 2009], to analyze the effect of affiliation networks on creating innovative ideas and forming a technological position [Weng et al., 2010], and so on.

Geographic Proximity. Proximity theories argue that people communicate most frequently with those to whom they are physically close [Monge and Contractor, 2003]. Hence, many methods have used this characteristic for SN analysis. For example, Kaltenbrunner et al. [2012] show that online social interactions are weakly affected by geographic distance and, once social connections are established, other factors may influence how users send messages to their friends. Brown et al. [2012] propose a way to extract place-focused communities from the social graph by annotating edges with check-in information. The potential implications is that better services and applications can be designed by focusing on users who frequently visit the same physical places.

Likewise, Volkovich et al. [2012] show that social connections between users inside the core of the SN tend to have shorter geographic spans than connections stretching outside the core. Geographic closeness not only increases the likelihood of connections, but also increases the likelihood that users belong to the same, tightly connected group of individuals. Instead, social ties outside the core tend to be much longer than the other links: the length of these bridge ties is thus creating not only network shortcuts, but also spatial shortcuts. The role of these spatially long bridges is crucial to spread information over the network and, at the same time, over space. Finally, Yu et al. [2011] recommended geographically related friends in social network combining GPS information and SN structures.

Analyzing these results, an important conclusion is that the geographic proximity alone does not influence social interactions, but it is very essential for the formation of new relationships. Note that in this context, geographic proximity refers to people who visit the same place at least once. This principle was investigated for message exchange [Kaltenbrunner et al., 2012] and friendship interactions [Liben-Nowell et al., 2005], but no work studies it for collaborations between researchers.

2.1.4 Social Principles

SNA includes principles defined by social theories such as homophily, proximity, contagion and exchange [Contractor et al., 2006]. Specifically, the homophily principle postulates that people tend to form links with other people who have similar characteristics (i.e. the tendency of *like* to associate with *like*) [David and Jon, 2010; McPherson et al., 2001]. Another social theory that influences the relationships between people is the proximity principle [Contractor et al., 2006], which affirms that proximity (physical or electronic) facilitates the likelihood of communication by increasing the probability that individuals will meet and interact [Monge and Contractor, 2003]. When such interactions occur, they allow individuals to know each other, discover common interests and share beliefs [Homans, 1950]. Likewise, individuals who are not proximate are deprived of the opportunity to explore these common interests and are, hence, less likely to a initiate relationship (new link in the network) [Monge and Contractor, 2003].

2.1.5 Academic Social Networks

Several research communities have used SNA to understand their own characteristics and behavior [Ding, 2011]. Examples include the communities of physics [Newman, 2001], mathematics [Barabasi, 2002] and digital libraries [Liu et al., 2005]. From those, a distinct type of social network has emerged: the *academic social networks*, in which social ties are given by research or academic collaboration. For instance, an academic tie exists between advisor and advisee [Wang et al., 2010], people from the same research group [Lopes et al., 2011] and coauthors [Laender et al., 2011; Tang et al., 2008].

Specifically, co-authorship networks are an important SN class and have been explored under different points of view. For instance, the study presented in [Ding, 2011] shows that prolific researchers usually collaborate with others who share common research interests. The study also analyzes citation patterns and shows that highly cited researchers do not usually collaborate. Likewise, other studies try to explore and visualize different co-authorship networks [Ganev et al., 2010; Laender et al., 2011; Tang et al., 2008]. Finally, it is also possible to rank graduate programs by analyzing their co-authorship networks [Lopes et al., 2011].

2.1.6 Link Prediction

Given a set of individuals organized in a social network, the link prediction problem infers which new connections are likely to occur in the near future. The term link prediction was coined by Liben-Nowell and Kleinberg [2003], a study on evaluating topological measures (e.g., a Jaccard coefficient) for classifying co-author collaborations. Likewise, Huang [2006] uses a topological measure to describe the occurrence of links. It is also possible to predict friendship relations in social media [Aiello et al., 2012]. Link prediction may also be mapped to recommending new connections [Lopes et al., 2010], which is discussed in Section 2.1.7.

2.1.7 Recommendation Definition

Recommending products (books, movies, music, hotels) to users by capturing the item-to-item and user-to-user similarity measures are tasks of traditional recommender systems (for example, Amazon, Netflix, Ringo) [Kutty et al., 2012]. The aim is to recommend items that match the preferences (likes or deslikes) of users [Cai et al., 2010].

Hence, as defined by Lopes et al. [2010], given a set of users (clients, customers) U and a set of items I (e.g., books, movies, music), a *typical recommendation method* has a recommendation function $f(u, i)$ that associates (u, i) pairs to application-oriented values (e.g., distance, profit, rating). The goal of these approaches is to find a set of items $i' \in I$ that maximize $f(u, i)$ for a user.

Another dimension of recommender systems is people-to-people recommendation. Recommending people connections has different challenges when compared to typical recommender systems [Guy et al., 2009], such as: in general, accepting a recommendation to connect with other people is less time consuming than following a recommendation to watch a movie and may thus be easier to attract to; it requires sending an invitation to another person whose reaction is unknown in advance; and the fact that the connection is typically exposed to the public may have social implications. Thus, as opposed to the traditional recommender system that considers relationships as means to provide better item recommendations, people recommendation uses relationships to recommend the related people themselves [Guy et al., 2009].

2.2 Combined Concepts

2.2.1 Recommendation in Social Networks

Existing recommendation approaches recommend items (e.g., music, hotel, club) and people (e.g. being friends, co-worker, lovers) to users in different settings as e-commerce websites, online dating, social networks, employment websites. Specifically, the interest of this work is in the social network setting.

The approaches presented by Freyne et al. [2010] and He and Chu [2010] recommend items based on information extracted from social networks. The difference between them is that the former recommends items considering the interactions of the individual with the SN, and the latter makes recommendations based on user's own preferences, the acceptance of the target item and the opinions from social friends.

Regarding people recommendation, Yang et al. [2012] propose a set of algorithms to infer circles of friends in online social networks. Likewise, Symeonidis et al. [2010] present a node similarity measure and an algorithm to recommend friends in SN. Guy et al. [2009] describe a novel system for providing users with recommendations of people to invite into their explicit enterprise SN. Finally, Lopes et al. [2010] present a new methodology for recommending collaborations in academic social networks. These approaches are related for making people-to-people recommendation, but the work of Lopes et al. [2010] differs from others due to the kind of relationship recommended.

The considerations to recommend friends are different from recommending people to work with. For example, Cai et al. [2010] recommend users to others based on similarity measures as *taste* (whom they like) and *attractiveness* (who likes them). However, this form such measuring similarity cannot be applied in the academic setting, because it is not possible to infer if a researcher likes (or not) another.

2.2.2 Social Principles in Link Prediction

Different social principles may influence on predicting links. For instance, recent studies show that the homophily principle can improve link prediction models [Aiello et al., 2012; Wang et al., 2011]. Aiello et al. [2012] developed an unsupervised model to estimate the strength of links based on users similarity and interaction activity. Wang et al. [2011] have explored many measures considering the homophily principle in human mobility to predict links. Others, like Quercia and Capra [2009] and Wang et al. [2011], predict new links in a social network considering both the homophily and the proximity principles (both use mobile phones to capture user trajectories). On the other hand, no work that uses the proximity principle for predicting links in an *academic* social network has been found so far.

2.2.3 Collaboration Recommendation in Social Networks

Collaboration recommendation is a specific recommendation problem in which two individuals are recommended to work together. In order to achieve relevant recommendations, it is necessary to consider aspects that influence collaboration relationships. For example, in CORALS (*Collaboration Recommendation on Academic Social Networks*) Lopes et al. [2010], a weight represents each relation between researchers and is defined for the measures: *cooperation* (Cp, how much the two researchers have collaborated), *correlation* (Cr, how similar the areas of the researchers are) and *social closeness* (Sc, a normalized variant of the shortest path metric). Cr e Sc are combined to form a single, weighted average measure. Furthermore, the *cooperation* between authors a and b is a value in the range [0,1] defined by the ratio of the number of papers that a has co-authored with b by the total number of a's papers. The *correlation* is defined by an equation that considers the researchers publications area and the vector space model (VSM) to compute the values between each pair of co-authors in the network.

2.3 Discussion on Contributions

In summary, social networks have been extensively applied by focusing on different types of SN including academic social networks, which is our context. Specifically, this work aims at recommending new links (or intensifying existing ones) between researchers in an academic SN using two new metrics based on homophily and proximity, given by affiliation and geographic location information.

The new metrics follow theoretical mechanisms (homophily and proximity principles) that have been used to explain the creation, maintenance, dissolution and reconstitution of social networks [Contractor et al., 2006; David and Jon, 2010; Homans, 1950; McPherson et al., 2001; Monge and Contractor, 2003]. These metrics explore weights and how different features on the SN (e.g., links semantics) affect the relationship between researchers. Determining such weights is a great challenge, because they should be closely related to the researchers profile, the type of data and the network model. In other words, recommending collaborations differs significantly from recommending items (e.g., Amazon, Netflix, Ringo). Indeed, people-to-people recommendation must consider different aspects from the social connections as well [Guy et al., 2009; Lopes et al., 2010; Symeonidis et al., 2010].

Finally, the work more related to ours is CORALS [Lopes et al., 2010], whose emphasis is also on recommending collaborations in academic social networks. Our work differs from CORALS for considering social theories in the definition of the collaboration weights: *(i)* the homophily principle, given by institutional affiliation of the researchers in *Affin*; and *(ii)* the proximity principle, represented by geographic location information on the researchers institutions in *GLI*. Furthermore, the experimental evaluation of CORALS considers only the accuracy of the recommendations. On the other hand, here we also define measures to evaluate accuracy, *novelty*, *diversity* and *coverage* of the recommendations generated by *Affin* and *GLI*. Such new ways of evaluating the recommendation results provide new insights on the quality of the recommendations. Moreover, whereas CORALS employs one dataset in the experimental evaluations, here we use two real datasets.

Chapter 3

Recommending Collaborations Using Link Semantics

This work aims to recommend collaborations by predicting links between researchers using two new metrics, *Affin* and *GLI*. Specifically, the metrics explore link semantics (affiliation and geographic location) to recommend collaborations in an academic (co-authorship) SN: *Affin* follows the homophily principle and considers that researchers collaborate with researchers from institutions with which they have already collaborated; and *GLI* follows the proximity principle and considers that researchers collaborate with researchers who are physically nearby.

3.1 Framework Description

Social Networks are formed by *actors* (people) and their *relational ties* (links) [Newman, 2003]. The importance of a relationship between its actors may be defined by a weight measure. Each weight is relevant because it reflects the link semantics, instead of just the network topological feature; i.e., the weight semantics provides rich information from the SN and its connections. We use an academic SN in which two researchers (actors in the network) are connected if they have co-authored a publication [Newman, 2003]. Although we focus on publication coauthorship, our metrics can be easily extended to work on similar relationships such as writing patents, editing books, proceedings and so on. The final goal is to recommend collaborations (new or intensification) over this network, which is mapped to predicting links in a SN.

Let $\mathcal{T}$ be a set of *target researchers* (i.e., the researchers that are going to receive the recommendations) and $\mathcal{R}$ the universe of researchers that will be evaluated (i.e., the researchers considered for the recommendation). Given a graph built with all

Figure 3.1. Framework to recommend collaborations.

researchers ($\mathcal{R} \cup \mathcal{T}$) and their connections (defined by their co-authorships), in which each link is associated to a set of weight values $\mathcal{W}$ that represents the semantics of the network (e.g., cooperation, affiliation). The weights are combined to form a metric $\mathcal{M}$, which is then employed by a recommendation function. The recommendation function $f(\mathcal{T}, \mathcal{R}, \mathcal{M})$ evaluates the two input sets according to the metric and returns a ranked list of recommended pairs $\langle t, r\rangle$ that maximizes the value of f.

Figure 3.1 shows the framework for generating the recommendations. First, a SN is built from the existing datasets. The link semantics of the social relations define the weights. Then, the weights may be further elaborated for defining the metric that composes the recommendation function. The recommendation function returns the ranked pairs of researchers. The hardest part of defining a recommendation function is choosing a proper metric. Next, we present both new metrics followed by a example.

3.2 Affin - Affiliation Metric

Affin is a metric that considers the homophily principle for recommending collaborations. In the academic context, this principle could be explored with different meanings, such as people who interested in the same research area, attending the same conferences and working in the same place. However, our interest is studying how the institutional affiliation increases the quality of the recommendations and influences in the collaboration. In this work, the homophily principle is derived from the institutional affiliation of the researchers and defined by the *affiliation weight* $Affin_{i,j}$, which represents the link semantics for any given pair of researchers $\langle i, j\rangle$ according to Equation 3.1,

$$Affin_{i,j} = \frac{NPI_{i,j}}{NT_i} \tag{3.1}$$

where $NPI_{i,j}$ is the number of publications of researcher i co-authored with people from j's institution, and NT_i is the total number of publications authored by i. *Affin* follows the natural intuition that an institution is more important to an author, if he has already collaborated with someone from that institution; hence, it is more likely to contact other researcher in the same institution.

However, recommending based solely on the researchers' affiliations is not enough, because it disregards the history of the researchers' collaborations. Therefore, we propose to combine it with existing metrics, in order to improve the recommendation function, as explained next.

Combining Affiliation and Cooperation. The first way of using affiliation is with cooperation. The *affiliation weight* $\mathit{Affin}_{i,j}$ is combined with the *cooperation weight* $(Cp_{i,j})$. Note that the cooperation value $Cp_{i,j}$ determines how much a pair of researchers has already collaborated (or not). The final goal is to have a recommendation function that is able to consider both affiliation and cooperation, in order to provide a better result and improve the overall connection of the academic social network.

In order to equally consider $\mathit{Affin}_{i,j}$ and $Cp_{i,j}$, *Affin* uses degrees to represent ranges of values: "high", "medium" and "low". The actual values for the ranges may follow a linear scale (e.g., *low* < 33% and *high* > 66%). Equation 3.2 shows the *recommendation function* that combines them and returns two recommended actions: "*Initiate_ Collaboration*" and "*Intensify_ Collaboration*",

$$r_{i,j} = \begin{cases} Initiate\_Collaboration, & \text{if } (Cp_{i,j} = 0) \wedge \\ & (\mathit{Affin}_{i,j} > \mathit{threshold}); \\ Intensify\_Collaboration, & \text{if } (Cp_{i,j} \in \{low, medium\}) \wedge \\ & (\mathit{Affin}_{i,j} \in \{medium, high\}); \end{cases} \tag{3.2}$$

where pairs of researchers with zero $Cp_{i,j}$ and non-zero $\mathit{Affin}_{i,j}$ (we choose "low" degree as threshold) are recommended to initiate collaborations; and pairs with "low" or "medium" $Cp_{i,j}$ and "medium" or "high" $\mathit{Affin}_{i,j}$ are recommended to intensify their collaborations.

Combining Affiliation, Cooperation, Social Closeness and Correlation. A better way of considering the affiliation aspect is combined with cooperation, social closeness and correlation aspects [Lopes et al., 2010]. This combination allows to consider different characteristics between researchers in the recommendation function. Following Lopes et al. [2010] that combines correlation and social closeness, we combine $\mathit{Affin}_{i,j}$ and $Sc_{i,j}$ to establish a single weight $\mathit{Affin\_Sc}_{i,j}$ defined by Equation 3.3,

$$\mathit{Affin\_Sc}_{i,j} = \frac{w_{Affin}.Affin_{i,j} + w_{Sc}.Sc_{i,j}}{w_{Affin} + w_{Sc}} \tag{3.3}$$

where given a network with authors i and j, $\mathit{Affin\_Sc}_{i,j}$ is a weighted average, w_{Affin} and w_{Sc} weights determine, respectively, the importance of $\mathit{Affin}_{i,j}$ and $Sc_{i,j}$ to the resulting value. Hence, the weights may be used for emphasizing either the affiliation or the social closeness; i.e., allowing to emphasize the homophily in different ways.

For each pair of researchers, the relationship among $\mathit{Affin}_{i,j}$, $Sc_{i,j}$, $Cp_{i,j}$ and $Cr_{i,j}$

determines if the collaboration should (or not) be initiated or intensified between them. Finally, Equation 3.4 shows the *recommendation function* with the weights (that define each metric) and recommendation actions:

$$r_{i,j} = \begin{cases} Initiate\_Collaboration, & \text{if } (Cp_{i,j} = 0) \wedge \\ & (\mathit{Affin\_Sc}_{i,j} > \mathit{threshold}); \\ Intensify\_Collaboration, & \text{if } (Cp_{i,j} \in \{low, medium\}) \wedge \\ & (\mathit{Affin}_{i,j} \in \{medium, high\}) \wedge \\ & (Cr_{i,j} \in \{medium, high\}); \end{cases} \tag{3.4}$$

where a pair of researchers with zero $Cp_{i,j}$ and non-zero $\mathit{Affin\_Sc}_{i,j}$ (we choose "low" degree as threshold) are recommended to create a collaboration; and pairs with "low" or "medium" $Cp_{i,j}$, "medium" or "high" $\mathit{Affin}_{i,j}$, and "medium" or "high" $Cr_{i,j}$ are recommended to intensify their existing collaborations.

It is important to notice that Equation 3.2 is a straightforward use of affiliation, whereas Equation 3.4 gives a more complete usage (because it considers more characteristics from the researchers relationship). We have performed a prior experimental evaluation comparing the use of both equations and the results showed that Equation 3.4 provides better results. Therefore, from now on, *Affin* refers to Equation 3.4.

After describing the recommendation actions, we now define a score to allow a final ranking of recommendations. Equation 3.5 shows that if the recommended action is to "Initiate_Collaboration", the recommendation score is equal to $\mathit{Affin\_Sc}_{i,j}$ and the recommended researchers are in *descending* order of this weight. If the recommendation action is to "Intensify_Collaboration", the ratio of cooperation and correlation is used and the recommended researchers are in *increasing* order of this ratio.

$$score_{i,j} = \begin{cases} \mathit{Affin\_Sc}_{i,j}, & \text{if } (r_{i,j} = Initiate\_Collaboration); \\ \frac{Cp_{i,j}}{Cr_{i,j}}, & \text{if } (r_{i,j} = Intensify\_Collaboration); \end{cases} \tag{3.5}$$

Finally, it is also important to notice that *Affin* is more complete than its predecessor CORALS, because it regards the homophily principle. Moreover, having an institution-oriented weight provides more information to the SNA, such as assisting in the search for collaborations with different institutions and analyzing the influence of the cooperation with an institution upon the collaborations.

3.3 GLI - Geographic Location Information

Based on the first law of geography, according to which "everything is related to everything else, but near things are more related than distant things" [Tobler, 1970], a new metric (*GLI*) that considers the geographic location is presented in this section. The *GLI* metric allows to study how the physical distance may increase the quality of the recommendations and influence the collaborations.

The *GLI* metric follows the proximity principle. The theoretical mechanisms of this principle (that considers the influence of distance in the relationships) can be captured in the SN's relational ties (links). In order to measure the physical proximity between pairs of researchers, we introduce the *geographic location weight* $GLI_{i,j}$ that considers the geographic location information for any given pair of researchers $\langle i, j\rangle$ defined by Equation 3.6

$$GLI_{i,j} = distance(GC_i, GC_j) \tag{3.6}$$

where GC_i and GC_j represent the geographical coordinates of the researchers i and j institutions, respectively, and *distance* is a selected function to compute the distance between locations.

In this work, we use geographic coordinates of the city in which the researcher's institution is located. The data with geographic location of the institutions was gathered from Wikimapia[1] and stored in a PostgreSQL[2] database. This DBMS (Database Management System) was chosen because it has an open source spatial database extension called PostGIS[3]. This extension provides operators and functions to manipulate geographic data. *ST_Distance* is one of these functions that returns spheroidal minimum distance between two geographies in meters, it was selected to approximate the distance between researchers.

In order to define a qualitative scale, we are interested in the *travel time* that covers the distance (represented by $GLI_{i,j}$) between researchers. It allows to specify how far two researchers are from to each other. Thus, given a pair of researchers $\langle i, j\rangle$, the *travel time* is defined by Equation 3.7,

$$\begin{cases} \mathit{if}\ GLI_{i,j} < \mathit{190\ km}, & \Delta t_{i,j} = \frac{GLI_{i,j}}{80(Km/h)} \\ \\ \mathit{else}, & \Delta t_{i,j} = \frac{GLI_{i,j}}{500(Km/h)} + 2h \end{cases} \tag{3.7}$$

[1] Wikimapia: `http://wikimapia.org`

[2] PostgreSQL: `http://www.postgresql.org`

[3] PostGIS: `http://www.postgis.org`

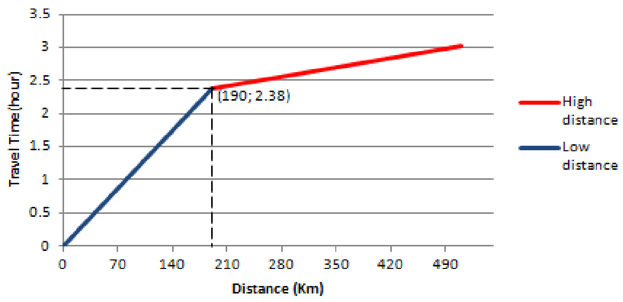


Figure 3.2. Distance versus Travel Time.

where $\Delta t_{i,j}$ represents the *travel time weight*.

Equation 3.7 was defined considering that people do not usually fly when the distance is less than $190Km$ (because it is very short). Using land transportation, the speed is approximately $80Km/h$, which indicates a travel time of approximately 2 hours. For longer distances, greater than or equal to $190Km$, air transportation is a better option because of the reduced travel time. Moreover, flying $500Km$ takes approximately 1 hour, plus 1 hour to arrive and to leave the airports, the travel time would be 3 hours or less.

Figure 3.2 shows that the intersection between high and low distance equations is $190Km$ and 2.38 hours. Hence, we define that researchers are near when *travel time* is less than 2.5 hours, and far from each other when *travel time* is greater than or equal to 2.5 hours. This defines a qualitative scale: "near" < 2.5 and "far" ≥ 2.5.

In order to recommend collaborations considering the geographic location information, Equation 3.8 shows the *recommendation function* that combines $\Delta t_{i,j}$, $Cp_{i,j}$ and $Cr_{i,j}$ and its recommended actions:

$$r_{i,j} = \begin{cases} Initiate\_Collaboration, & \text{if } (Cp_{i,j} = 0) \wedge \\ & (\Delta t_{i,j} \in \{near\}); \\ Intensify\_Collaboration, & \text{if } (Cp_{i,j} \in \{low, medium\}) \wedge \\ & (\Delta t_{i,j} \in \{near, far\}) \wedge \\ & (Cr_{i,j} \in \{medium, high\}); \end{cases} \tag{3.8}$$

where pairs of researchers with zero $Cp_{i,j}$ and "near" $\Delta t_{i,j}$ are recommended to create a collaboration; and pairs with "low" or "medium" $Cp_{i,j}$, "near" or "far" $\Delta t_{i,j}$, and "medium" or "high" $Cr_{i,j}$ are recommended to intensify it.

Finally, Equation 3.9 presents the calculation of the recommendation score. If the recommendation action is to "Initiate_Collaboration", the recommendation score is calculated by $\Delta t_{i,j}$ and the recommended researchers are in *increasing* order of this weight. If the recommendation action is to "Intensify_Collaboration", the ratio of cooperation and correlation is used and the recommended researchers are in *increasing* order of this ratio.

$$score_{i,j} = \begin{cases} \Delta t_{i,j}, & \text{if } (r_{i,j} = Initiate\_Collaboration); \\ \frac{Cp_{i,j}}{Cr_{i,j}}, & \text{if } (r_{i,j} = Intensify\_Collaboration); \end{cases} \tag{3.9}$$

Hence, the *GLI* metric recommends collaborations using $GLI_{i,j}$ and $\Delta t_{i,j}$ indexes that follow the proximity principle. Also, a geographic location information oriented-based weight gives rich information to SNA. Finally, the recommendation list is sorted according to a score for each action.

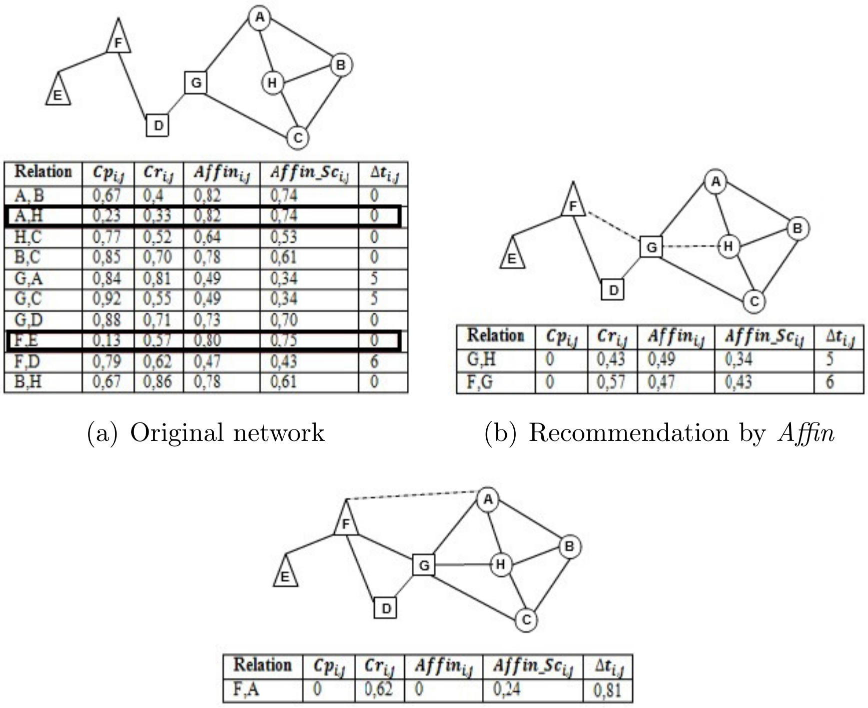
3.4 Example of using the Affin and GLI metrics

Figure 3.3 shows an example of the use of the new metrics. Consider the academic social network as in Figure 3.3(a), in which collaborations can be recommended to initiate or to intensify. In this SN, nodes with similar form belong to the same institution, and the weights of each relation are described in the table below. In order to simplify the explanation, the weights refer to only one direction, for example, the relation of A to B, but not B to A (depending of the direction, the weights may vary). The *Affin* and *GLI* metrics are then applied to make this SN more connected.

Figure 3.3(b) presents the recommendation generated by *Affin*. This metric considers not only the relation between pairs of researchers, but also the relation of each researcher with other researchers from the same institution with which the former has already collaborated. Thus, the pairs of researchers $\langle F, G \rangle$ and $\langle G, H \rangle$ are recommended to initiate collaboration, because there is no cooperation between them ($Cp_{i,j} = 0$) and $Affin\_Sc_{i,j}$ is greater than "low". In other words, F has collaborated with researchers from G's institution, and G with researchers from H's institution.

Likewise, Figure 3.3(c) shows the recommendation made by the *GLI* metric. In this case, the relation is established (or not) considering the physical distance and the travel time. Thus, the pair of researcher $\langle F, A \rangle$ is recommended to collaborate, because there is no cooperation between them ($Cp_{i,j} = 0$) and the weight ($\Delta t_{i,j}$) is "near".

Regarding the recommendation to intensify collaboration, the pairs of researchers $\langle A, H \rangle$ and $\langle F, E \rangle$ are recommended by *Affin* and *GLI*. Both relations have weights that satisfy the two metrics ($Affin_{i,j} \in \{medium, high\}$ and $\Delta t_{i,j} \in \{near, far\}$). Moreover, the researchers of the two pairs are correlated, i.e., they work in similar research areas.



Relation	$Cp_{i,j}$	$Cr_{i,j}$	$Affin_{i,j}$	$Affin\_Sc_{i,j}$	$\Delta t_{i,j}$
A,B	0,67	0,4	0,82	0,74	0
A,H	0,23	0,33	0,82	0,74	0
H,C	0,77	0,52	0,64	0,53	0
B,C	0,85	0,70	0,78	0,61	0
G,A	0,84	0,81	0,49	0,34	5
G,C	0,92	0,55	0,49	0,34	5
G,D	0,88	0,71	0,73	0,70	0
F,E	0,13	0,57	0,80	0,75	0
F,D	0,79	0,62	0,47	0,43	6
B,H	0,67	0,86	0,78	0,61	0

(a) Original network

Relation	$Cp_{i,j}$	$Cr_{i,j}$	$Affin_{i,j}$	$Affin\_Sc_{i,j}$	$\Delta t_{i,j}$
G,H	0	0,43	0,49	0,34	5
F,G	0	0,57	0,47	0,43	6

(b) Recommendation by *Affin*

Relation	$Cp_{i,j}$	$Cr_{i,j}$	$Affin_{i,j}$	$Affin\_Sc_{i,j}$	$\Delta t_{i,j}$
F,A	0	0,62	0	0,24	0,81

(c) Recommendation by *GLI*

Figure 3.3. Example using the *Affin* and *GLI* metrics.

3.5 Concluding Remarks

In this Chapter, we have formally defined the concept of collaboration recommendation and described the proposed framework for generating recommendations. Specifically, we showed the importance of the weights, metrics and recommendation functions.

Based on these concepts, we described two new metrics that consider the link semantics of the networks to recommend collaborations. These metrics follow social principles (homophily and proximity) and can also be used to SNA. Furthermore, in the formulation of *Affin* and *GLI*, we try to consider factors that influence in the cooperation between researchers. This is very important to have quality recommendations.

Regarding the recommendation scores, *Affin* and *GLI* order their recommendations to intensify collaboration using the proportion between cooperation and correlation. The proportion was proposed by Lopes et al. [2010], and our previous experiments showed that it provides better results in the ranking of the recommendations.

Chapter 4

Evaluation Metrics

Evaluating the quality of recommendations and the effectiveness of recommendation functions is a very difficult task, mainly for two reasons [Fouss and Saerens, 2008]: *(i)* different algorithms may have different performance on different datasets, and *(ii)* the goals for which an evaluation is performed may differ. Many studies focus on evaluating the accuracy of recommendations, such as [Lopes et al., 2010] and [Huang, 2006]. Having a high accuracy is important, but *insufficient* to ensure the quality of the recommendations [Fouss and Saerens, 2008; Shani and Gunawardana, 2011].

In order to define which metrics to use for evaluating recommendations, we have studied the compiled list presented by Shani and Gunawardana [2011]. Among several evaluation metrics we have concluded that precision, recall, novelty, diversity and coverage are more appropriate to evaluate the recommendation of collaborations. Metrics such as confidence, trust, utility and risk are not appropriate, because prior information about researchers' preferences is necessary and beyond our reach. The robustness metric is also not appropriate, because the datasets (used in the experimental evaluation and described in Chapter 5) do not present much noise.

It is important to note that in [Lopes et al., 2010], CORALS is evaluated in relation to the accuracy of the recommendations, but it does not represent their quality. Thus, this is the first time that *novelty*, *diversity* and *coverage* are used to evaluate the recommendation of academic collaborations. Next, we detail each metric and show how each of them is employed for evaluating the recommendation lists.

4.1 Accuracy

The accuracy of most recommender systems is evaluated according to precision and recall. However, calculating these metrics for a recommender algorithm presents some problems [Baeza-Yates and Ribeiro-Neto, 2011; Herlocker et al., 2004]. First, these metrics require knowing whether each resulting item is relevant. In general, it is very difficult to define an item relevance. Second, there is in general a small number of relevant items in a item set. Third, it is necessary to consider resulting (recommended) items that are selected from a much larger set.

Therefore, the focus of this paper is on recall because: *(i)* in general, the networks are very sparse and the total number of possible links is large (as shown in Section 5); *(ii)* the *Affin* and *GLI* metrics aim to make networks more connected, as opposed to totally connected; and *(iii)* high recall indicates that the metrics provide correct recommendations. Just to give an idea of result size, the average number of recommendations for each researcher is 176 in *CiênciaBrasil* and 22 in DBLP (details of these datasets will be presented in Section 5) out of thousands of possibilities.

This decision (of focusing on recall) is also emphasized in the literature. Specifically, Menzies et al. [2007] present many examples of situations where high recall (and low precision) are useful, including: a commercial Web search engine like *Google* that reports more than 10^9 Web pages to a query with the word "software", and the effort involved in looking at a page is so low that users do not mind examining false results; and Cleland-Huang et al. [2006] won the best paper award at the 2006 IEEE Requirements Engineering conference with a data mining method exhibiting precision of about 0.25 (even with low-precision, the analysis of results suggests that the proposed classification algorithm can detect many different types of non-functional requirements).

4.2 Novelty

New recommendations are indications of items that users do not know and would not know in the absence of a recommender algorithm. The *novelty* metric aims to quantify the "novel" or "original" characteristic in a recommendation list [Fouss and Saerens, 2008]. In order to compute this metric, we have adapted the idea proposed in [Fouss and Saerens, 2008] for the setting of an academic SN.

Algorithm 1 shows how to compute the *novelty* of recommendations. Lines 4 and 5 describe the frequency calculation of each recommended researcher in the recommendation list. This frequency represents the popularity degree of the researchers, i.e., researchers with high frequency are likely to be known. In this case, we consider that the less popular a recommended researcher is (included in the recommendation list), the most probable he/she is unknown to a target researcher.

In line 8, the median is used as a central tendency metric to represent the fre-

Algorithm 1 Calculate-Novelty

```
1: {**Input**: Recommendation list $\mathcal{L}$, number of researchers $n$}
2: {**Output**: Average frequency $\mu_{f_m}$}
3: {For each recommended researcher $r$ in $\mathcal{L}$}
4: for $r \in \mathcal{L}$ do
5:     $f_r$ := calculate_frequency($r$)
6:     $\Im := \Im \cup f_r$ {Define the frequency of each recommended researcher to a target
       researcher $t$ in $\mathcal{L}$}
7: end for
8: $f_m$ := calculate_median($\Im$)
9: $\mu_{f_m}$ := calculate_average($f_m$, $n$)
10: return $\mu_{f_m}$
```

quencies (following the proposal by Fouss and Saerens [2008]). Finally, in line 9, the frequency median of the recommended researchers is divided by the total number of target researchers. Hence, it provides the distribution of the frequency median in relation to target researchers.

The resulting value represents the *novelty* in a recommendation list. The *novelty* metric varies in the range [0,1], in which values near zero represent greatest novelty and the opposite when approaching one.

4.3 Diversity

Diversity is generally defined as the opposite of similarity [Shani and Gunawardana, 2011]. In some cases, suggesting a set of similar items may not be useful. For example, considering a collaboration recommendation where an algorithm should recommend researchers. Presenting a list with 10 researchers, all from the same institution or research groups may not be as useful as recommending researchers from various places. This follows the intuition that researchers from the same institution have a higher probability of already knowing each other.

The most explored method to measure *diversity* in a recommendation list is using the intra-list similarity metric [Shani and Gunawardana, 2011]. We use this method based on the approach presented by Ziegler et al. [2005], which evaluates traditional recommender systems. In addition, some changes have been made in this approach to evaluate collaborations recommendations.

Given a set of all target researchers $\mathcal{T}$ and a recommendation list $\mathcal{L}$, Algorithm 2 describes how to calculate the *diversity* using the intra-list similarity metric. Line 4 defines how to measure the similarity between recommended researchers in a recommendation list. In general, this similarity is defined by *Pearson's correlation* or *cosine distance* [Ziegler et al., 2005]. However, in this work, the correlation (defined by Lopes et al. [2010]) among researchers that represents the semantic of the SN relations (links) has been used to calculate this similarity. Line 5 describes the calculation of the intra-list similarity metric that considers the similarity between researchers. In line 7, the *diversity* metric is computed, where high values indicate low diversity.

The resulting values for *diversity* are not in a specific range. Thus, after computing this metric for different databases, the values are normalized linearly within the interval $[0, 1]$ (line 8).

Algorithm 2 Calculate-Diversity

```
1: {Input: Set of all target researchers $\mathcal{T}$, recommendation list $\mathcal{L}$, number of researchers $n$ }
2: {Output: Average intra-list similarity $\mu_{\mathcal{S}_{in-list}}$}
3: for $t \in \mathcal{T}$ do
4:     $\mathcal{S}_{\mathcal{L}}$ := calc_similarity($t$, $\mathcal{L}$) {Calculate the similarity among recommended researchers to $t$ in $\mathcal{L}$};
5:     $\mathcal{S}_{in-list}$ := calc_similarity_intraList($\mathcal{S}_{\mathcal{L}}$, $\mathcal{L}$) {As defined in [Ziegler et al., 2005]};
6: end for
7: $\mu_{\mathcal{S}_{in-list}}$ := calculate_average($\mathcal{S}_{in-list}$, $n$);
8: $\mu_{\mathcal{S}_{in-list}}$ := normalize($\mu_{\mathcal{S}_{in-list}}$) {a linear normalization};
9: return $\mu_{\mathcal{S}_{in-list}}$;
```

4.4 Coverage

The term coverage refers to distinct properties of a recommender system such as *item space coverage* and *user space coverage*. This work needs only the property *item space coverage*, because computing *user space coverage* requires knowledge about users preferences which are beyond our scope. The term *item space coverage* refers to the proportion of items that the recommender system can recommend [Shani and Gunawardana, 2011]. This property may be represented by metrics that compute how unequally different items are recommended to users. Two different metrics are used to compute this distributional inequality: *Gini index* (GI) and *Shannon Entropy* (SE) [Shani and Gunawardana, 2011]. Algorithms 3 and 4 describe how to compute these two metrics for academic collaborations. Furthermore, the approaches of the *Gini index* and *Shannon Entropy* presented in [Shani and Gunawardana, 2011] have been considered as the base for these algorithms.

Algorithm 3 Calculate-GiniIndex

```
1: {**Input**: Recommendation list $\mathcal{L}$}
2: {**Output**: Gini index $\mathcal{G}$}
3: $\mathcal{T}otal_{\mathcal{L}}$ := calcTotal_recommendedResearchers($\mathcal{L}$)
4: $\mathcal{T}otDif_{\mathcal{L}}$ := calcTotDif_recommendedResearchers($\mathcal{L}$)
5: {For each recommended researcher $r$ in $\mathcal{L}$}
6: for $r \in \mathcal{L}$ do
7:     $f_r$ := calculate_frequency($r$)
8:     $p_r := f_r / \mathcal{T}otal_{\mathcal{L}}$
9:     $\mathcal{P} := \mathcal{P} \cup p_r$ {Compute the proportion of each recommended researcher to $t$ in
       $\mathcal{L}$}
10: end for
11: increasing_order($\mathcal{P}$)
12: $j = 0$ {Count the total of each recommended researcher to $t$ in $\mathcal{L}$}
13: $sum = 0$
14: {Compute *Gini index* as defined by [Shani and Gunawardana, 2011]}
15: for $p_r \in \mathcal{P}$ do
16:     $sum := sum + ((2 * j) - \mathcal{T}otDif_{\mathcal{R}} - 1) * p_r$
17:     $j := j + 1$
18: end for
19: $\mathcal{G} := (1/(\mathcal{T}otDif_{\mathcal{R}} - 1)) * sum$;
20: return $\mathcal{G}$;
```

Given a recommendation list $\mathcal{L}$, Algorithm 3 computes the Gini index. Line 3 computes the total number of the recommended researchers in the recommendation list. Line 4 calculates the total number of different recommended researchers. The next step calculates the proportion of each recommended researcher p_r, lines 6-10. Following Shani and Gunawardana [2011], the set of proportion $\mathcal{P}$ is ordered according to increasing values p_r (line 11). Finally, in lines 15-19, the *Gini index* is computed according to [Shani and Gunawardana, 2011]. The index is zero when all researchers are recommended equally often, and one when a single researcher is always recommended.

In this work, we are interested in recommendations with *Gini index* near zero, which represents that each researcher receives distinguished recommendations according to his/her characteristics (affiliation, geographic localization and similar research area). If all researchers receive the same recommendations, the recommendations may be wrong, since each researcher has different characteristics.

Likewise, given a recommendation list, Algorithm 4 shows how to compute the Shannon Entropy. The initial steps of Algorithm 4 are similar to Algorithm 3 (lines 1-9), because it also needs to know the proportion of each recommended researcher. Following Shani and Gunawardana [2011], lines 12-15 computes the Shannon Entropy.

Algorithm 4 Calculate-ShannonEntropy

```
{**Input**: Recommendation list $\mathcal{L}$}
{**Output**: Shannon Entropy $E$}
$\mathcal{T}otal_{\mathcal{L}}$ := calcTotal_recommendedResearchers($\mathcal{L}$)
{For each recommended researcher $r$ in $\mathcal{L}$}
**for** $r \in \mathcal{L}$ **do**
    $f_r$ := calculate_frequency($r$)
    $p_r := f_r / \mathcal{T}otal_{\mathcal{L}}$
    $\mathcal{P} := \mathcal{P} \cup p_r$ {Compute the proportion of each recommended researcher to $t$ in $\mathcal{L}$}
**end for**
$sum = 0$
{Compute Shannon Entropy as defined in [Shani and Gunawardana, 2011]}
**for** $p_r \in \mathcal{P}$ **do**
    $sum := sum + (p_r * \log p_r)$
**end for**
$E := -sum$;
**return** $E$;
```

The entropy is zero when a single researcher is always recommended, and $\log n$ when n researchers are recommended equally often (n is the total number of distinct recommended researchers in the recommendation list). Similarly to the *Gini index*, here we are also interested in a recommendation list with many different researchers, i.e., Shannon Entropy near $\log n$.

4.5 Concluding Remarks

This Chapter addressed the problems of evaluating recommender algorithms. We described different metrics that can be used to evaluate recommender algorithms. However, it is necessary to consider the aim of the application when choosing more appropriate evaluation metrics. As discussed over this section, we have chosen accuracy, novelty, diversity and coverage to evaluate recommendations of collaborations.

Chapter 5

Experiments and Results

As previously discussed, this work also contributes to the way a recommendation function is evaluated. The previous section reviewed the accuracy metric (traditionally employed for evaluating recommender systems) and also introduced novelty, diversity and coverage. This chapter details the datasets employed in our experimental evaluation (Section 5.1), shows how the weights are defined (Section 5.2), and then presents the evaluation results (Section 5.3) and a graph analysis of the datasets (Section 5.4).

5.1 Dataset Details

The experiments were performed using two real datasets that were built from *CiênciaBrasil*[1] and DBLP academic social networks, as detailed next.

5.1.1 Real Dataset 1: CiênciaBrasil

The *CiênciaBrasil* dataset contains Lattes[2] résumés of Brazilian researchers from selected research groups [Laender et al., 2011]. The academic social network built from this dataset included **340** Computer Science researchers connected by the relation of co-authorship. If all researchers collaborated with each other, the number of relations (links) between them would be 57,630. Moreover, we have also limited the set of publications to include only those published from 2000 to 2011.

[1] CiênciaBrasil: `http://pbct.inweb.org.br`

[2] Lattes: `http://lattes.cnpq.br`

5.1.2 Real Dataset 2: DBLP

Based on the DBLP digital library dataset, we built an academic social network for **629** researchers from 45 Brazilian institutions and their publications from 1971 to 2012. If all researchers collaborated with each other, the number of relations would be 197,506. In order to provide a more homogeneous dataset, we have also limited the publications to those published in conference proceedings and journals (*i.e.*, dataset elements *inproceedings* or *article*). This way, both *CiênciaBrasil* and DBLP cover the same types of publication, although in different time intervals and for different universe of researchers (such differences enrich our results, as discussed ahead).

It is important to notice that, contrary to the *CiênciaBrasil* dataset, the DBLP dataset provides no information on the researcher's affiliation. Given that *Affin* and *GLI* need such an information, we have manually defined it for each of the 629 researchers with data extracted from CAPES[3].

5.1.3 Building the Academic Social Networks

The algorithm for building an academic social network based on co-authorship relations is simple: each researcher becomes a node in the network; for each pair of researchers a and b, if they have co-authored at least one publication, then an edge is added between their nodes. The social networks were built based strictly on the given datasets. In other words, it is possible that only a subset of the researchers' publications is represented in the SN, provided that the researchers may have other publications that are outside the datasets. However, given the coverage of both datasets in terms of conferences and journals, we believe that the most relevant part of the researchers' publications is reflected in the datasets and is enough for providing good recommendations. Finally, the focus of this chapter is in comparing the results across different metrics, not the absolute results themselves.

Each of the two datasets was divided in two parts (based on the concept of *split* [Baeza-Yates and Ribeiro-Neto, 2011]): 90% of the data as characterization set of the recommendations, and the remaining 10% for validation. The first part (the largest percentage of the data) was explored to create the researchers' profile and the social network. The second, smallest part is the testing one, which means that it contains the expected results a recommender system should provide. Furthermore, both parts also follow the time interval distribution, where the first part considers publications prior to the second part. In other words, the second part represents the "future" of

[3] CAPES: `http://www.capes.gov.br`

Table 5.1. Information about the networks.

	CiênciaBrasil		***DBLP***	
Information	**90%**	**10%**	**90%**	**10%**
Period in years	2000-2009	2009-2011	1971-2011	2011-2012
Total of publications	11,598	1,289	9,583	1,064
Publications avg. by researcher	34.11	3.79	15.24	1.69
Number of co-authorship relations	454	75	517	105

Note: Avg. = Average

the first one, and hence allows us to see what recommendations would be more useful. Note that this is one way to evaluate the recommender system while avoiding an actual feedback from the users – which is another way of doing so.

Table 5.1 describes the splits from both datasets and their social networks. It is clear that each SN is sparse (less than 460 relations of co-authorships from a possible total of 57,630 for *CiênciaBrasil*, and less than 520 from possible 197,506 for DBLP). Consequently, there are many possible results for a recommendation function to consider (approximately, 57,630 - 460 = 57,170 for *CiênciaBrasil* and 197,506 - 520 = 196,986 for DBLP).

Comparing the two parts (in different time intervals) of each dataset shows that new collaborations have started during the second interval. The number of co-authorship relations (Table 5.1) in the second network of each dataset represents only new co-authorships, excluding existing relations from the first network. These new co-authorships would already be in the first SN, making it more connected, if a good recommendation system was used.

Furthermore, we compare the results of *Affin* and *GLI* with *CORALS*. *CORALS* builds the SN for each dataset considering the publications of all researchers with one relevant difference: it includes researchers correlated by researchers area and some level of social closeness. On the other hand, both *Affin* and *GLI* will consider the same universe of researchers that *CORALS* plus the researchers correlated by affiliation when building their SN. In order to provide a better comparison, we have combined *CORALS* and *Affin* in a new metric, called *CORALS+Affin*, that works on a SN built as *CORALS* including all researchers correlated by research area, social closeness and affiliation. Note that we did not consider combining *CORALS* and *GLI* because, as shown in the next section, there is no relation between cooperation and location.

5.2 Setting the Weights for the Affin Metric

The aim of this section is to analyze the results of Cr and Sc from CORALS, and *Affin* for generating ranked recommendations. This study helps to assign values for weights w_{Affin} and w_{Sc} from Equation 3.3. It is important to attribute a correct value, because the recommendation score depends of such values (as shown in Section 3.2).

Tables 5.2 and 5.3 show the number of relevant recommendations retrieved (underlined), the total number of recommendations retrieved (in parentheses) and recall results for the individual and combined metrics using *CiênciaBrasil* and DBLP, respectively. *Affin*, Cr and Sc are considered separately; *Affin* and Sc are combined using intersection, ordered by one of the metrics; and the union of *Affin* and Sc is defined by *Affin_Sc*. In addition, *Affin* and Sc are not combined with Cr, because Cr retrieves many pairs of researchers and is not much more relevant than *Affin* and Sc.

According to Tables 5.2 and 5.3, Sc presents recall greater than *Affin*. Hence, in order to increase recall, w_{Sc} must be greater than w_{Affin}.

Table 5.2. *CiênciaBrasil*: relevant/retrieved recommendations and recall

Method	Relevant/Retrieved	Recall
Affin	57 (3,981)	0.760
Cr	68 (20,652)	0.906
Sc	69 (12,530)	0.920
Affin $\cap$ Sc (ordered by *Affin*)	54 (3,066)	0.720
Affin $\cap$ Sc (ordered by Sc)	54 (3,066)	0.720
Affin_Sc ($w_{\mathit{Affin}} = 1$, $w_{Sc} = 150$)	64 (6,087)	0.853

Table 5.3. DBLP: relevant/retrieved recommendations and recall

Method	Relevant/Retrieved	Recall
Affin	10 (5,950)	0.710
Cr	13 (44,414)	0.920
Sc	12 (28,646)	0.850
Affin $\cap$ Sc (ordered by *Affin*)	10 (5,059)	0.710
Affin $\cap$ Sc (ordered by Sc)	10 (5,059)	0.710
Affin_Sc ($w_{\mathit{Affin}} = 1$, $w_{Sc} = 25$)	12 (12,422)	0.857

Table 5.4 shows recall results when varying w_{Sc} values for a fixed $w_{\mathit{Affin}} = 1$ in *CiênciaBrasil* and DBLP networks. Regarding *CiênciaBrasil*, recall stabilize around $w_{Sc} = 25$. Thus, any value above 25 may be selected. In the DBLP network, precision and recall also stabilize around $w_{Sc} = 25$. Thus, in the experimental evaluation, we chose $w_{Sc} = 150$ for *CiênciaBrasil* and $w_{Sc} = 25$ for DBLP.

Table 5.4. Recall for different w_{Sc} ($w_{Affin} = 1$)

$\mathbf{w_{Sc}}$	***CiênciaBrasil***	**DBLP**
0	0.173	0.142
1	0.320	0.285
25	0.853	**0.857**
50	0.853	0.857
75	0.853	0.857
100	0.853	0.857
125	0.853	0.857
150	**0.853**	0.857
175	0.853	0.857
200	0.853	0.857
...	....	...

5.3 Evaluation Results

We have grouped our experimental results as follows. Section 5.3.1 presents the results when the metrics are used to recommend new collaborations. Likewise, Section 5.3.2 shows the results when the metrics are used to recommend "intensifiable" collaborations (i.e., those existing collaborations that can be further intensified).

5.3.1 Recommending New Collaborations

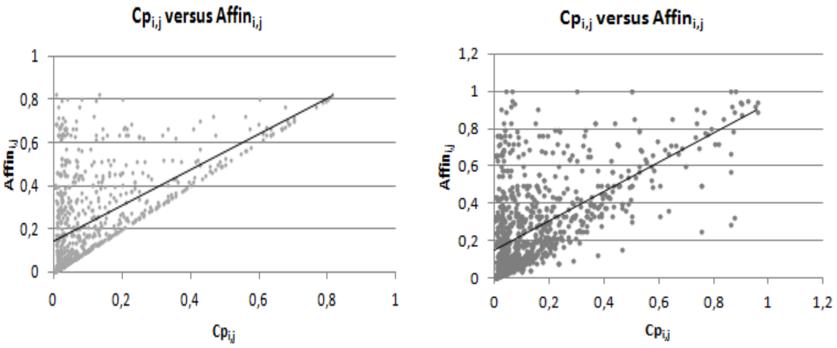
This first set of experiments considers both the *CiênciaBrasil* and DBLP datasets. Then it evaluates *Affin* and *GLI* versus CORALS and *CORALS+Affin*. Table 5.5 presents the results of recall (in percentage) of the experiments. Note that, as discussed in Section 4.1, accuracy is given by recall only. The results show that using institutional affiliation leads to an improvement in accuracy. Thus, in *CiênciaBrasil* and DBLP, *Affin* performs better than *GLI* and CORALS. The recall of *CORALS+Affin* is equal to *Affin*, because *Affin* adds affiliation to the original *CORALS*.

Table 5.5. New collaborations - Recall

Network	**Affin**	**GLI**	**CORALS**	**CORALS+Affin**
CiênciaBrasil	0.8533	0.6666	0.7733	0.8533
DBLP	0.8571	0.7647	0.8571	0.8571

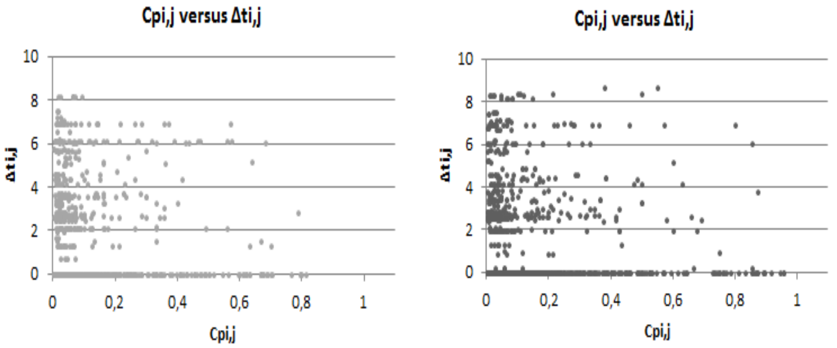
A complementary result is illustrated in Figure 5.1. It shows that affiliation and cooperation ($Affin_{i,j}$ and $Cp_{i,j}$) are directly related. As presented by Cohen [1988] and Hopkins [2002], the correlation coefficient (CC) in both SN is large, i.e. greater than 0.5. This fact explains why *Affin* provides more accurate recommendations.

Regarding geographic location, *GLI* presents the worst accuracy results. For better understanding, the graphics in Figure 5.2 show that intensifying cooperation and improving *travel time* ($Cp_{i,j}$ and $\Delta t_{i,j}$) are not related. This is clear when observing that there are pairs of researchers (points) indicating high cooperation in high *travel time* and low cooperation in low *travel time*. Furthermore, the correlation coefficient is in the range [-0.09; 0.0] which indicates the lack of correlation [Cohen, 1988]. This is a common behavior in both *CiênciaBrasil* and DBLP datasets.



(a) *CiênciaBrasil*: CC = 0.573 (b) DBLP: CC = 0.601

Figure 5.1. The (clear) relation between *Affin* and Cooperation for *CiênciaBrasil* and DBLP.



(a) *CiênciaBrasil*: CC = -0.047 (b) DBLP: CC = -0.090

Figure 5.2. The (non-existant) relation between travel time and Cooperation for *CiênciaBrasil* and DBLP.

Table 5.6 shows the results to the *novelty* and *diversity* metrics, in which the values in parentheses represent the *diversity* normalized in $[0, 1]$ (note that zero and one are only representative values to compare the metrics). In both social networks, *GLI* provides recommendations with more *novelty* and *diversity*. *Affin* presents the second best value for *diversity* and the same result as *CORALS+Affin* for *novelty*.

Table 5.6. New collaborations - *Novelty* and *Diversity*

	CiênciaBrasil		**DBLP**	
Metric	**Novelty**	**Diversity**	**Novelty**	**Diversity**
Affin	0.139	0.75	0.124	0.96
GLI	0.1233	0.0	0.044	0.0
CORALS	0.137	1.0	0.124	1.0
CORALS+Affin	0.139	0.78	0.124	1.0

Note: the higher the values, the worse the results

Table 5.7. New collaborations - *Coverage*

	CiênciaBrasil		**DBLP**	
Metric	**Gini I.**	**Shannon E.**	**Gini I.**	**Shannon E.**
Affin	0.416	4.93	0.492	5.214
GLI	0.385	5.19	0.473	5.46
CORALS	0.445	4.85	0.490	5.217
CORALS+Affin	0.424	4.92	0.490	5.218

Table 5.7 shows that using geographic location leads to an improvement in *coverage*. *GLI* generates a recommendation list with more unequally different researchers, and presents the best results for *Gini index* and *Shannon Entropy* (as detailed in Section 4.4) in *CiênciaBrasil* and DBLP. *Affin* presents the second best result in *CiênciaBrasil* and the worst in DBLP. Moreover, *CORALS+Affin* shows results better than CORALS for *coverage* in both SN, because *CORALS+Affin* considers more researchers than CORALS in the recommendations, which increases the difference between them.

Overall, the comparative analysis of Tables 5.5, 5.6 and 5.7 shows that even though *GLI* presents the worst results for accuracy, it presents the best ones for *novelty*, *diversity* and *coverage*. The reasoning for such results is as follows. Each target researcher receives recommendations considering similarity criteria (e.g., homophily or proximity principles); and increasing the number of recommended researchers (in this work, it increases the accuracy) also improves the similarity between them; hence, decreasing both *novelty* and *diversity*. Moreover, the number of researchers is finite, which means that the greater the number of recommended researchers, the less different they are in the resulting recommendation list; thus, the lower the *coverage*.

5.3.2 Recommending Intensifiable Collaborations

As previously discussed, besides recommending new collaborations, we also work on recommending (existing) collaborations that can be further intensified, i.e., the intensifiable collaborations. In order to evaluate such recommendations, we consider only accuracy. Note that other evaluations do not apply for intensifiable collaborations, because *novelty*, *diversity* and *coverage* cannot be established for existing collaborations.

Table 5.8 presents the results for accuracy of the recommendations to intensify

Table 5.8. Intensify collaborations - Recall

Network	Affin	GLI	CORALS	CORALS+Affin
CiênciaBrasil	0.8831	0.9805	0.7467	0.7467
DBLP	0.7714	0.9518	0.7619	0.7619

collaborations. *GLI* shows recommendations with the best recall (for both networks), which is justified because it distinguishes researchers with near and far *travel time*, increasing the number of relevant results. *Affin* presents the second best recall (for the two social networks). This shows that the affiliation can improve the accuracy of the recommendations. Finally, CORALS and *CORALS+Affin* present the same results.

5.4 Graph Analyses

In this section, we study the interactions within the datasets and analyze the properties of their networks based on two different graph analyses. First, in Section 5.4.1, we infer the strength of ties between researchers that have collaborated (i.e., to infer how connected two researchers are) [Gupte and Eliassi-Rad, 2012]. This study aims to show that our metrics recommend collaborations that will be weak ties, which are important for establishing bridges within the network. Then, in Section 5.4.2, we build and study social networks (using *CiênciaBrasil* and DBLP datasets) that represent the collaboration among researchers grouped by their institutions. This study is important to understand our metrics (*Affin* and *GLI*) that are based on the institutions. Overall, these two studies contribute to understand the results of the recommendations functions and, at the same time, provide further evidence that validates our metrics.

5.4.1 Tie Strength Represented by Absolute Cooperation

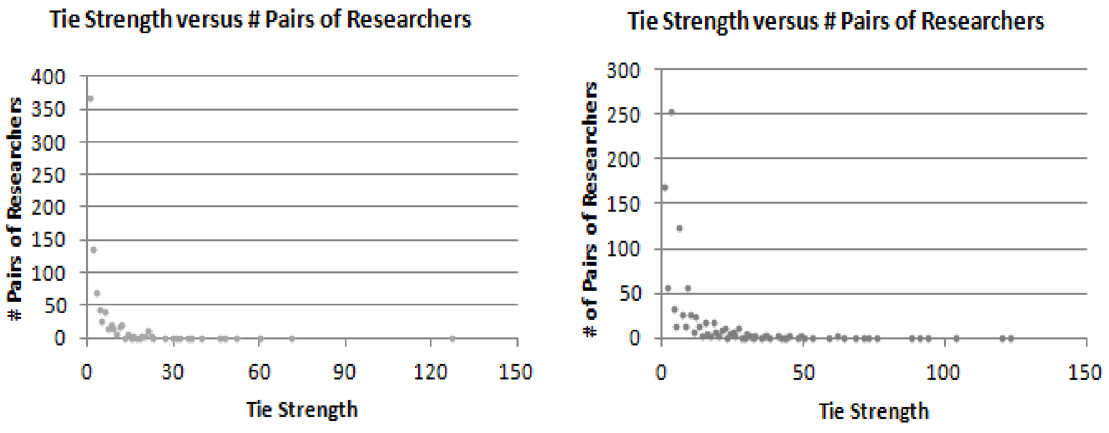
Given a graph that represents a social network, the tie strength measures how close the graph vertices are according to properties that are implicit in the graph [Gupte and Eliassi-Rad, 2012][4]. The strength of the ties has been studied, for example, in information diffusion [Granovetter, 1973; Bakshy et al., 2012], question answers [Panovich et al., 2012] and detection of important links in social networks [Gupte and Eliassi-Rad, 2012]. Here, we are interested in the strength of the cooperation between researchers and study the relation between tie strength and our metrics (*Affin* and *GLI*).

Each tie between two vertices may be defined as strong, weak or absent (including both the lack of relationship and ties without substantial significance) [Granovetter, 1973]. Weak ties are more likely to link people of different groups than strong ones. In

[4] Tie strength is not the same as edge weight, which is explicit in the graph.

this sense, weak ties act as bridges whereas strong ties lead to overall fragmentation. Furthermore, the more local bridges in a community organization network, the more cohesive the community and the more capable of acting in consonance [Granovetter, 1973]. Therefore, given the importance of having weak ties, we study their presence in the *CiênciaBrasil* and DBLP networks, and verify whether the recommendation metrics can generate weak ties.

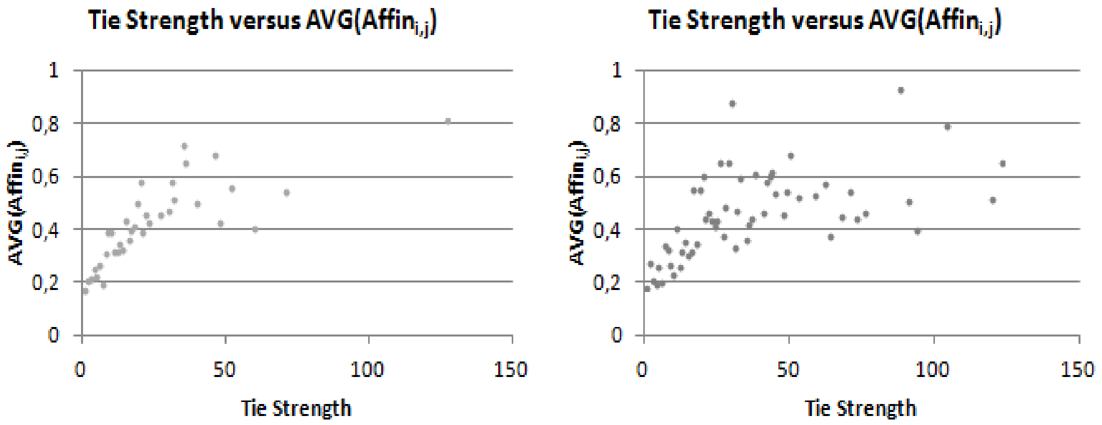
There are many metrics of tie strength [Gupte and Eliassi-Rad, 2012]. Common neighbors is the simplest metric and is used in this work. Given a pair of researchers, the common neighbors metric represents the absolute cooperation and measures the total number of papers that these researchers have co-authored. Figure 5.3(a) shows that most of the pairs of researchers have weak ties in *CiênciaBrasil* and DBLP.



(a) *CiênciaBrasil* (b) DBLP

Figure 5.3. Ties strength represented by the absolute cooperation.

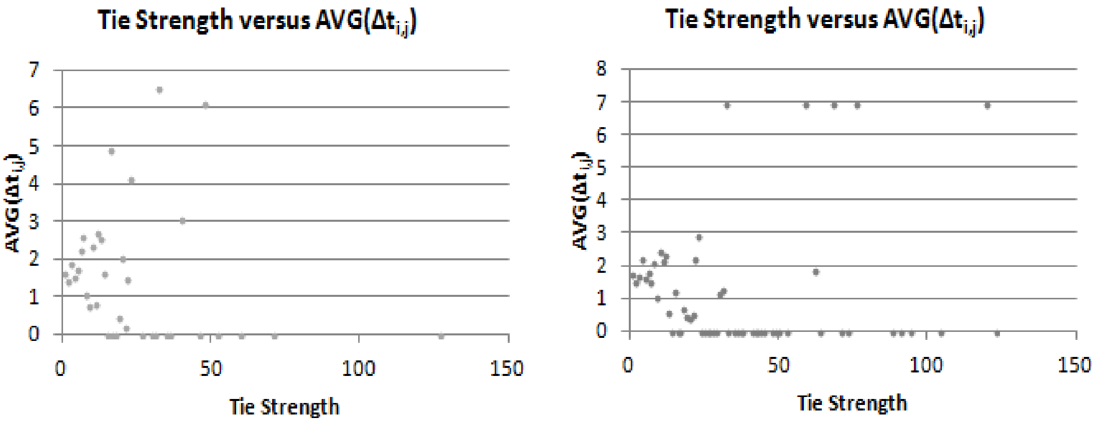
One important question is whether *Affin* and *GLI* can recommend collaborations that will also be weak ties. To answer this question, consider Figures 5.4 and 5.5. In these figures, $AVG(Affin_{i,j})$ and $AVG(\Delta t_{i,j})$ represent the average of the *affiliation weight* and the *travel time weight*, respectively, in relation to all pairs of researchers with the same tie strength. Figure 5.4 and the correlation coefficient shows that affiliation and tie strength are directly related. Moreover, as discussed in Chapter 3, the recommendation function needs the $Affin_{i,j}$ (combined with $Sc_{i,j}$) to be better than "low" to recommend new collaborations. Thus, as illustrated in the figures, weak ties are common when the value of $Affin_{i,j}$ is a little larger than the threshold that is specified in Section 3.2. This fact shows that *Affin* can recommend collaborations that will probably be weak ties between researchers. In other words, as already shown in Figure 5.3, the networks are already formed by weak ties. Now, Figure 5.4 shows that the new recommended collaborations will preserve such an important feature.



(a) CiênciaBrasil: CC = 0.759

(b) DBLP: CC = 0.564

Figure 5.4. The (clear) relation between *Affin* and tie strength for *CiênciaBrasil* and DBLP.



(a) CiênciaBrasil: CC = -0.189

(b) DBLP: CC = 0.101

Figure 5.5. The (non-existant) relation between *GLI* (represented by *travel time*) and tie strength for *CiênciaBrasil* and DBLP.

On the other hand, Figure 5.5 shows that *travel time* and *tie strength* are not related. Therefore, it is not possible to predict whether *GLI* will give recommendations that will be weak ties between researchers.

5.4.2 Aggregated Cooperation

One way to measure the cooperation between researchers from different institutions i and j is defined by Equation 5.1.

$$Cgroup_{i,j} = \frac{NPI_{i,j}}{NT_i} \tag{5.1}$$

where $NPI_{i,j}$ is the number of papers of researchers from institution i co-authored with people from institution j, and NT_i is the total number of papers authored by researchers from institution i. Equation 5.1 is very similar to Equation 3.1 that defines the affiliation index $Affin_{i,j}$. The difference is that $Cgroup_{i,j}$ provides the cooperation in relation to all researchers from the institutions. Here, $\langle i,j \rangle$ represents pairs of institutions, not pairs of researchers.

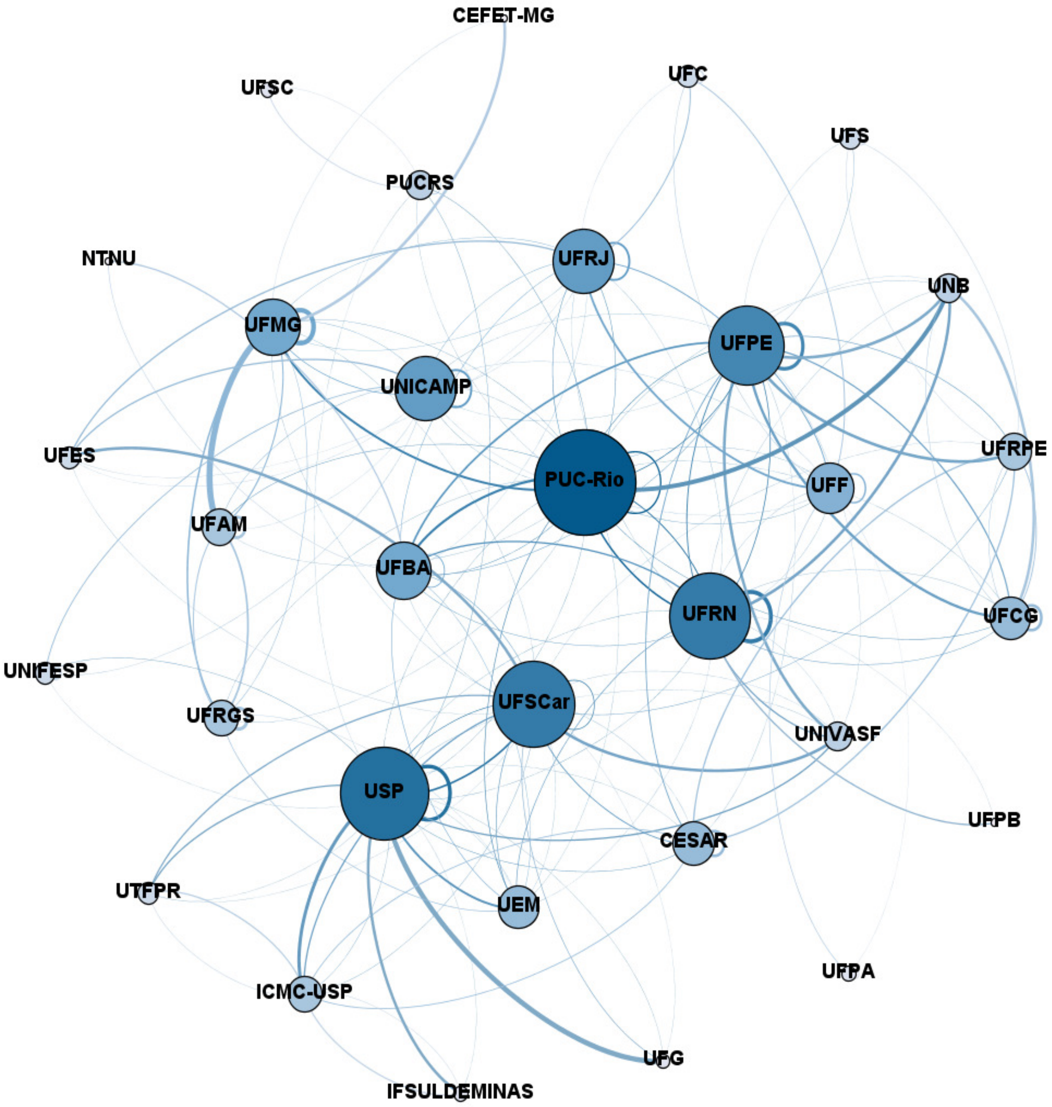


Figure 5.6. Co-authorship network between institutions - *CiênciaBrasil*

Figures 5.6 and 5.7 show the institution-level co-authorship graph from *CiênciaBrasil* and DBLP, respectively. Each graph represents the cooperation between researchers from pairs of institutions, where two institutions are connected if any two

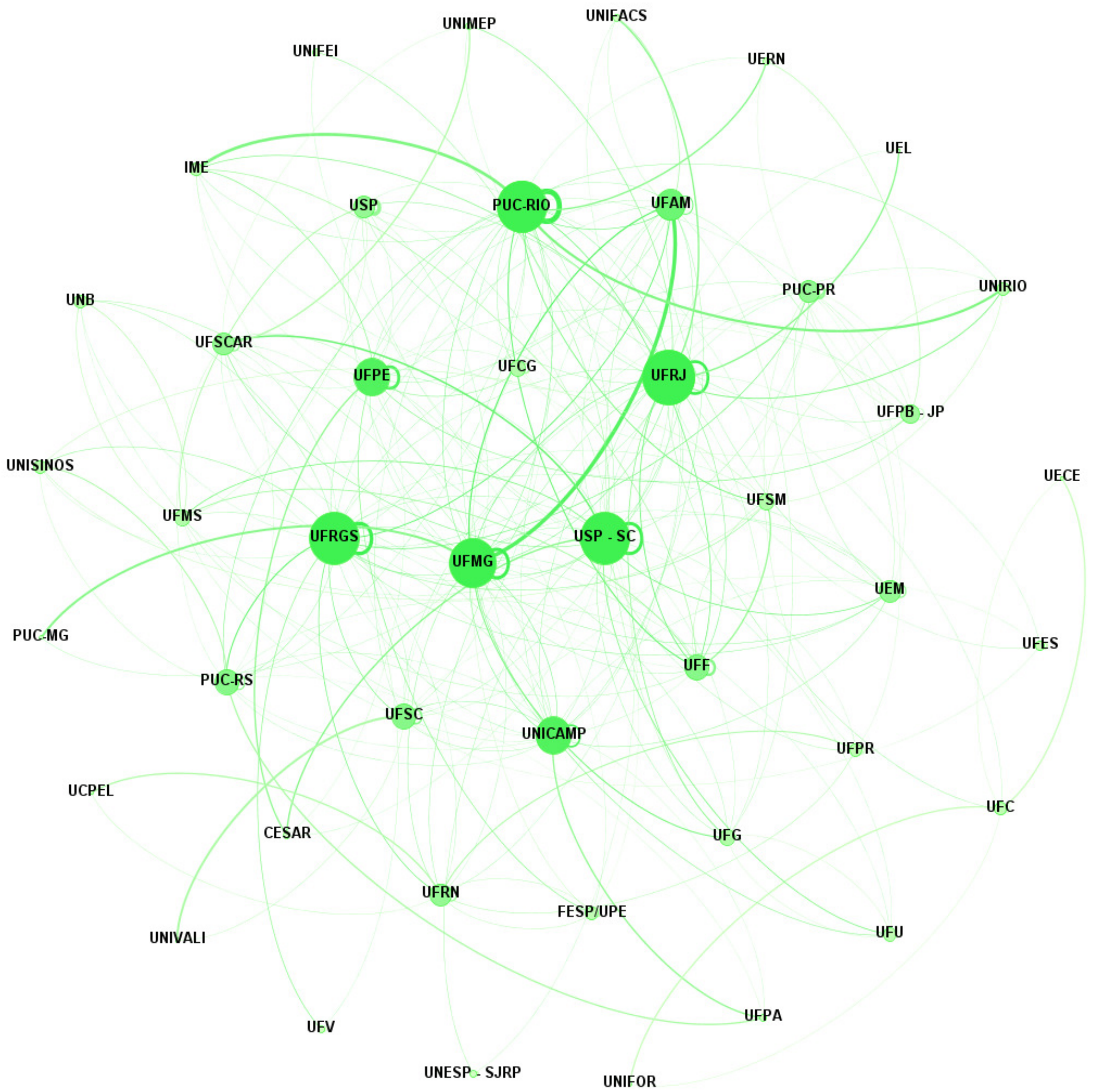


Figure 5.7. Co-authorship network between institutions - DBLP

of their researchers have coauthored a publication (both graphs were built using the testing set described in Table 5.1). In these figures, Equation 5.1 ($Cgroup_{i,j}$) defines the weight of the edges, where thicker edges have higher cooperation levels. For example, in *CiênciaBrasil*, there are 13 collaborations between institutions UFMG and UFAM and 1 between UFMG and UFSCar. Furthermore, the size of each node varies depending on the number of researchers represented by that institution. For example, in *CiênciaBrasil*, PUC-Rio has 31 researchers whereas UFPB has 13. We are interested in studying the properties of these networks and investigating if these networks follow the small-world phenomenon. This study shows how the affiliation with the institution influences in the cooperation between pairs of researchers. Hence, we can validate our metrics based on the institutions.

Section 2.1.2 showed that random networks have been used as benchmark to verify small-world properties in real networks. We have built two random networks in Gephi[5] with the properties described in Table 5.9 in order to compare their properties with the real networks and verify if the real networks follow small-world phenomenon. Each random network, Random1 and Random2, has the same number of nodes that *CiênciaBrasil* and DBLP, respectively. In these two networks, the probability of connection between nodes is defined by a binomial distribution [Erdös and Rényi, 1959]. We defined *0.5* for this probability so that the random networks have symmetrical distribution of the edges, no high clustering and high degree. Note that using *0.1* for the probability makes the random network more skewed. In other words, the probability of connecting two nodes is very small and makes the degree no more than four. For probability near *1*, all edges are reconnected, and a random network is obtained, with all corresponding properties such as low clustering and short path lengths. Thus, the ideal value to build these random networks is intermediate values of probability [Watts and Strogatz, 1998]. Next, we analyze each property in Table 5.9.

Table 5.9. Properties of the networks

Information	**Random1**	***CiênciaBrasil***	**Random2**	**DBLP**
Number of Nodes	32	32	44	44
Number of Edges	248	180	465	261
Density	0.25	0.181	0.246	0.138
Connected Components	1	1	1	1
Avg. Clustering Coefficient	0.25	0.454	0.248	0.373
Diameter	3	4	4	4
Average Degree	7.75	5.625	10.56	5.932
Average Path length	1.48	2.30	1.48	2.39

Table 5.9 shows that *CiênciaBrasil*, DBLP, Random1 and Random2 networks are not dense. In this context, *density* measures how many institutions are cooperating versus the total number of possible cooperations. This result shows that there is room for many collaborations to emerge between researchers from different institutions (i.e., *GLI* would favor such new collaborations).

In addition, the four networks have only one *connected component.* Such a feature ensures that all researchers from different institutions have a chance to cooperate with everyone else in the network [Wiemken et al., 2012].

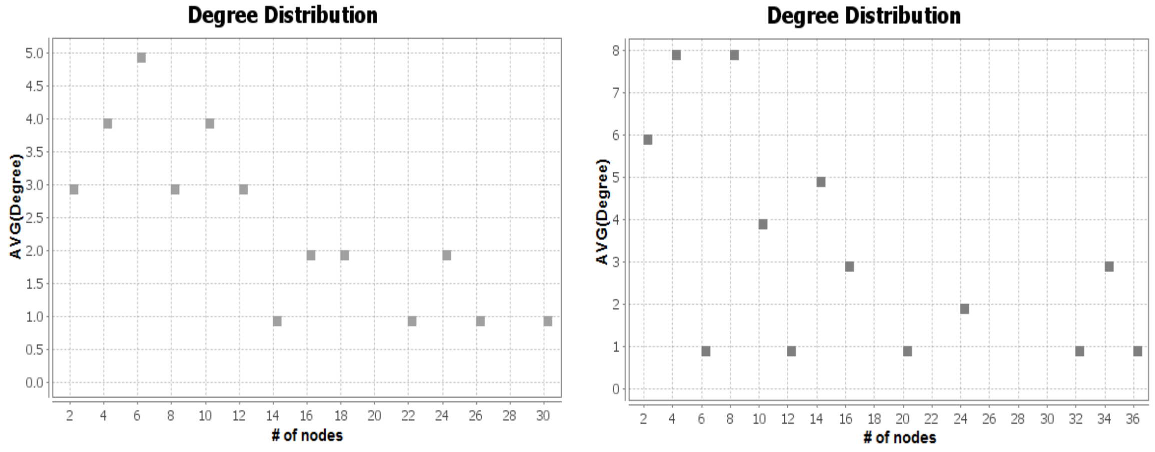
[5] Gephi: `https://gephi.org`

Regarding the *clustering coefficient*, we can see that there is a strong clustering effect in *CiênciaBrasil* and DBLP networks: two researchers from different institutions have more than 30% probability of collaborating if both have collaborated with a third institution (this claim follows the idea presented in [Newman, 2001]). Moreover, the clustering coefficient of these two real networks is higher than the clustering coefficient of the random ones, which satisfies one of the conditions for the networks to have small-world properties (as shown in Section 2.1.2).

The *diameter* of the two real social networks is very small, because it is necessary only four steps to get from one side of the network to the other [Mislove et al., 2007]. The diameter of Random1 network is the lowest. This can occur because the number of edges in Random1 is higher than *CiênciaBrasil* network.

Figure 5.8 presents the relation between a set of nodes and the *average degree* for *CiênciaBrasil* and DBLP networks. This fact shows that most nodes have only few links (low degree), but there are few nodes that are extremely linked (high degree). Hence, using a recommendation system has a real potential to increase the degree of these low degree nodes.

Furthermore, the *average path length* of both Random1 and Random2 is lower than *CiênciaBrasil* and DBLP. This is also justified by the number of edges and the *average degree*. Thus, the nodes in the random networks are more connected, and the distance between pairs of nodes is lower.



(a) *CiênciaBrasil*

(b) DBLP

Figure 5.8. Degree distribution: the average degree for a set of nodes in *CiênciaBrasil* and *DBLP*.

Therefore, we can conclude that the networks that represent the cooperation between Brazilian computer scientists from different institutions have the small-word properties: they present small diameter, small average path length (real networks' average path length $\geq$ random networks' average path length) and exhibit high clustering (real networks' clustering coefficient $\gg$ random networks' clustering coefficient) [Mislove et al., 2007; Watts and Strogatz, 1998]. These results agree with those from [Elmacioglu and Lee, 2005] that considered 32,689 authors and 38,773 papers from 1968 to 2003 of DBLP. This conclusion may help ensure that the physical distance alone does not influence in the cooperation between researchers (computer scientists) from different institutions. By contrast, this conclusion validates the *Affin* metric, because *Affin* considers that researchers from different institution can cooperate.

5.5 Concluding Remarks

In this chapter, we have presented and discussed an extensive evaluation study considering two real datasets and compared our metrics with the state-of-the-art. The results show that using the *Affin* metric leads to an improvement in accuracy of the recommendations. Furthermore, even though using the *GLI* metric presents the worst accuracy in the recommendation to initiate collaboration, it has a very positive impact when recommending intensifiable collaborations.

Regarding *novelty*, *diversity* and *coverage*, *GLI* presents the best results and *Affin* the second best. As complementary experiments, we have also evaluated the relations between *Affin* and Cooperation, and *travel time* (important to *GLI*) and Cooperation. The results showed a true relation between the first pair, and that the second pair is unrelated. Overall, the new metrics generate recommendations with more quality than the state-of-the-art (CORALS). In the experimentation, *Affin* was also combined with CORALS to improve CORALS' results (providing a more fair comparison).

Finally, we have also analyzed the properties of the co-authorship networks from *CiênciaBrasil* and DBLP by analyzing their weak ties and grouping researchers in their institutions. The analyses showed that *Affin* makes recommendations that will generate weak ties between researchers, which keeps the status of the existing ties. Furthermore, these networks have small-world properties that validate *Affin* and show that the physical distance (considered in *GLI*) does not influence in the cooperation.

Chapter 6

Conclusion

This work introduced two new metrics for recommending collaborations in an academic social network. Given a recommendation system, the hardest part is to define which metric should the recommendation function rely upon when producing the results. The base of our work is to consider the social aspects when recommending collaborations to researchers. Specifically, we consider the institutional affiliation aspect (*Affin*) and the geographic localization information (*GLI*) of all researchers in the social network. Both metrics focus on social principles: *Affin* on homophily and *GLI* on proximity. Besides providing these two new metrics, we have also proposed new ways for evaluating the recommendation results. Instead of relying only on precision and recall (the traditional ways), we have also proposed evaluation algorithms that consider novelty, diversity and coverage. Concluding this dissertation, next, we list our main contributions and plans for future work.

6.1 Contributions

The main contributions of this dissertation were:

- **An analysis of different approaches for social network studies, link prediction and recommendation of collaboration.** In Chapter 2, we presented many fundamental concepts of social networks, link prediction and recommendation systems. Furthermore, we discussed different approaches of these areas and how social theories influence such approaches.

- **A formal definition for collaborations recommendation.** We described the steps to recommend collaborations in an academic social network (Section

3.1). Moreover, we formally defined the recommendations of collaborations emphasizing the importance of the recommendation functions.

- ***Affin*, a metric for recommending collaborations.** In Section 3.2, we presented the metric *Affin* that considers the concept of institutional affiliation and the homophily principle. This metric was combined in a new recommendation function with cooperation, correlation and social closeness.

- ***GLI*, a second metric for recommending collaborations.** We proposed another metric to recommend collaborations called *GLI* (Section 3.3). This metric uses the geographic location of the researchers' institution to compute the travel time between institutions and is based on the proximity principle. *GLI* was also combined in a new recommendation function with cooperation and correlation.

- **The utilization of *novelty*, *diversity* and *coverage* measures for analyzing the quality of the recommendations.** In Chapter 4, we discussed the difficulties of analyzing recommendation algorithms and their results. Thus, we developed new algorithms to use *novelty*, *diversity* and *coverage* in our experimental evaluation. This is the first time that those measures are used for evaluating recommendation of collaborations.

- **An experimental evaluation using two real academic social networks and comparison to the state-of-the-art.** Two datasets were used and described in the experimentation (Section 5.1). Each dataset was divided in 90% for characterizing the recommendations and 10% for validation. Both sets also followed the time interval distribution, where the training set considered publications prior to the testing one. Many experiments were performed and the results showed that *Affin* and *GLI* can provide, in general, recommendations with more quality than the state-of-the-art. In addition, it was presented the clear relation between *Affin* and Cooperation, and the non-existence one between *GLI* (represented by *travel time*) and Cooperation. This fact reveals that the physical proximity does not influence in the intensity of the cooperation between researchers (Section 5.3).

- **An analysis of two academic social networks.** We built the co-authorship networks using two datasets, grouping researchers by the institutions. Our analysis showed that these networks follow the small-world properties and that *Affin* can recommend collaborations that probably will be weak ties. These facts validate *Affin*, at the same time, reaffirm that only the physical proximity does not influence in the cooperation (Section 5.4). Therefore, it is necessary to study others metrics to combine with *GLI*.

- **Publications.** The results of this dissertation are published in [Brandão and Moro, 2012a], [Brandão and Moro, 2012b] and [Brandão et al., 2013].

6.2 Future Work

Ideas for extending and improving this work include:

- **Refining the recommendation function.** Other link semantics can be extracted from the academic social network and used to recommend collaborations. For example, participation in events and courses. They could all be used for refining the existing recommendation function.

- **Studying other geographic factors that influence in the cooperation.** There are some factors (such as attending the same committees or/and conferences, studying or/and working in the same college) that can potentially contribute for two researchers to initiate collaboration.

- **Considering other metrics to evaluate the quality of the recommendations.** There are many measures proposed in the literature to evaluate the recommendations. It is necessary to investigate these measures and possibly verify how they perform in our experimental setup.

Bibliography

Adamic, L. A., Buyukkokten, O., and Adar, E. (2003). A social network caught in the web. *First Monday.*

Adams, J. (2012). Collaborations: The rise of research networks. *Nature*, pages 335--336.

Aiello, L. M., Barrat, A., Toulon, S., Schifanella, R., Cattuto, C., Markines, B., and Menczer, F. (2012). Friendship prediction and homophily in social media. *ACM Transactions on the Web*, pages 9:1--9:33.

Akcora et al, C. G. (2012). Privacy in Social Networks: How Risky is Your Social Graph? In *Proceedings of the IEEE International Conference on Data Engineering*, pages 9–19, Washington, DC, USA.

Albert, R., Jeong, H., and Barabasi, A. L. (1999). The diameter of the world wide web. *Nature*, pages 130--131.

Backstrom, L., Boldi, P., Rosa, M., Ugander, J., and Vigna, S. (2012). Four degrees of separation. *ACM International Conference on Web Science.*

Baeza-Yates, R. A. and Ribeiro-Neto, B. A. (2011). *Modern Information Retrieval - the concepts and technology behind search.* Pearson Education Ltd., Harlow, England.

Bakshy, E., Rosenn, I., Marlow, C., and Adamic, L. (2012). The role of social networks in information diffusion. In *Proceedings of the 21st International Conference on World Wide Web*, pages 519--528, New York, NY, USA.

Banyai, M., Nepusz, T., Negyessy, L., and Bazso, F. (2009). Convergence properties of some random networks. In *Proceedings of the 7th International Symposium on Intelligent Systems and Informatics*, pages 241 –245.

Barabasi, A.-L. (2002). *Linked: The New Science of Networks.* Perseus Books Group, Cambridge, MA, USA.

Bigonha, C., Cardoso, T. N. C., Moro, M. M., Gonçalves, M. A., and Almeida, V. A. F. (2011). Sentiment-based influence detection on twitter. *Journal of the Brazilian Computer Society*, pages 1--15.

Brandão, M. A. and Moro, M. M. (2012a). Affiliation Influence on Recommendation in Academic Social Networks. In *Proceedings of the 6th Alberto Mendelzon International Workshop on Foundations of Data Management*, pages 230–234, Belo Horizonte, MG, Brazil.

Brandão, M. A. and Moro, M. M. (2012b). Recomendação de colaboração em redes sociais acadêmicas baseada na afiliação dos pesquisadores. *Simpósio Brasileiro de Banco de Dados.*

Brandão, M. A., Moro, M. M., Lopes, G. R., and de Oliveira, J. P. M. (2013). Using Link Semantics to Recommend Collaborations in Academic Social Networks. In *WWW Workshops*, Rio de Janeiro, Brasil.

Brown, C., Nicosia, V., Scellato, S., Noulas, A., and Mascolo, C. (2012). The importance of being placefriends: discovering location-focused online communities. In *Proceedings of ACM Workshop on Online Social Networks*, pages 31--36, New York, NY, USA.

Cai, X., Bain, M., Krzywicki, A., Wobcke, W., Kim, Y. S., Compton, P., and Mahidadia:, A. (2010). Learning collaborative filtering and its application to people to people recommendation in social networks. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, pages 743--748, Washington, DC, USA.

Cleland-Huang, J., Settimi, R., Zou, X., and Solc, P. (2006). The Detection and Classification of Non-Functional Requirements with Application to Early Aspects. In *Proceedings of the 14th IEEE International Requirements Engineering Conference*, pages 39 –48, Washington, DC, USA.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* Routledge Academic, second edition.

Contractor, N. S., Wasserman, S., and Faust, K. (2006). Testing Multitheoretical, Multilevel Hypotheses about Organizational Networks: An Analytic Framework and Empirical Example. *The Academy of Management Review*, pages 681–703.

David, E. and Jon, K. (2010). *Networks, Crowds, and Markets: Reasoning About a Highly Connected World.* Cambridge University Press, New York, NY, USA.

Ding, Y. (2011). Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of Informetrics*, pages 187 -- 203.

Elmacioglu, E. and Lee, D. (2005). On six degrees of separation in dblp-db and more. *SIGMOD Rec.*, pages 33--40.

Erdös, P. and Rényi, A. (1959). On random graphs, i. *Publicationes Mathematicae (Debrecen)*, pages 290--297.

Figueiredo, D. R. (2011). Introdução a redes complexas. *Jornada de Atualização em Informática*, pages 303--358.

Fouss, F. and Saerens, M. (2008). Evaluating Performance of Recommender Systems: An Experimental Comparison. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 735--738, Washington, DC, USA.

Freyne, J., Berkovsky, S., Daly, E. M., and Geyer, W. (2010). Social networking feeds: recommending items of interest. In *Proceedings of the 4th ACM Conference on Recommender Systems*, pages 277--280, New York, NY, USA.

Ganev, V., Guo, Z., Serrano, D., Barbosa, D., and Stroulia, E. (2010). Exploring and visualizing academic social networks. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 1963--1964, New York, NY, USA.

Granovetter, M. S. (1973). The Strength of Weak Ties. *The American Journal of Sociology*, pages 1360--1380.

Gupte, M. and Eliassi-Rad, T. (2012). Measuring tie strength in implicit social networks. In *Web Science*, pages 109–118, Evanston, IL, USA.

Guy, I., Ronen, I., and Wilcox, E. (2009). Do you know?: recommending people to invite into your social network. In *Proceedings of the 14th International Conference on Intelligent User Interfaces*, pages 77--86, New York, NY, USA.

He, J. and Chu, W. W. (2010). A Social Network-Based Recommender System (SNRS) Data Mining for Social Network Data. In *Data Mining for Social Network Data*, pages 47--74. Springer US, Boston, MA.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *Journal of ACM Transactions on Information Systems*, pages 5--53.

Homans, G. C. (1950). *The human group.* Harcourt, Brace and Company, New York, USA.

Hopkins, W. G. (2002). A scale of magnitudes for the effect statistics. *New View Stat.*

Huang, Z. (2006). Link Prediction Based on Graph Topology: The Predictive Value of Generalized Clustering Coefficient. In *Proceedings of the Workshop on Link Analysis: Dynamics and Static of Large Networks*, Philadelphia, PA, USA.

Kaltenbrunner, A., Scellato, S., Volkovich, Y., Laniado, D., Currie, D., Jutemar, E. J., and Mascolo, C. (2012). Far from the eyes, close on the web: impact of geographic distance on online social interactions. In *Proceedings of the 2012 ACM Workshop on Online Social Networks*, pages 19--24, New York, NY, USA.

Kitahara, T. and Yoshikai, N. (2010). Organization structure analysis based on an affiliation network, and verification of its effectiveness. In *Proceedings of the 8th Asia-Pacific Symposium on Information and Telecommunication Technologies*, pages 1 –6.

Kleinberg, J. (2000). The small-world phenomenon: An algorithmic perspective. In *Proceedings of the 32nd ACM Symposium on Theory of Computing*, pages 163--170, Ithaca, NY, USA.

Kutty, S., Chen, L., and Nayak, R. (2012). A people-to-people recommendation system using tensor space models. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pages 187--192, New York, NY, USA.

Laender, A. H. F., Moro, M. M., da Silva, A. S., Jr., C. A. D., Gonçalves, M. A., Galante, R., Silva, A. J. C., Bigonha, C. A. S., Dalip, D. H., Barbosa, E. M., Borges, E. N., Cortez, E., Jr., P. S. P., de Alencar, R. O., Cardoso, T. N. C., and Salles, T. (2011). CiênciaBrasil-The Brazilian Portal of Science and Technology. In *Proceedings of Seminário Integrado de Software e Hardware*, Natal, RN, Brasil.

Lee, S., Kim, P., and Jeong, H. (2006). Statistical properties of sampled networks. *Physical Review E*, pages 102--109.

Levorato, V. and Petermann, C. (2011). Detection of communities in directed networks based on strongly p-connected components. In *Proceedings of International Conference on Computational Aspects of Social Networks*, pages 211--216.

Li, S. and Chen, Y. (2009). New algorithm for degree of network relation coupling in complex networks. In *Proceedings of IEEE International Conference on Grey Systems and Intelligent Services*, pages 1618 --1623.

Liben-Nowell, D. and Kleinberg, J. M. (2003). The link prediction problem for social networks. In *Proceedings of the 20th International Conference on Information and Knowledge Management*, pages 556--559, New York, NY, USA.

Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., and Tomkins, A. (2005). Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, pages 11623--11628.

Liu, X., Bollen, J., Nelson, M. L., and Van de Sompel, H. (2005). Co-authorship networks in the digital library research community. *Information Processing and Management*, pages 1462 -- 1480.

Lopes, G. R., Moro, M. M., da Silva, R., Barbosa, E. M., and de Oliveira, J. P. M. (2011). Ranking Strategy for Graduate Programs Evaluation. In *Proceedings of International Conference on Information Technology and Application*, Sydney, Australia.

Lopes, G. R., Moro, M. M., Wives, L. K., and de Oliveira, J. P. M. (2010). Collaboration Recommendation on Academic Social Networks. In *ER Workshops*, pages 190--199, Vancouver, Canada.

M'Chirgui, Z. (2010). Small-world or scale-free phenomena in internet. In *Proceedings of IEEE International Conference on Management of Innovation and Technology*, pages 78 -- 83.

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks . *Annual Review of Sociology*, pages 415--444.

Menzies, T., Dekhtyar, A., Distefano, J., and Greenwald, J. (2007). Problems with Precision: A Response to "Comments on 'Data Mining Static Code Attributes to Learn Defect Predictors' ". In *Proceedings of IEEE Transactions on Software Engineering*, pages 637--640, Piscataway, NJ, USA.

Milgram, S. (1967). The small world problem. *Psychology Today*, 2:60–67.

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. (2007). Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29--42, New York, NY, USA.

Monge, P. R. and Contractor, N. (2003). *Theories of communication networks*. Oxford University Press, New York, NY, USA.

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, 45(2):404--409.

Newman, M. E. J. (2003). The Structure and Function of Complex Networks. *SIAM Review*, pages 167--256.

Panovich, K., Miller, R., and Karger, D. (2012). Tie strength in question & answer on social network sites. In *Proceedings of ACM CSCW12 Conference on Computer-Supported Cooperative Work*, pages 1057--1066, New York, NY, USA.

Pizzato, L. A. and Silvestrini, C. (2011). Stochastic matching and collaborative filtering to recommend people to people. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, pages 341--344, New York, NY, USA.

Qi, G.-J., Aggarwal, C. C., and Huang, T. S. (2012). Community Detection with Edge Content in Social Media Networks. In *Proceedings of the 2012 IEEE 28th International Conference on Data Engineering*, pages 534--545, Arlington, USA.

Qiong-Fang, B., Ting-ting, H., Hui, L., and Mu-yun, F. (2010). Research on the diameter and average diameter of undirected double-loop networks. In *Proceedings of the 2010 Ninth International Conference on Grid and Cloud Computing*, pages 461--466, Washington, DC, USA.

Quercia, D. and Capra, L. (2009). FriendSensing: recommending friends using mobile phones. In *Proceedings of the Third ACM Conference on Recommender Systems*, pages 273--276, New York, NY, USA.

Shani, G. and Gunawardana, A. (2011). Evaluating Recommendation Systems. In *Recommender Systems Handbook*, pages 257--297. Boston, MA, USA.

Sharara, H., Singh, L., Getoor, L., and Mann, J. (2009). The dynamics of actor loyalty to groups in affiliation networks. In *Proceedings of International Conference on Advances in Social Network Analysis and Mining*, pages 101 --106, Washington, DC, USA.

Subramani, M. R. and Rajagopalan, B. (2003). Knowledge-sharing and influence in online social networks via viral marketing. *Communications of the ACM*, 46(12):300--307.

Symeonidis, P., Tiakas, E., and Manolopoulos, Y. (2010). Transitive node similarity for link prediction in social networks with positive and negative links. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, pages 183--190, Barcelona, Spain.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., and Su, Z. (2008). ArnetMiner: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 990--998, Las Vegas, NV, USA.

Tobler, W. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, pages 234--240.

Ugander, J., Karrer, B., Backstrom, L., and Marlow, C. (2011). The anatomy of the facebook social graph. *Computing Research Repository.*

Wang, C., Han, J., Jia, Y., Tang, J., Zhang, D., Yu, Y., and Guo, J. (2010). Mining advisor-advisee relationships from research publication networks. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 203--212, Washington, DC, USA.

Wang, D., Pedreschi, D., Song, C., Giannotti, F., and Barabási, A.-L. (2011). Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1100--1108, San Diego, USA.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications.* Cambridge University Press, Cambridge, UK.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of'small-world'networks. *Nature*, pages 409--10.

Weng, C., Chen, W.-Y., and Lo, T.-W. (2010). The effect of affiliation network on technological innovation. In *Proceedings of Technology Management for Global Economic Growth*, pages 1--5.

Wiemken, T. L., Ramirez, J. A., Polgreen, P., Peyrani, P., and Carrico, R. M. (2012). Evaluation of the knowledge-sharing social network of hospital-based infection preventionists in kentucky. *American Journal of Infection Control*, 40(5):440 -- 445.

Wu, B., Ke, Q., and Dong, Y. (2011). Degree and similarity based search in networks. In *Proceedings of Eighth International Conference on Fuzzy Systems and Knowledge Discovery*, pages 1267--1270.

Yagan, O. and Makowski, A. (2009). Random key graphs can they be small worlds? In *Proceedings of First International Conference on Networks and Communications*, pages 313 --318.

Yan, J. and Assimakopoulos, D. (2007). The small world and scale-free structure of an internet technical community. In *Proceedings of the 2007 Symposium on Computer Human Interaction for the Management of Information Technology*, New York, NY, USA.

Yang, X., Steck, H., and Liu, Y. (2012). Circle-based recommendation in online social networks. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1267--1275, New York, NY, USA.

Yiran, G. and Wenwen, Z. (2012). Improved search algorithm based on probability of node's degree on complex networks. In *Proceedings of Third International Conference on Digital Manufacturing and Automation*, pages 485--488.

Yu, X., Pan, A., Tang, L. A., Li, Z., and Han, J. (2011). Geo-friends recommendation in gps-based cyber-physical social network. In *Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining*, pages 361--368, Washington, DC, USA.

Yuan, M., Chen, L., and Yu, P. S. (2010). Personalized Privacy Protection in Social Networks. In *Proceedings of Very Large Data Base Endowment*, pages 141--150.

Ziegler, C.-N., McNee, S. M., Konstan, J. A., and Lausen, G. (2005). Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web*, pages 22--32, New York, NY, USA.