A FORÇA DOS RELACIONAMENTOS EM REDES SOCIAIS DE COAUTORIA: ANÁLISES, MÉTRICAS E UM NOVO MODELO COMPUTACIONAL

MICHELE AMARAL BRANDÃO

A FORÇA DOS RELACIONAMENTOS EM REDES SOCIAIS DE COAUTORIA: ANÁLISES, MÉTRICAS E UM NOVO MODELO COMPUTACIONAL

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

Orientadora: Mirella Moura Moro

Belo Horizonte

Abril de 2017

MICHELE AMARAL BRANDÃO

TIE STRENGTH IN CO-AUTHORSHIP SOCIAL NETWORKS: ANALYSES, METRICS AND A NEW COMPUTATIONAL MODEL

Thesis presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Doctor in Computer Science.

Advisor: Mirella Moura Moro

Belo Horizonte

April 2017

Brandão, Michele Amaral

B817t Tie strength in co-authorship social networks: analyses, metrics and a new computational model / Michele Amaral Brandão. — Belo Horizonte, 2017
xxx, 166 f. : il. ; 29cm

Tese (doutorado) — Universidade Federal de Minas Gerais – Departamento de Ciência da Computação

Orientadora: Mirella Moura Moro

1. Computação - Teses. 2. Redes de relações sociais. 3. Sistemas de recomendação. 4. Coautoria. I. Orientadora. II. Título.

CDU 519.6*73(043)

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

FOLHA DE APROVAÇÃO

Tie strength in co-authorship social networks: analyses, metrics and a new computational model

MICHELE AMARAL BRANDÃO

Tese defendida e aprovada pela banca examinadora constituída pelos Senhores:

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

PROF. ANA PAULA COUTO DA SILVA
Departamento de Ciência da Computação - UFMG

DRA. JONICE DE OLIVEIRA SAMPAIO
Departamento de Ciência da Computação - UFRJ

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

PROF. PEDRO OLMO STANCIOLI VAZ DE MELO
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 20 de abril de 2017.

I dedicate this work to everyone who believed in me.

Acknowledgments

My warm thanks to everyone who encouraged me and contributed in some way to the development of this work. Firstly, I want to thank God, my fountain of faith, light and hope. Thank you God for all your blessings to me and my family. For the strength You give me each day and for all the people around me. Without You, nothing would be possible.

Also, my huge thanks to my parents, Ivan and Margarete, for all the support, love, care and stimuli unconditionally dedicated in my entire life. If I'm taking another important step in my life, this is a result of the education that you gave me.

I also want to thank my lovely husband Rafick, for being always present, sharing the joys in moments of success and helping me to overcome the moments of discouragement. For understanding my absence and believing in me. Also, by the right word at the right time and the hug that makes everything seems better.

Additionally, I thank my family and dear friends, specially, my grandmother Janete and my godfather Brandão, my brother Tauan, my lovely uncles and cousins, Delza, Jorge, Tânia, Lívia, Taiana, Iolanda, Ludivan, Manuela, Renata, Leonardo and Sophia for the affection, encouragement and friendship.

I also thank my dear friends who make me feel at home in Belo Horizonte: Juliana Padilha, Laís Mota, Natália Machado, Guilherme Vezula, Danilo Seufitelli, Elaine Muniz, Mônica Machado, Arielle Moura, Ricardo Bigarella and Jaqueline Sauer.

I want to especially thank my advisor Mirella M. Moro for her teachings, disposition, patience, confidence, willingness, opportunities and friendship. Her guidance and knowledge were very essential. Words are few to describe the eternal gratitude for having devoted so much, and for so long, to my guidance.

I also want to thank the faculty at UFMG and UESC. Specially, professor Pedro O. S. Vaz de Melo for his wise ideas, suggestions, disposition and great opportunity of cooperation; professor Ana Paula Couto for all important discussions about social network analyses; professors Jussara Almeida and Wagner Meira Jr. (specifically at the beginning of my PhD) for their willingness to clarify doubts and give suggestions.

My thanks to the friends and colleagues for helping me during my studies. Specially, Luciana Maroun, Gabriela Brant, Natércia Aguilar, Daniel Hasan, Michele Brito, Guilherme Augusto, Matheus Diniz, Sérgio Canuto, Vitor Mangaravite, Júlio Reis, Daniel Xavier and Jeancarlo Leão.

I thank the professors, Dr. Ana Paula Couto, Dr. Jonice Oliveira, Dr. José Palazzo M. de Oliveira and Dr. Pedro O. S. Vaz de Melo, for accepting to be part of the defense committees, even in the face of so many commitments and activities.

I thank CAPES, for the financial support in the form of a scholarship. This work was also partially supported by CNPq and InWeb (Brazil). These financial supports are gratefully acknowledged.

I also want to thank my little dog Malú, for all her love, friendship and calming me down when I'm stressed and worried.

Finally, I thank the staff of the Computer Science Department (UFMG) for being always available to solve doubts and meeting requests. Specially, Sônia, Linda, Claudia, Sheila and Cida.

This work would not be possible without all the people in my life. For me, all of you are the best. **Thank you very much**!

"Success is going from failure to failure without losing your enthusiasm."
(Winston Churchill)

Resumo

O estudo de relacionamentos sociais tem sido utilizado para construir modelos rigorosos que revelam a evolução de redes sociais e seus dinamismos. Uma propriedade dos relacionamentos sociais é a força, a qual tem sido aplicada em diferentes contextos como por exemplo: difusão de informação, análises de padrões em logs de comunicação e avaliação da produtividade científica de pesquisadores. Especialmente, analisar a força dos relacionamentos permite investigar como diferentes relacionamentos desempenham papéis distintos e identificar o impacto em nível micro e macro na rede. O objetivo desta tese é medir a força dos relacionamentos de coautoria em redes sociais acadêmicas não-temporais e temporais. As principais contribuições são: (1) uma revisão do estado-da-arte e uma taxonomia para redes sociais profissionais, que contextualizam o problema abordado neste trabalho; (2) uma análise de como propriedades topológicas relacionam-se com a força dos relacionamentos, pois nossos resultados mostram que diferentes propriedades topológicas explicam variações na força dos relacionamentos de coautoria, dependendo da área de pesquisa; (3) uma nova métrica chamada *tieness* que é fácil de calcular e melhor diferencia a força dos relacionamentos em diferentes níveis em redes sociais de coautoria não-temporais; (4) uma análise da dinâmica das forças dos relacionamentos ao longo do tempo por meio de dois algoritmos, um já existente e um proposto aqui, chamado STACY (*Strength of Ties Automatic-Classifier over the Years*); e (5) um novo modelo computacional chamado *temporal_tieness* que diretamente classifica com baixo custo computacional a força dos relacionamentos em redes sociais temporais de coautoria.

Palavras-chave: Rede Social, Força dos Relacionamentos, Redes Temporais.

Abstract

The study of social ties has lead to build rigorous models that reveal the evolution of social networks and their dynamism. A property related to social ties is the strength of ties, which has been largely explored in different contexts, such as information diffusion, analyses of patterns in communication logs and evaluation of scientific researchers productivity. Specially, analyzing tie strength allows investigating how distinct relationships play different roles and identifying impact at micro-macro levels in the network. In this thesis, the goal is to measure the strength of co-authorship ties in non-temporal and temporal real academic social networks. In summary, the main contributions are: (1) a survey and a taxonomy of social professional networks that contextualize the problem addressed in this work; (2) an analysis of how topological properties relate to the strength of ties in non-temporal social networks, as our results show different topological properties explain variations in the strength of co-authorship ties, depending on the research area; (3) a new metric called *tieness* that is easy to calculate and better differentiates tie strength in different levels in non-temporal co-authorship social networks; (4) an analysis of tie strength dynamism over time by measuring such strength with an existing algorithm in the state of the art and a new one proposed here, called STACY (*Strength of Ties Automatic-Classifier over the Years*), which better identifies strong ties; and (5) a new computational model called *temporal_tieness* that directly classifies the strength of ties in temporal co-authorship social networks with low computational cost.

Keywords: Social Network, Tie Strength, Co-authorship, Temporal Networks.

List of Figures

List of Tables

Contents

List of Acronyms and Abbreviations

Acronym	*Description*
APS	*American Physical Society*
CNARe	*Co-authorship Networks Analysis and Recommendations*
coAfrequency	*Co-authorship Frequency*
CCDF	*Complementary Cumulative Distribution Function*
CPM	*Clique Percolation Method*
CS	*Computer Science*
DBLP	*Digital Bibliography & Library Project*
DCWN	*Dynamic Complex Wireless Networks*
ECDF	*Empirical Cumulative Distribution Function*
fast-RECAST	*fast Random rElationship ClASsifier sTrategy*
LM	*Louvain Method*
MCL	*Markov Cluster Algorithm*
Med	*Medicine*
NO	*Neighborhood Overlap*
per	*Edge Persistence*
RECAST	*Random rElationship ClASsifier sTrategy*
RQ	*Research Question*
SN	*Social Network*
Soc	*Sociology*
SPN	*Social Professional Network*
STACY	*Strength of Ties Automatic-Classifier over the Years*

Chapter 1

Introduction

Social networks (SN) are complex structures that describe individuals in any social context. Theoretically, they can be mapped to graphs where nodes represent the individuals and edges connect nodes according to the individuals relationships. Then, properties and features can be extracted from the graph as well as metrics can be applied to nodes and edges in order to better understand the individuals social behavior [Barabási, 2016]. Finally, there are many interesting applications based on such networks, including (but definitely not limited to) ranking individuals and their groups, link prediction, information diffusion, recommendation and pattern analysis (e.g., [Bagci and Karagoz, 2016; Brandão and Moro, 2017a; Brandão et al., 2013; Freire and Figueiredo, 2011; Hristova et al., 2016; Luna et al., 2013; Seo et al., 2017]).

Furthermore, Social Networks Analysis has evolved from a Social Sciences research area to a Computer Science-based Multidisciplinary research area. Despite the many analyses possible, there are two main aspects to any research at both perspectives (Social and Computer Science): *(i)* how to collect and manage social data, and *(ii)* how to build and analyze the social networks derived from such data.

Also, a specific perspective of evaluation is given by *academic social networks*, in which nodes represent researchers and edges their co-authorships and academic relations. Building the structure of such networks is relatively simple, as the nodes are given by any set of researchers who are connected through their common published work, for example. However, one central aspect of more complex analysis is *the strength of the ties* among researchers, as pairs of researchers have stronger or weaker connections depending on the degree of academic relationship. Such degree of relationship (or tie strength) may be defined according to Granovetter's theory: the ties are *weak* when they serve as bridges in the network by connecting users from different groups, and *strong* when they link individuals in the same group (community) [Granovetter, 1973].

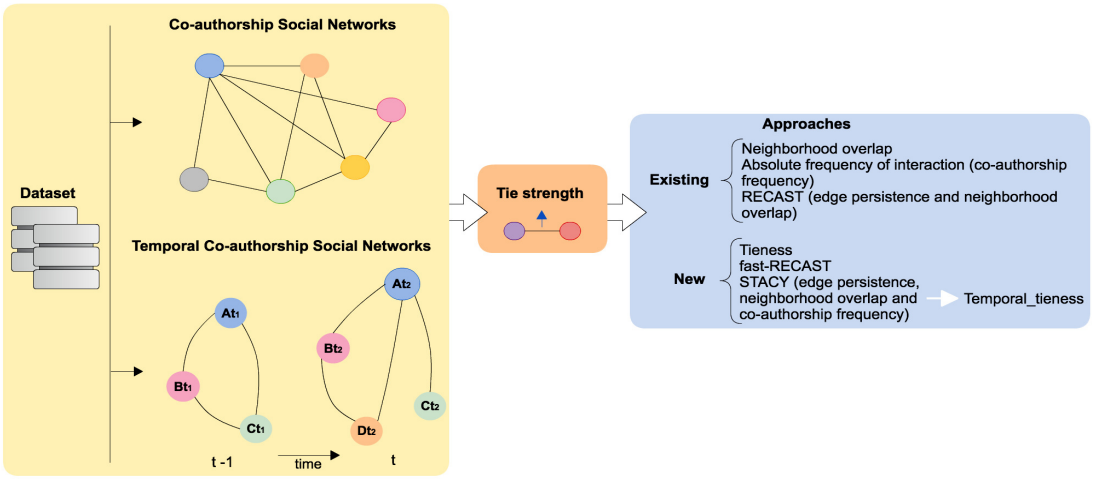


Figure 1.1: Thesis overview: measuring tie strength in non-temporal and temporal co-authorship social networks.

Starting from measuring tie strength in non-temporal (static) co-authorship SNs, this thesis moves forwards to *temporal (dynamic) co-authorship social networks*. Non-temporal co-authorship social networks do not consider time as an aspect of the relationships, whereas temporal social network do. In this context, studying the strength of ties allows to identify patterns of co-authorship over time, to detect aspects that influence it, to determine a limit of co-authorship in a period, and so on.

Tie strength may be measured by a combination of the amount of time, the cooperation intensity and the reciprocal services that characterize the tie [Granovetter, 1973; Rana et al., 2014]. Such strength may also be measured by using the *neighborhood overlap* metric (also known as topological overlap or Jaccard Similarity Coefficient) [Easley and Kleinberg, 2010], a numerical quantity that captures the total number of collaborations between the two ends of each edge. This metric has been used for uncovering the community structure [Li et al., 2012], analyzing structural properties of a large network of mobile phone users [Akoglu and Dalvi, 2010] and measuring tie strength [Brandão and Moro, 2015; Easley and Kleinberg, 2010; Onnela et al., 2007; Pan and Saramäki, 2012; Vaz de Melo et al., 2015]. In this thesis, neighborhood overlap is the base to the development of new tie strength metrics and algorithms.

Figure 1.1 summarizes the overview of this thesis. Given a dataset, we build a non-temporal co-authorship SN and a temporal co-authorship social network by considering the frequency of interactions between pairs of people as the weight of the edges. Note that in temporal social networks, each node is a person and there is an edge between two nodes in a given time if they share any particular relationship in that time. Then, we measure the strength of edges (ties) by considering six approaches: three

existing ones (neighborhood overlap, absolute frequency of interaction and RECAST – Random rElationship ClASsifier sTrategy), and three proposed by us (tieness[1], fast-RECAST and STACY – Strength of Ties Automatic-Classifier over the Years). Note that RECAST (consequently, fast-RECAST) and STACY use an additional feature called edge persistence [Nicosia et al., 2013; Vaz de Melo et al., 2015]. Such feature has been used to measure tie strength in mobile networks [Akoglu and Dalvi, 2010; Raeder et al., 2011], but not applied in co-authorship social networks. Thus, in this thesis we define a computational model and algorithm to automatically infer the strength of the ties in non-temporal and temporal co-authorship social networks.

1.1 Relevance

Extracting and analyzing relevant knowledge from social networks provide many challenges for developers, users and technology. For developers, after collecting data from collaborators, it is necessary to model, store and manage them within databases with proper interface to whatever application uses them. For users, when they need to obtain relevant knowledge from these networks. For technology, which should provide the necessary support for implementation of methodologies. For instance, exploring collaborative relations can improve the accuracy and quality of existing methods that combine bibliometry and academic social analysis. In such a context, this thesis presents two main contributions: (*i*) to define a metric that represents the strength of the relationship between pairs of collaborators in non-temporal co-authorship social network; and (*ii*) to develop an algorithm that automatically classifies the strength of ties in temporal co-authorship social networks. Moreover, we derive a computational model with low computational cost from this algorithm that can be used to measure the strength of co-authorship ties in temporal networks.

Specifically, initial studies of social networks have emphasized the importance of properly measuring the strength of social ties to understand social behaviors [Bruggeman, 2016; Granovetter, 1973; Newman, 2001a]. Also, the study of social ties has been used to build rigorous models that reveal the evolution of social networks and the dynamics of information exchange [Aiello et al., 2014]. More recently, analyzing tie strength has allowed to investigate the different roles of relationships including ranking for influence detection [Freire and Figueiredo, 2011], identify impact at micro-macro levels in the network [Burt, 2010], its influence in patterns of communications [Wiese et al., 2015] and team formation [Castilho et al., 2017].

[1]Tieness is an inspirational neologism based on the quality of being connected, tied.

Despite the importance of analyzing the strength of ties, there are not many studies on evaluating how to measure it in *scientific collaboration networks* (also called *co-authorship networks*). In such networks, nodes are researchers and there is an edge between those pairs that have co-authored at least one scientific publication. Specifically, studying the strength of co-authorship ties may reveal how its behaviors relate to research, and any application based on co-authorship patterns may benefit. For instance, new strength-related metrics could help existing works on measuring research productivity [Chan et al., 2016; Ductor, 2015], ranking researchers [Freire and Figueiredo, 2011] and their graduate programs [Lopes et al., 2011], as well as recommending collaborations [Brandão et al., 2013].

Furthermore, properly measuring the strength of co-authorship ties may help to identify which collaborations are more influent to each researcher. For example, if a researcher A collaborates with other researchers B and C, the strength of ties reveals which one is more important to A, then allowing different studies, such as team formation analyses. Also, researchers that form mostly weak (or strong) ties in the social network may indicate different collaboration patterns. For example, a researcher who has many collaborators through single papers, i.e., that person has collaborated only once with many people.

Formally, we consider two definitions of tie strength in this thesis. The first one is for non-temporal social networks, in which given a non-temporal graph $G(\mathcal{V}, \mathcal{E})$, a tie (i, j) is likely to be strong if it has a high number of common neighbors or a large co-authorship frequency. On the other hand, the tie (i, j) is likely to be weak if it has few common neighbors or small co-authorship frequency. The second definition is for temporal social networks, in which given a temporal graph $G_k(\mathcal{V}_k, \mathcal{E}_k)$, where k is the time step in which a co-authorship occurs, a tie (i, j) is likely to be strong if it is present in G_k for most values of k, and likely to be weak if it is present in G_k for just a few values of k.

One of the first notable studies covering tie strength was published by Granovetter [1973]. He presents the importance of weak ties in SNs for various aspects, such as the spread of information. Since then, the strength of ties has been studied in different contexts with distinct goals [Brandão and Moro, 2015; Gupte and Eliassi-Rad, 2012; Lopes et al., 2011; Silva et al., 2014]. However, few studies have addressed the strength of ties in temporal social networks [Dasgupta et al., 2008; Karsai et al., 2014; Kostakos, 2009; Laurent et al., 2015; Nicosia et al., 2013].

Given the addition of temporal aspects, computing social networks properties and their time-varying behavior constitutes a new challenge. For example, the clustering coefficient of a network in time t_1 is not necessarily the same in time t_2, because some

interactions appear and others perish over time. Also, social networks properties are employed over different domains for distinct purposes, including (but not limited to) recommend collaborators [Brandão et al., 2013; de Sousa et al., 2015; Lopes et al., 2010], viral marketing [Subramani and Rajagopalan, 2003] and graduate programs evaluation [Lopes et al., 2011]. Hence, those properties should represent reality in the best way and consider the time dimension - as time is part of most realities. Furthermore, adding time to any social model may interfere in the process of computing not only properties and features but also complex calculations such as defining communities.

1.2 Hypothesis and Goals

Tie strength can be computed in different ways (for example, amount of time, cooperation intensity and reciprocal services). Then, our main hypothesis is that such strength can be better computed by considering the neighbors of individuals involved in the relationship and combining it with other social networks properties (such as the absolute frequency of interaction and edge persistence). Also, we claim that the time is an important aspect to consider when measuring tie strength. Thus, the main goal of this research is to propose, apply and validate new strategies to measure the strength of co-authorship ties in non-temporal and temporal social networks. These are not easy tasks as there is no ground truth to automatically evaluate the strength of ties metrics. Hence, this general objective can be divided in five specific goals, defined by the following research questions:

- *Research Question 1 (RQ1): How to identify which aspects impact on the strength of co-authorship ties?* The analysis of aspects that affect the strength of collaboration ties is important to better measure and represent such strength. According to Granovetter's theory, aspects related to the strength of ties are the amount of time, the cooperation intensity and the reciprocal services. Indeed, analyzing such strength based on a single absolute value from a metric may provide misleading interpretations. We use statistical techniques to answer this question.

- *Research Question 2 (RQ2): How to measure the strength of co-authorship ties in non-temporal social networks?* We note the strength of ties has long been studied in different contexts, for example, to infer close relationships based on communication logs [Wiese et al., 2015] and to investigate the spread of information in social networks [Miritello et al., 2011]. However, previous works do not analyze the best way to measure the strength of co-authorship ties. In this

thesis, we compare three non-temporal approaches: neighborhood overlap, absolute frequency of interaction (also known as co-authorship frequency) and tieness (our new metric). Note that we compare neighborhood overlap and tieness by analyzing how both differentiate the strength of ties between pairs of researchers.

- *Research Question 3 (RQ3): How to measure the strength of co-authorship ties in temporal social networks?* Considering the temporal aspect to measure social network properties is a challenge due to the dynamism of nodes and their interactions over time. In this thesis, we measure the strength of co-authorship ties in temporal networks by using tree different approaches: RECAST and fast-RECAST (a multiprocessing version of RECAST), and STACY (our new algorithm). Then, we derive a computational model from STACY to measure the strength of ties. Also, we compare RECAST and STACY by investigating how they classify ties that persist over time.

- *Research Question 4 (RQ4): How is tie strength defined for temporal networks?* There are several measures of tie strength for non-temporal networks. For instance, Dasgupta et al. [2008] consider an edge with high call frequency or call volume (weight) as a strong tie, whereas Brandão and Moro [2015] define strong ties as edges with high neighborhood overlap (also known as topological overlap). In temporal networks, such definitions do not hold, since these values may vary over time. In this thesis, we consider that a strong tie characterizes interactions that are likely to appear in the future, whereas a weak tie occurs sporadically.

- *Research Question 5 (RQ5): How much does the strength of ties vary over time?* Nicosia et al. [2013] claim that if two nodes are strongly (or weakly) connected in a time t_1, they will also be strongly (or weakly) linked in a time t_2 where $t_2 > t_1$. Here, we challenge such claim in the context of temporal co-authorship SN. Investigating the strength of co-authorship ties may show how the authors ties relate to research, and any application based on co-authorship patterns may benefit. For example, new strength-related metrics could help existing works on measuring research productivity [Ductor, 2015], ranking researchers [Freire and Figueiredo, 2011] and their graduate programs [Lopes et al., 2011]. Moreover, many studies observe edge features that are good indicators of tie strength, such as edge persistence [Nicosia et al., 2013; Vaz de Melo et al., 2015], neighborhood overlap [Brandão and Moro, 2015; Easley and Kleinberg, 2010; Vaz de Melo et al., 2015] and Adamic Adar [Liben-Nowell and Kleinberg, 2007; Zignani et al., 2016]. In this thesis, we analyze the dynamism of tie strength by observing four

edge classes composed from two of the aforementioned metrics, namely edge persistence and neighborhood overlap (the two metrics are part of RECAST and fast-RECAST). These properties represent the regularity of interaction and the similarity between people in a relationship. Then, we compare the results generated by these two metrics, which give four tie strength classes, with the results of eight edge classes from the combination of edge persistence, neighborhood overlap and co-authorship frequency, which compose STACY.

1.3 Contributions

The main contributions of this thesis are summarized as follows.

1. A new general taxonomy to social networks that helps to identify related work in the area according to their main goal (necessary to all research questions).

2. An analysis of how nine topological properties affect the strength of co-authorship ties when measured by neighborhood overlap (RQ1). Here, we present a multiple regression model to predict the value of neighborhood overlap by using different topological properties.

3. A nominal scale to neighborhood overlap for classifying a tie as weak or strong (RQ2). We define such scale by analyzing the distribution of neighborhood overlap and comparing the values of neighborhood overlap with the absolute frequency of interaction. Then, we present four case studies that show problems of measuring the strength of ties with only neighborhood overlap or absolute frequency of interaction. Easley and Kleinberg [2010] claim neighborhood overlap can be used to measure the strength of ties. We verify that such metric can also be applied in co-authorship SN. However, such metric presents limitations when applied alone.

4. A new metric to measure the strength of ties in non-temporal social networks called *tieness*, resulting from a combination of a modified neighborhood overlap with absolute frequency of interaction (or co-authorship frequency) (RQ2). We also define a nominal scale to tieness based on the values of modified neighborhood overlap and absolute frequency of interaction.

5. A new algorithm called STACY (Strength of Ties Automatic-Classifier over the Years) that automatically classifies the strength of ties in temporal co-authorship social networks (RQ3). We also derive a computational model from STACY named *temporal_tieness*.

6. A set of eight tie strength classes identified by STACY (RQ3). We have characterized each class according to the number of researchers' publication.

7. An analysis of how tie strength is defined over time (RQ4). To do so, we improve an existing algorithm (RECAST) that we call as fast-RECAST. We also use STACY to do such analysis.

8. An analysis of how such strength varies through the years (RQ5). Our results show that most ties, even the strong ones, tend to perish over time. Also, real co-authorship social networks from different research areas have more weak and random ties than strong and bridge ties. Finally, STACY is able of better identify strong ties than fast-RECAST.

1.4 Thesis Organization

The rest of this thesis is organized as follows. We present a taxonomy for existing work on social professional networks in Chapter 2. Then, Chapter 3 states the background for developing this thesis as well as its related work. Chapter 4 characterizes the strength of ties using neighborhood overlap and absolute frequency of interaction. In turn, the results from Chapter 4 are the base for a new metric (tieness) described in Chapter 5. Also, Chapter 6 presents our new algorithm STACY and all the strength of ties analyses to temporal co-authorship social networks. Finally, Chapter 7 concludes this thesis and discusses future work.

Chapter 2

General Taxonomy for Social Networks

The Web has introduced different and new ways in which professionals can easily share their work, publish content, find job opportunities, interact with other professionals, and so on. Besides general purpose social networks (SN), such as Facebook and Twitter, there are online social professional networks (SPN) whose focus is on those activities. Indeed, there are currently more than 20 websites for social professional purposes. Furthermore, as pointed out by Yang et al. [2014], online social networks as a type of communication networks enable straightforward information access. Finally, with a big volume of data available, researchers have used the data from those sites to study SPN characteristics and discover behavioral patterns.

However, there are many challenges in working with social networks [Kleinberg, 2007; Knoke and Yang, 2008]: collecting the data, inferring social process from the data, keeping individual privacy, choosing the best technique to select the data, among others. The social professional networks have an additional challenge that is modeling user emotion. For instance, it is hard to differ if a professional behavior is based on emotional reasons or not. Therefore, in this chapter, we help to identify possible solutions in the literature for these challenges by categorizing existing work according to the social professional network type, goal and stage of development. Note that co-authorship social network is a type of social professional network. Thus, this chapter helps to situate the tie strength research in the state of the art.

Specifically, we define that different research topics address social professional networks and are divided in *issues* and *tasks*. The issues emerge from the need for crawling, storing, managing and treating the data from the networks [Carpineto and Romano, 2012; Chau et al., 2007; Garcia-Molina et al., 2000; Gjoka et al., 2011b;

Han et al., 2011; Harth et al., 2006; Huynh et al., 2012; Mihalcea and Radev, 2011; Rezvanian and Meybodi, 2015; Russell, 2013; Vural et al., 2014; Zaki and Meira Jr, 2014; Zhuang et al., 2005]. Then, the tasks represent the ways that such networks can be analyzed, used, improved and applied in different contexts [Aral and Walker, 2012; Arnaboldi et al., 2016; Brandão and Moro, 2015; Easley and Kleinberg, 2010; Elmacioglu and Lee, 2005; Guille et al., 2013; Kadushin, 2012; Kempe et al., 2003; Kramer, 2010; Murray, 2013; Pak and Paroubek, 2010; Park et al., 2015; Scott and Carrington, 2011; Trusov et al., 2009; Wasserman, 1994; Weng et al., 2010].

In this chapter, we propose a general taxonomy considering issues and tasks as a first-level classification. Overall, the issues are problems *within* social networks regarding their maintenance and usage, whereas the tasks are problems whose solutions benefit from *using* SN data. We describe works related to both, but we focus on research topics related to tasks that are specific to social professional networks (note that issues relate to any type of SN). Hence, we further analyze works whose main tasks are: *(i)* grouping people or items on SPN, *(ii)* recommending people or items, and *(iii)* applying ranking strategies to improve the last two topics.

For the first task, the strategy to form groups is called *clustering* and has been largely presented in different contexts [Ahmed et al., 2014; Backstrom et al., 2006; Blondel et al., 2008; Fortunato, 2010; Girvan and Newman, 2002; Gómez et al., 2015; Keyes, 2015; Palla et al., 2005; Palla et al., 2007; Sales-Pardo et al., 2007; Tang et al., 2007; Xie et al., 2013]. We note that organizing data into groups is one of the most fundamental ways of understanding and learning about patterns intra and inter communities. In turn, considering the variety of information available on social networks that potentially overwhelms users, generating recommendations becomes crucial [Brandão et al., 2013; Lopes et al., 2010; Lops et al., 2011; Schall, 2014; Sharma and Yan, 2013; Yang et al., 2015, 2014; Yu et al., 2016b; Zhang et al., 2014]. At that end, ranking strategies are important to improve clustering approaches [Ahmed et al., 2014; Baumes et al., 2005; Sun et al., 2009] and essential to present the recommendations in a proper order [Brandão et al., 2013; Fouss and Saerens, 2008; Liben-Nowell and Kleinberg, 2007; Lopes et al., 2010; Pu et al., 2012; Schall, 2014; Shani and Gunawardana, 2011; Sharma and Yan, 2013; Song et al., 2011; Xia et al., 2014; Yang et al., 2015; Zhang et al., 2014].

Such taxonomy is a result of a systematic review process performed in five steps [Khan et al., 2003; Kitchenham et al., 2009]: *(Step 1)* Defining questions for the review: How is the publication activity after the first publication in computer science area? What research themes are being investigated and covered? What are the limitations and open problems of current research?; *(Step 2)* Finding relevant work: we initially

search by publications on Google Scholar[1] using the keyword "Social Network". Then, we consider only publications in relevant venues of Computer Science area (specifically, publications from ACM, SPRINGER, ELSEVIER, IEEE); *(Step 3)* Evaluating study quality: we compare different studies grouping them by their main characteristics and analyzing the year of publication, venue, number of citations and abstract; *(Step 4)* Summarizing the evidences: we summarize the evidences in distinct categories that allow to identify the main research topics in SN; *(Step 5)* Interpreting the findings: systematic review findings allow to generate our new taxonomy.

In this chapter, our main contributions are: an overview about social networks and a categorization of social professional networks (Section 2.1); a general taxonomy for social professional networks (Section 2.2); a summary of clustering algorithms (Section 2.3) and recommendation approaches (Section 2.4) applied to social professional networks grouped by their stages of development; a discussion on ranking strategies applied to clustering and recommendation approaches (Section 2.5); and insights over future directions (Section 2.6).

2.1 Main Definitions

In this section, we focus on social professional networks, a specific type of social networks in which the relationships go beyond simply friendship and acquaintances. Thus, we first present an overview of social networks (Section 2.1.1) and then describe the social professional networks with their main types and features (Section 2.1.2).

2.1.1 Social Networks Overview

Any society or social interaction can be mapped to a SN. Then such a network can be analyzed to reveal hidden information of all types, from how a disease has spread and major political views to who the new criminals are.

A social network is defined as a $graph(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes (or vertices) representing individuals (persons, organizations, countries, etc), and $\mathcal{E}$ is the set of edges (or links) constituting their relationships, given by an $n \times n$ matrix in which $e_{i,j} \in \mathcal{E}$ is the (weighted or not, directed or not) relation between nodes i and j [Barabasi, 2002; Newman, 2003; Wasserman and Faust, 1994]. In such definition, the set of nodes and edges can be of a single type, representing a *homogeneous* social network model. For example, all nodes represent persons and all edges their friendship.

[1]Google Scholar: `https://scholar.google.com.br/`

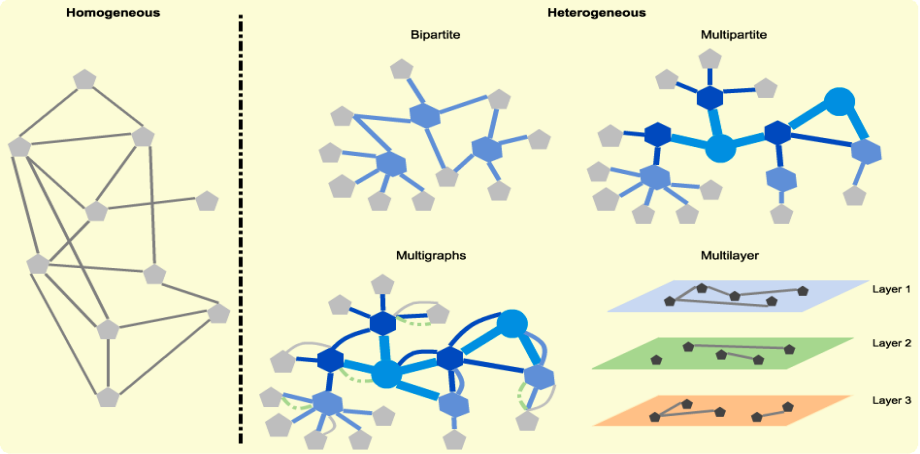


Figure 2.1: Examples of homogeneous and heterogeneous (bipartite, multipartite, multigraphs and multilayers) SN models.

Likewise, having more than one type of nodes defines a heterogeneous social network, e.g., nodes can be people and blogs. The edges between people result from a comment of a person in a post of another person, and the edges from people to blogs represent a person who posted in that blog. There are also edges between blogs, if they have related topics. In both types of social networks (homogeneous or heterogeneous), the edges can be directed or not, and weighted or not. To better illustrate the differences, Figure 2.1 shows generic structures of social networks for homogeneous and heterogeneous models.

Heterogeneous social networks can also be modeled by bipartite or multipartite graphs [Ghosh and Lerman, 2009]. Bipartite graphs are formally defined as $graph(\mathcal{V}_1 \cup \mathcal{V}_2, \mathcal{E})$, in which a node i in $\mathcal{V}_1$ can be only connected to another node j in $\mathcal{V}_2$, i.e. there is no connection among nodes within $\mathcal{V}_1$ (or $\mathcal{V}_2$). Figure 2.1 presents an example of an undirected bipartitie graph, in which, for instance, the nodes represented by pentagons can be women in social networks and the nodes illustrated by cubes can be events that they attended. Likewise, multipartite graphs with n different types of nodes are formally described as $graph(\mathcal{V}_1 \cup \mathcal{V}_2 \cup ... \cup \mathcal{V}_n, \mathcal{E})$ such that for each edge between i and j, $i \in \mathcal{V}_k$ and $j \in \mathcal{V}_{k+1}$ for some $k \in \{1, ..., n-1\}$ [Dawande et al., 2001]. Figure 2.1 also presents a structure of undirected multipartite graphs, in which, for example, nodes represented by pentagons are developers, nodes identified by circles are development projects, and the cubes are a set of commits that a developer did in a project. The thickness of the edges represents the strength of the relationship.

Other models of heterogenous social networks are the multigraphs and multilayer graphs. The former is also interpreted as a combination of single graphs with multiple types of edges [Gjoka et al., 2011a]. Formally, a multigraph is defined as $graph(\mathcal{V}_1 \cup \mathcal{V}_2 \cup$

$...\cup\mathcal{V}_n, \mathcal{E}_1\cup\mathcal{E}_2\cup...\cup\mathcal{E}_m)$, where $\mathcal{V}$ is the set of nodes of n types (representing individuals or items, e.g., events) and belongs to $\mathcal{V}_k...\mathcal{V}_n$ for some $k \in \{1, ..., n-1\}$, and edges between nodes in $\mathcal{V}$ can be of m types and belong to $\mathcal{E}_k...\mathcal{E}_m$ for some $k \in \{1, ..., m-1\}$. For instance, the multigraph in Figure 2.1 may represent a knowledge-sharing network (e.g. Stack Overflow[2]), in which the nodes are different individuals and they are connected by distinct relationships, such as making questions, answering questions, commeting questions or answers. The latter represents networks where nodes are in several layers of the graph, and nodes from a layer are connected to nodes in another one [Bianconi, 2013; Kivelä et al., 2014; Lotero et al., 2016].

In a formal way, multilayer graphs are determined by $graph(\mathcal{V}^\alpha, \mathcal{E}^\alpha, L_{|L|})$, in which $\alpha \in \{1, ..., |L|\}$ and $|L|$ is the number of layers. Each layer may represent nodes from distinct social networks, social status, organizations, among others. For example, each layer of the multilayer graph in Figure 2.1 represents a social network: Layer 1 is GitHub[3], Layer 2 is AngelList[4] and Layer 3 is ResearchGate[5]. Moreover, according to Salehi et al. [2015], multilayer is a generic term that refers to models involving multiple graphs, such as interconnected [Hristova et al., 2016], multiplex (different types of relationships) [Meng et al., 2016], interdependent [Dickison et al., 2016], multisliced [Mucha and Porter, 2010], multidimensional [Ahmed et al., 2016], multiple [Zhang et al., 2016b], multilevel [Wang et al., 2016] networks, and networks of networks [D'agostino and Scala, 2014]. Kivelä et al. [2014] also consider that all these networks are types of multilayer network, but with a few distinct properties, such as the adjacency of nodes, the set of nodes and edges, the number of possible layers, and so on.

There are other models that consider the dynamics of temporal and spatial information [Kivelä et al., 2014]. Considering the temporal aspect, social networks evolve as relationships may appear or disappear over time [Kostakos, 2009]. The analysis of temporal graphs (also called as time-varying networks) may reveal publications patterns [Wang et al., 2016] and users' interactions classes (random or social relationship) [Vaz de Melo et al., 2015], for example. Another dynamic aspect is the spatial information, in which the nodes have locations and the existence of edges are described by those locations [Dale and Fortin, 2010]. Spatial networks (also known as location-based social networks) have been investigated to improve link prediction algorithms [Scellato et al., 2011] and to discovery malicious accounts [Xuan et al., 2016], among others.

Overall, social networks is a very prolific research area. Indeed, looking for "social

[2]Stack Overflow: `stackoverflow.com`
[3]GitHub: `github.com`
[4]AngelList: `angel.co`
[5]ResearchGate: `www.researchgate.net`

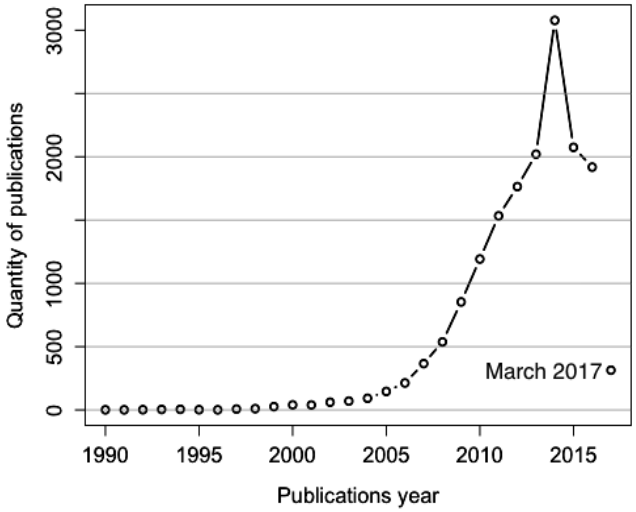


Figure 2.2: DBLP results when searching publications with the term "social network".

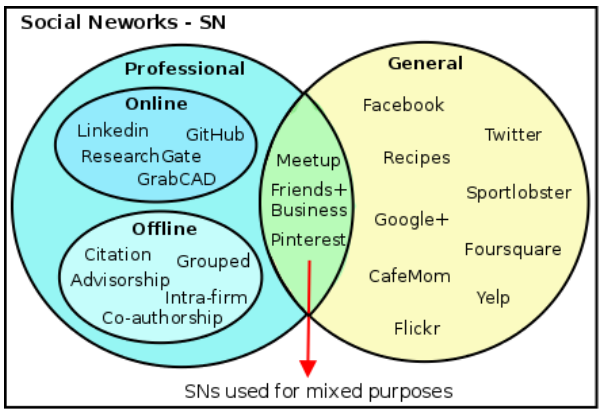


Figure 2.3: Real SN classified by their main purpose.

network" on DBLP[6] returned over 16,380 entries in March 2017. Considering them, Figure 2.2 presents the noncumulative distribution of the increasing amount of publications over the years. Focusing in 2017, there are 314 publications by March 29th.

Among such publications, there are studies addressing different types of social networks with specific goals. Considering social interaction aspects, a high-level classification for such networks is: *online* and *offline*. An online SN is a website or web-based service that allows people to interact with others, i.e., the relationship between users is caracterized by the presence of online communications and not necessarily face-to-face contact [Arnaboldi et al., 2016]. On the other hand, an offline SN is characterized by the absence of online communications to mediate the relationship between users. Also, it is built to represent social relationships in order to allow the study of structural and semantic properties.

[6]DBLP: `www.dblp.org/db`

Table 2.1: Topological properties and concepts on social networks: Given a graph $G(\mathcal{V}, \mathcal{E})$ with a set of nodes $\mathcal{V}$ and edges $\mathcal{E}$, in which i and j are nodes in $\mathcal{V}$.

Property	Definition	Examples of application
Degree (k_i)	$k_i = \sum_{j \in \mathcal{V} \wedge i \neq j} edge(i,j)$, in which $edge(i,j)$ is 1 when there is an edge connecting i to j and 0, otherwise.	Performing decentralized search in networks [Wu et al., 2011].
Path length ($l_G(i,j)$)	$l_G(i,j) = \sum_{i \neq j} distance(i,j)$, in which $distance(i,j) = 0$ if j cannot be reached by i.	Improving collaboration recommendations quality [Brandão et al., 2013; Lopes et al., 2010].
Density (D_G)	$D_G = \frac{Actual\ Edges}{Potential\ Edges}$, where $Potential\ Edge = \frac{\lvert\mathcal{V}\rvert(\lvert\mathcal{V}\rvert-1)}{2}$.	Evaluating knowledge-sharing in social networks [Wiemken et al., 2012], assessing quality of graduate programs [Lopes et al., 2011] and studying communities structures [Newman, 2003].
Community, group or cluster	Subsets of nodes in which the connections among nodes intra communities are dense, but between different communities are less dense.	Interpreting social mobility within the United States [Melamed, 2015], investigating the time dependence and evolution of overlapping communities [Palla et al., 2007] and assessing researchers' productivity [Silva et al., 2015a].
Modularity (Q_G)	$Q_G = \frac{1}{4m} \sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) s_i s_j$, where m is the total number of edges in the network, A_{ij} is the total of edges between i and j, and k_i (k_j) is degree of node i (j). For a dividion of the network into two groups, let $s_i = 1$ if i belongs to group 1 and $s_i = -1$ if it belongs to group 2.	Determining if there are subgroups that should be connected or addressed separately in a research project [Valente et al., 2015].
Common neighbors	$score(i,j) = \Gamma(i) \cap \Gamma(j)$, in which $\Gamma(i)$ ($\Gamma(j)$) is the set of neighboors of i (j).	Being one of predictors to link prediction model [Liben-Nowell and Kleinberg, 2007] and a feature to a community detection algorithm [Xu et al., 2007].
Betweenness ($BC(i)$)	$BC(i) = \sum_{i \neq j \in N} \delta_i(j)$, where $\delta_i(j) = \sum_{w: j \in pred(i,w)} \frac{\delta_{ij}}{\delta_{iw}} (1 + \delta_i(w))$, $pred(i,w)$ is the set of predecessors of w in the shortest paths from i to w, δ_{ij} is the number of shortest path between i and j and $\delta_i(w)$ is the number of shortest path through w.	Defining a new algorithm to detect communities [Girvan and Newman, 2002] and identifying central scholars in database communities [Elmacioglu and Lee, 2005].
Random networks	It is built based on an original network. Thus, a random network has the same number of nodes, edges and empirical degree distribution as the original network. The difference between them is the way that the nodes are connected to each other.	Verifying whether networks have small-world properties [Watts and Strogatz, 1998] and classifying the relationship between users in a mobile social network [Vaz de Melo et al., 2015] by comparing them to random networks.

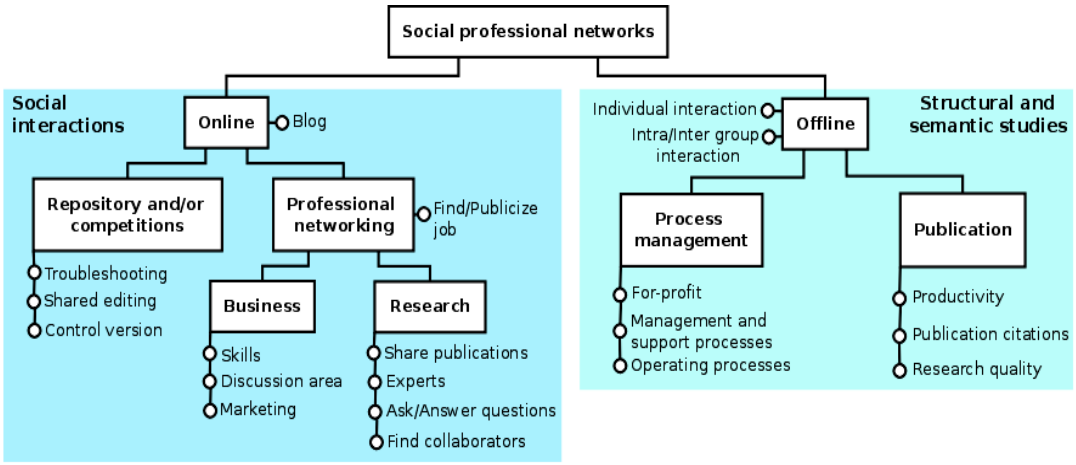


Figure 2.4: Hierarchical diagram of social professional networks types.

Figure 2.3 shows examples of real social networks. Specifically, we have classified them as professional or general, with an intersection representing mixed purposes. For example, Meetup[7] can be used to set up meetings to discuss research or to schedule a school class reunion. Our classification considers the original goal with which a social network was developed. For instance, Flickr has two main goals that are to make photos and videos available and to provide ways for organizing them. This kind of network can be used by non-professionals and professionals (as photographers). Thus, Flickr is not specifically aimed at professionals as LinkedIn. Overall, our focus is social professional networks, and we present examples of general networks only to account for their existence.

Finally, Table 2.1 describes concepts related to social networks that are used throughout this thesis. Note that degree, path length, density, modularity, common neighbors and betweenness represent topological properties from social networks. Thus, they can be computed in different kinds of social networks. Communities and random networks are also concepts related to the network topology. However, there are algorithms to detect communities that also consider semantic properties (theoretical characteristics related to the nodes and edges, for example, geographic location). Such aspects depend on the information available in the social networks.

2.1.2 Social Professional Networks Types

Social professional networks serve different purposes including sharing code and papers, solving doubts, etc. Given their similarities, Figure 2.4 presents the main types of social professional networks organized hierarchically and their main features: the rectangles

[7]Meetup: `www.meetup.com`

describe the main goal of each network type, and each small circle represents networks' features. For instance, *blog* is a functionality of all SPN, and *find/publicize job* is one feature of networks aiming at professional networking. Next, we summarize the SPN types and their features.

Repository and/or competition. This kind of network allows to share code, files and datasets, discuss solutions, discover trends on industry problems, access resources and tools, and/or participate in competitions. Examples include GitHub, Kaggle[8], SourceForge[9], E.Factor[10] and GrabCAD[11]. Regarding research on the topic, Dabbish et al. [2012] study how individuals interpret and use information about other users' actions on code in GitHub. Likewise, Thung et al. [2013] use GitHub to explore the relationship between developers and projects, whereas Bartusiak et al. [2016] predict developers collaboration in the same network. Finally, Narayanan et al. [2011] describe a solution to link prediction problem on Flickr that resulted from a competition in Kaggle social network [Narayanan et al., 2011].

Professional networking. Designed for business or research purposes, its main goal is to provide a platform for professional networking and interaction. Examples are Classemates[12], LinkedIn[13], ResearchGate[14], StartupNation[15] and AngelList. Regarding research, Skeels and Grudin [2009] investigate if such sites enhance productivity, are useful for enterprises and cause issues for new user populations (e.g. stresses from combining personal and professional people). Also, Russell [2013] describes how to crawl LinkedIn data and suggests types of analyses including a histogram of companies in which the contacts have already worked, counting job titles for the technology industry, etc. Moreover, Yu et al. [2016a] evaluate ResearchGate metrics that are used to quantify the performance of researchers and institutions.

Process management. It is a type of offline SN in which the nodes can be companies, employees, sellers, suppliers, customers and/or other business entities that are connected by business relationships. These networks main purpose is usually analyzing interfirm relationships role within a marketing perspective [Heide and Wathne, 2006]. Additionally, Burt [2010] uses industry networks to analyze theories at micro (individuals or small groups) and macro (collective) levels.

[8]Kaggle: `www.kaggle.com`
[9]SourceForge: `sourceforge.net`
[10]E.Factor: `www.efactor.com`
[11]GrabCAD: `grabcad.com`
[12]Classemates: `www.classmates.com`
[13]LinkedIn: `www.linkedin.com`
[14]ResearchGate: `www.researchgate.net`
[15]StartupNation: `startupnation.com`

Publication. They represent relationships among academic entities such as publications, authors, specialists, advisor/advisee and so on. As the others, the academic networks have been largely studied. Examples are co-authorship, citation, advisorship and teachers' professional interactions social networks. Focusing on research, Fu et al. [2014] define different metrics to rank authors, publication venues and institutions. Silva et al. [2015b] consider different properties to analyze researchers' behavior and their publication dynamics in different venues classes. Lee [2015] investigates the multidisciplinary characteristics of technology management research through journal citation network analysis. Brandão et al. [2013] and Lopes et al. [2010] recommend potential collaborators to researchers.

Overall, SN have improved professional activities as job search, contact making, networking, productivity evaluations, etc. Such improvements are clear by the number of existing social professional networks and the large volume of users, data and interactions. Hence, studying such networks has the potential to improve even further their reach and benefits. Moreover, having so many networks requires a proper classification in order to compare them, leading to our proposed taxonomy next.

2.2 General Taxonomy for Social Networks

We propose a taxonomy based on the tasks and issues of social networks. By analyzing the publications on the area, we have identified two main tasks (analysis and application) and two main issues (data acquisition and preparation, and data storage), as illustrated in Figure 2.5 and detailed next.

Data acquisition and preparation. The focus is obtaining the data from social networks or other sources. Current approaches include data from social networks websites [Chau et al., 2007; Gjoka et al., 2011b; Rezvanian and Meybodi, 2015; Russell, 2013], digital libraries [Carpineto and Romano, 2012; Huynh et al., 2012; Zhuang et al., 2005] and the web [Harth et al., 2006; Vural et al., 2014] (i.e., researchers use data from digital libraries and the web to build the structure of social networks).

Often, such data need a pre-processing for cleaning, feature extraction and selection, normalization, transformation, and so on [Kotsiantis et al., 2006]. Thus, techniques of natural language processing have been explored to identify concepts, sentiments, topics and similarities before or after building a social network structure. According to Chen and Ji [2010], Rosenberg and Hirschberg [2007] and Turian et al. [2010], *clustering* may solve natural language processing problems. Its main goal is to

Figure 2.5: Main social networks topics: the tasks refer to using social networks to solve problems, and the issues address problems related to managing social networks.

find a suitable, useful, meaningful and valid organization of nodes and edges. Here, we focus on graph clustering, because entities represented by a social network tend to form clusters, and the number of existing clustering algorithms as well as applications is high. We further describe them considering the SPN context in Section 2.3.

Data storage. Social networks require storing and accessing their data. Specifically for data storage, there are *plenty* of approaches on how to efficiently store data for different purposes. As examples, Garcia-Molina et al. [2000] extensively discuss types of data storage, Corbellini et al. [2017] and Han et al. [2011] describe the background, basic characteristics and data model of NoSQL databases, and Cellary et al. [2014] focus on concurrency control in distributed database systems. Furthermore, there are studies addressing how to deal with the large volume of data that comes from social networks [Almeida, 2013; Yu et al., 2017].

Analysis. The goal is to examine nodes and their interactions in SN towards a specific goal, including to discover influential people [Aral and Walker, 2012; Peng et al., 2017; Weng et al., 2010] and to analyze users and communities sentiments [Kramer, 2010; Pak and Paroubek, 2010; Park et al., 2015; Wang et al., 2017]. Indeed, there is a whole research area on SN analysis that provides algorithms, methods, features to be considered, information diffusion algorithms, models of graph and game theories, etc

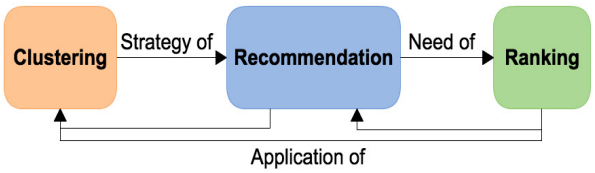


Figure 2.6: Relationship among clustering, recommendation and ranking algorithms.

[Easley and Kleinberg, 2010; Guille et al., 2013; Kadushin, 2012; Scott and Carrington, 2011; Wasserman, 1994]. Tie strength studies are also categorized as this kind of task [Alves et al., 2016; Brandão and Moro, 2015; Brandão et al., 2016; Castilho et al., 2017; Gupte and Eliassi-Rad, 2012; Wiese et al., 2015; Vaz de Melo et al., 2015]. Moreover, we emphasize that clustering algorithms also serve for social network analysis and further discuss them in Section 2.3.

Application. The goal is to use SN to develop methods, features and programs to benefit users. The main types of SN applications include marketing, visualization, recommendation and ranking. Note that marketing [Kempe et al., 2003; Subbian et al., 2017; Trusov et al., 2009] and visualization [Giridhar et al., 2017; Guerra-Gomez et al., 2016; Murray, 2013; de Sousa et al., 2015] may motivate SN analysis or be an application. Therefore, in this chapter, we focus on recommendation (Section 2.4) and ranking (Section 2.5) strategies in social professional network context. Both applications have motivated competitions such as Netflix Prize, CAMRA, the Yahoo! Music KDD Cup 2011 and Kaggle's competitions.

Overall, solutions applied to the first two issues (data acquisition/preparation and data storage) may be adapted to the SN context as they are *not* exclusive for social professional networks. The same cannot be said about the two tasks (analysis and application), which we study specifically in the context of social professional networks. Hence, our research focuses on the latter (i.e., the tasks), and the next sections cover research topics related to clustering, recommendation and ranking algorithms.

In summary, Figure 2.6 shows the relationships for clustering, recommendation and ranking on social professional networks. Specially, clustering algorithms can be part of a recommendation method [Sharma and Yan, 2013] and improved by a ranking function. The recommendation algorithms need ranking strategies, as their results are sorted by considering certain aspects (e.g., relevance). Hence, clustering and recommendation algorithms are applications of ranking strategies. Although we have empirically observed such relationships on social professional networks, we believe they are also valid to other works that address clustering, recommendation and ranking.

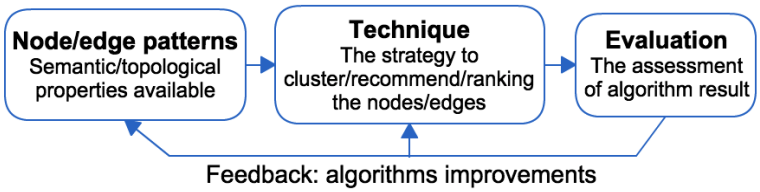


Figure 2.7: Stages of clustering, recommendation and ranking in social network, based on Jain et al. [1999].

2.3 Clustering

Most things in nature and society can be separated into groups according to similarities and properties. A key question is how to group different things in specific contexts. Specially in Computer Science, this question may be rewritten as how to automatically cluster data by using a computer. This task is called *clustering*, whose main goal is to find a useful and valid organization of the input data. There are many algorithms for clustering data. Here, we focus on those that aim to cluster nodes and edges in a graph (i.e., are easily applied to SN).

Clustering techniques have been largely used to detect communities in social networks [Ahmed et al., 2016; Fortunato, 2010; Fortunato and Hric, 2016; Girvan and Newman, 2002; Palla et al., 2005; Tabarzad and Hamzeh, 2017; Xie et al., 2013]. Detecting communities is a task in social networks as it allows the analysis of the interactions and relatedness among users. According to Lancichinetti et al. [2009], nodes in a community are more connected to each other than to the remain of the social network. Note that in the social network context, clusters are also called as community [Girvan and Newman, 2002]. Therefore, we use both terms interchangeably to maintain the nomenclature of the clustering algorithms' authors.

Figure 2.7 presents the stages of clustering in social networks that may also be applied to recommendation and ranking. The following sections detail each level for clustering by discussing existing works.

2.3.1 Nodes and Edges Patterns

This section overviews concepts necessary for understanding clustering techniques. In a graph, nodes and edges of social networks can be clustered for different purposes and by considering different characteristics. We detail and exemplify works divided into three categories as follows.

Spatial patterns. Here, nodes and edges compose clusters according to geographic location. For example, Selassie et al. [2011] propose an algorithm that considers di-

rectional lanes (a property of edge direction) to present visualizations of clusters in a graph. Such clusters represent the amount of GitHub follower data in cities of the United States. Wal et al. [2009] use SN analysis to study inter-firm networks in clusters, regional innovations system and agglomeration economies. Sorenson [2005] also considers SN concepts to investigate how industries in concentrated regions might increase their production. Finally, for the three approaches, the nodes/edges spatial patterns define clusters based on their common geographic regions.

Collaborative interactions. The relationship between nodes is collaborative when they interact to reach a common goal or do a task in an intellectual endeavor. In this context, the main goal of clustering is to group individuals that collaborate. For instance, [Newman, 2001b] analyzes co-authorship SN from biomedical, theoretical physics, high-energy physics and computer science research areas. He shows there is a very strong clustering effect in such scientific communities. Różewski et al. [2015] present a model that combines concepts of knowledge workers to form clusters within an organizational SN. The goal is to increase the competence of knowledge workers' collaborative learning. Kshitij et al. [2015] study how patterns of collaboration in cancer research impact on research policy in India. Such study considers co-authorship SN built from publications in cancer research. By applying a clustering algorithm, the authors reveal the presence of small clusters of researchers connected to one or more highly central researchers.

Different categories. Clusters can group and categorize different types of entities. For example, given a set of nodes and edges representing people in a university, a cluster can represent the class of either professors, students or staff. Considering such a categorization aspect, Ahmed et al. [2014] propose a clustering algorithm to gather LinkedIn users. The clustering criteria are factors significant to users for building groups in such a network, including: area of expertise, job openings, security and time. Melamed [2015] uses eigenspectrum decomposition for community detection in social mobility data. The author considers six categories based on social class (e.g., employers and employees) as nodes, and there is a weighted relation between them when people change the social category.

2.3.2 Clustering Techniques

According to Jain and Dubes [1988], there are six types of clustering techniques: exclusive or nonexclusive, intrinsic or extrinsic, and partitional or hierarchical. Those six are the base for a varied set of clustering techniques. Moreover, there is no "one

size fits all" here, as the choice of clustering type highly depends on the problem and the properties of the input data. For example, if the data have category labels and the problem requires a solution considering specific properties of such categories, the extrinsic approach is more suitable. Traditional clustering algorithms include k-means, k-spectral clustering, DBSCAN and clique percolation [Zaki and Meira Jr, 2014]. As each of them clusters the data according to a different set of features, their resulting clusters for the same input may differ significantly. Next, we discuss all six types of clustering techniques and summarize how they have been applied to SPNs.

Exclusive or nonexclusive. An object may belong to exactly one cluster (exclusive) or to more than one (nonexclusive). Considering exclusive algorithms, Fortunato [2010] overviews exclusive clustering detection on SN. The author provides an example of the division of a co-authorship social network in disjoint clusters by describing the algorithm proposed by Girvan and Newman [2002] (which uses edge betwenness metric to form clusters). Regarding nonexclusive approaches, Xie et al. [2013] compare 14 overlapping clustering algorithms. Among such algorithms, [Palla et al., 2005] propose one that creates overlapping groups by considering the set of nodes' statistical features on SN. Then, Palla et al. [2007] use clique percolation to study the time dependence of overlapping communities (groups) on a co-authorship social networks then characterizing communities evolution.

Intrinsic or extrinsic. Given a set of objects as input to a clustering algorithm, a proximity matrix is computed by measuring the distance between them. In such a context, intrinsic is a kind of unsupervised learning based solely on the proximity matrix to perform classification. On the other hand, besides such matrix, extrinsic uses category labels on the objects as well. Regarding intrinsic techniques, Keyes [2015] applies k-means (unsupervised technique) to geographically cluster LinkedIn users and plots the results on Google Maps[16] or Google Earth[17]. Sales-Pardo et al. [2007] also propose an intrinsic algorithm that forms overlapping clusters by extracting nested hierarchical organization. On the extrinsic perspective, in order to form groups with LinkedIn users, Ahmed et al. [2014] propose a semi-supervised clustering approach that uses partially labeled data. Tang et al. [2007] also use an extrinsic technique to cluster publications into exclusive groups by using characteristics of the publications.

Partitional or hierarchical. A partitional clustering assigns each resulting cluster to a single partition. Then, a hierarchical clustering nests a set of partitions in different levels. Backstrom et al. [2006] study group formation and evolution in co-authorship

[16]Google Maps: `maps.google.com`
[17]Google Earth: `www.google.com/earth`

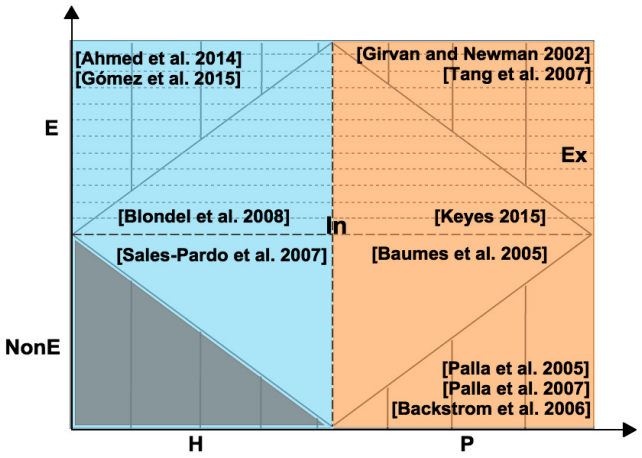


Figure 2.8: Clustering techniques and their overlaps: E - exclusive, NonE - nonexclusive, In - intrinsic, Ex - extrinsic, H - hierarchical and P - partitional.

and conference publications in DBLP. They consider a conference represents a cluster of researchers. There are overlaps among communities and the clusters are partitional. Likewise, Baumes et al. [2005] propose a partitional algorithm that aims to detect overlapping communities based on the clusters density. They experiment in synthetic (random graphs) and real data (DBLP). On the other hand, Ahmed et al. [2014] use hierarchical clustering technique to extract groups from LinkedIn users based on their profiles. Also, Gómez et al. [2015] propose a divisive algorithm based on hierarchical clustering technique for a network of papers and their citations. First, the edges are ordered from the most to the least divisive, then ordered from the most to the least similar. Blondel et al. [2008] propose an algorithm that combines hierarchical technique with modularity optimization to form communities in different networks, including papers and their citations.

Finally, Figure 2.8 summarizes the works that are related to the intersection of the three aforementioned dimensions of clustering techniques. Specifically, each dimension defines a pair of mutually exclusive techniques: exclusive and nonexclusive; intrinsic and extrinsic; partitional and hierarchical. For example, a clustering technique cannot be partitional and hierarchical at the same time. Also, we could not find any work that addresses nonexclusive, extrinsic and hierarchical techniques at the same time on social professional networks. The reason is possibly the increasing time complexity of the clustering algorithm.

2.3.3 Clustering Evaluation

All clustering algorithms should satisfy requirements such as finding clusters with arbitrary shape, dealing with different types of features, treating outliers and noise, handling high dimensionality and insensitivity to the order of input data, scalability, among others [Zaki and Meira Jr, 2014]. Moreover, each algorithm probably requires a different evaluation technique. Specifically, cluster evaluation verifies whether the resulting clusters make sense in a specific context. Even for a dataset with no natural cluster structure, almost every clustering algorithm will find clusters in it [Tan et al., 2006]. Then, a cluster evaluation technique may consider internal or/and external criteria to estimate clusters quality.

For *internal criteria*, two types of metrics are commonly used to estimate resulting clusters quality: distance (how close two objects are to each other) and similarity (how similar/distinct two objects are). The resulting clusters are considered good when the similarity among objects inside the cluster is high (and the distance is low), whereas the similarity among objects from different clusters is low (and the distance is high) [Tan et al., 2006; Zaki and Meira Jr, 2014].

Given a cluster (or community) $C = \{C_1, ..., C_r\}$ representing a clustering of a dataset into r clusters. Let M be the adjacency matrix of the graph and M_{ij} the weight of the edge between nodes i and j. The distance $d(i, j)$ is the dissimilarity between i and j, which can be computed by, for example, edge path ($d(i, j) = 1/M_{ij}$), shortest path distance (using Dijkstra's Shortest Path algorithm) or adjacency relation distance ($d(i, j) = \sqrt{\sum_{k \neq j,i} (A_{ik} - A_{jk})^2}$) [Rabbany et al., 2014]. Then, the most common metrics to internal clustering evaluation criteria are as follows.

Modularity. It calculates the difference between the proportion of edges that are intracluster and the expected such proportion in the case of random distributed edges [Newman and Girvan, 2004; Rabbany et al., 2014]. Let E be the number of edges in the social network, i.e., $E = \frac{1}{2} \sum_{ij} M_{ij}$, then

$$Q_{modularity} = \frac{1}{2E} \sum_{l=1}^{k} \sum_{i,j \in C_l} [M_{ij} - \frac{\sum_k M_{ik} \sum_k M_{kj}}{2E}],$$

in which $l \in \{1, 2, ..., r\}$. The smaller the modularity, the better the resulting clusters, because the distance intracluster is lower than expected.

BetaCV. It is the ratio between the intracluster's edges and intercluster's edges coefficient of variation (CV): $BetaCV = intra\_CV/inter\_CV$ [Zaki and Meira Jr, 2014]. The lower the BetaCV, the better the clustering, because such result shows that intr-

acluster distances are smaller than intercluster distances.

C-index. It measures to what extent the clustering puts together the total number of intracluster edges that are the closest across the identified clusters in a social network [Rabbany et al., 2014]:

$$CIndex = \frac{\theta - min\theta}{max\theta - min\theta},$$

in which $\theta = \frac{1}{2}\sum_{l=1}^{k}\sum_{i,j \in C_l} d(i,j)$, $min\theta$ and $max\theta$ are calculated by summing the $m_1 (m_1 = \sum_{l=1}^{k} \frac{|C_l|(C_l-1)}{2})$ smallest and largest distance, respectively, between every edge. Clustering techniques with smaller C-Index are better, because they return clusters with relatively smaller distances intracluster rather than interclusters.

Silhouette Width Criterion. It measures cohesion and separation of clusters and regards the difference between the average distance of edges interclusters and edges intracluster [Rabbany et al., 2014]:

$$Silhouette = \frac{1}{NE}\sum_{l=1}^{k}\sum_{i \in C_l} \frac{min_{m \neq l} d(i,C_m) - d(i,C_l)}{max\{min_{m \neq l} d(i,C_n), d(i,C_l)\}},$$

in which NE is the total number of intracluster edges and $d(i, C_l) = \frac{1}{|C_l|}\sum_{j \in C_l} d(i,j)$. The higher the silhouette, the better the clustering.

For *external criteria*, the evaluation measures how close the clustering is to the pre-determined data. However, there is no consensus on whether such evaluation is adequate for real data, or only for synthetic data. The clusters are evaluated based on data that was not used in the clustering process. Such data can be a set of pre-classified objects, which is often created by experts. Also, if the clustering goal is to discover new knowledge, the comparison of the resulting clusters with pre-determined data may not necessarily provide the intended result – this comparison may result in the reproduction of known knowledge.

Given a dataset with n points in a d-dimensional space $D = \{x_i\}_{i=1}^{n}$, divided into k clusters. Let $y_i \in \{1, 2, ..., k\}$ represent the ground-truth cluster for each point. The ground-truth clustering is given as $\mathcal{T} = \{T_1, T_2, ..., T_k\}$, in which the cluster T_j denotes all the points with label j, i.e., $T_j = \{x_i \in D | y_i = j\}$. Furthermore, let $C = \{C_1, ..., C_r\}$ represent a clustering of the same dataset into r clusters, resulted from a clustering algorithm. Following Zaki and Meira Jr [2014], all external measures depend on the $r \times k$ *contingency table* $\mathbf{N}$. Such table is composed by clustering C and

the ground-truth partitioning $\mathcal{T}$, defined as $\mathbf{N}(i,j) = n_{ij} = |C_i \cap \mathcal{T}_j|$. Nonetheless, examples of external metrics most used for evaluating clustering in social professional networks include the following.

Purity. It measures the extent to which a cluster C_i has points from only one partition [Zaki and Meira Jr, 2014]:

$$purity = \frac{1}{n} \sum_{i=1}^{r} max_{j=1}^{k} \{n_{ij}\}$$

The larger the purity, the more the result agrees with the ground-truth.

F-Measure. It is the harmonic mean of precision (the proportion of points in C_i from the partition $\mathcal{T}_{j_i}$) and recall (the proportion of points in partition $\mathcal{T}_{j_i}$ in common with cluster C_i) values for each cluster [Chen et al., 2009; Zaki and Meira Jr, 2014]. Given a cluster C_i, let j_i be the partition that has the maximum number of points from C_i, which is, $j_i = max_{j=1}^{k}\{n_{ij}\}$, $F_i = \frac{2n_{ij_i}}{n_i + mj_i}$, in which $mj_i = \mathcal{T}_{j_i}$. The F-measure for the clustering C is the mean of F-measure values of each cluster: $F = \frac{1}{r} \sum_{i=1}^{r} F_i$. A clustering is considered perfect when the value of F-measure is 1.

Entropy-based measures. They measure the homogeinity (the class distribution intracluster should be assigned to a single class) and completeness (it is symmetrical to homogeinety, i.e., all points member of a single class must be assigned to a single cluster) of the resulting clusters [Dom, 2002; Meilă, 2007; Rosenberg and Hirschberg, 2007; Zaki and Meira Jr, 2014]. Examples of most common entropy-based measures are: V-Measure [Rosenberg and Hirschberg, 2007] - an entropy-based measure that considers homogeneity (each cluster has datapoints of a single class) and completeness (all the datapoints from a given class are elements of the same cluster); Q_0 [Dom, 2002] - a measure for non-hierarchical clustering that uses conditional entropy - *H(number of classes/number of clusters)* - to calculate the correctness of a clustering algorithm result; variation of information (VI) [Meilă, 2007] - a distance metric that uses entropy for comparing different clusters: *VI(number of classes, number of clusters) = H(number of classes/number of clusters) + H(number of clusters/number of classes).*

2.3.4 Clustering Overview on Social Professional Networks

In summary, clustering techniques have been extensively applied to social professional networks for different purposes. We could identify three main directions: using social professional networks to validate a proposed algorithm [Baumes et al., 2005; Blondel

et al., 2008; Gómez et al., 2015; Palla et al., 2005; Selassie et al., 2011]; applying the existing clustering algorithms to analyze social professional networks [Backstrom et al., 2006; Keyes, 2015; Kshitij et al., 2015; Melamed, 2015; Newman, 2001b; Palla et al., 2007; Sorenson, 2005; Wal et al., 2009]; and designing algorithms, models and approaches exclusive for SPNs [Ahmed et al., 2014; Różewski et al., 2015].

At that end, metrics to evaluate clusters in social professional networks are the same for evaluating clusters in other SN. Although evaluating if a clustering algorithm forms clusters in a consistent way is important, such evaluation is not common in works that address social professional networks.

2.4 Recommendation

Social professional networks provide information from entities and interactions between them. Due to the large volume of data, it is often hard to find an expert on a particular subject, an online media (e.g. book, video), etc. In this context, using recommendation systems is important to find adequate information from social networks [Sun et al., 2015]. The main goal of such systems is to predict the preference of a user for an item (term used to define things that a system recommends to a user) or people, i.e. recommend or suggest items/people that are relevant to a user. The recommender systems have various practical applications. For example, helping parents of children with Autism Spectrum Disorders to find a community of related parents based on assessment (from clinical) of such disease [Song et al., 2011], LinkedIn users to join a group [Sharma and Yan, 2013], project managers in GitHub to get potential reviews to a new pull-request (when a user pushes changes to a repository in GitHub) [Yu et al., 2014], objects (e.g. movies or songs) to receive relevant tags [Belém et al., 2016], users to reach their favorite songs [Fujino et al., 2017], and researchers to find collaborators [Brandão et al., 2013; de Sousa et al., 2015]. There are many techniques and algorithms to recommend a particular item. Here, we focus on those that recommend based on social networks data.

As already mentioned, Figure 2.7 also shows the stages of developing a recommendation algorithm. We detail each stage next.

2.4.1 Topological and Semantic Features

Social professional networks have different properties that can be represented by topological or semantic features. The topological features (some described in Table 2.1) refer to the structure of nodes and/or edges on the social network. For example, there

are SN in which nodes and edges represent (respectively): developers and their commitment to the same project (GitHub); musicians/bands and their collaboration in composing or performing a similar music [Cano et al., 2006]; and authors and their co-authored papers [Liben-Nowell and Kleinberg, 2007]. The topological features used to analyze such networks and propose recommendations include: common neighbors, degree distribution and node proximity.

On the other hand, semantic features represent theoretical concepts related to the existence of nodes and/or edges and interactions among them. For instance, Yu et al. [2014] consider semantic feature to recommend reviewers of incoming pull-requests in GitHub. The semantic feature (extracted from developers' comments on GitHub projects) represents the common interest of each reviewer. Additionally, Chen et al. [2015] and Xia et al. [2014] propose algorithms (AVER and MVCWalker, respectively) to recommend collaborators by considering semantic features, such as co-publication frequency, weights of relations and researchers' academic level in AVER, and co-author order, latest collaboration time and number of collaboration in MVCWalker.

There are also approaches that combine topological and semantic features. For example, Brandão et al. [2013] and Lopes et al. [2010] propose algorithms (Affin and CORALS) to recommend collaborators. Affin combines shortest path (topological feature) with researchers' institutional affiliation, whereas CORALS merges such topological feature with the research area of researchers. In addition, Schall [2014] presents a new algorithm for recommending relevant users to follow on GitHub by considering semantic (user behavior and shared interests) and topological (follower degree, triadic closure and network centrality) features.

2.4.2 Recommendation Techniques

Recommender systems use a large range of techniques. According to Adomavicius and Tuzhilin [2005], such systems are classified as: content-based, collaborative filtering and hybrid, as detailed next.

Content-based technique. This relies on prior preferences or behavior of a user to recommend items/people [Lops et al., 2011]. For example, the algorithm recommends an article on a social network to a user if the article has features similar to others that the user has read. Sharma and Yan [2013] combine pairwise for preference learning with such technique to recommend groups to LinkedIn users. Also, Zhang et al. [2014] consider user's behavior on GitHub to recommend open source projects for developers.

Collaborative filtering technique. It suggests items/people to a user considering

items/people previously rated by or related to others. For example, given two users with similar taste, a collaborative filtering technique considers the liked articles of one user and recommends such articles to the other user. Also, Schall [2014] uses the collaborative filtering approach to recommend users to follow on GitHub. In such context, if a user A watches the same repository as a user B, then A might follow B because both have similar interest. Finally, Yang et al. [2014] present a literature review of recommender systems that use collaborative filtering approaches based on social interactions between users in online social networks.

Hybrid technique. The most common hybrid techniques combine the content-based and collaborative approaches, which helps to avoid the limitations of both methods such as recommending item/people already used or known by a user, using only imprecise content analysis, considering just similar users' ratings, and so on. For example, Yang et al. [2015] propose a hybrid approach that combines research topic network, researcher collaboration network and institution network with SVM-Rank to recommend collaborators. Affin [Brandão et al., 2013] and CORALS [Lopes et al., 2010] are also hybrid algorithms, because they consider researchers' properties and the researchers' relations with others.

2.4.3 Recommendation Evaluation

Evaluating the effectiveness of recommender systems and the quality of the resulting recommendations is a hard task, mainly for three reasons: different algorithms may have divergent performance on distinct datasets; the goals for which an evaluation is performed may differ; and the recommendations that are "good" for a set of users are not necessarily good for another set. Evaluating the quality of the generated recommendations means to identify how "good" the recommendations are regarding different criteria, such as diversity, novelty, accuracy, coverage, serendipity, utility, and so on. Many studies have focused on evaluating only the accuracy (i.e. evaluating the genereted recommendations regarding a ground-truth and using metrics as mean absolute error, recall and precision [Ge et al., 2010; Shani and Gunawardana, 2011]) of recommendations, for example, [Chen et al., 2015] and [Lopes et al., 2010]. Although having a high accuracy is important, it is also insufficient to ensure the quality of the recommendations [Belém et al., 2016; Brandão et al., 2013; Fouss and Saerens, 2008; Shani and Gunawardana, 2011].

Likewise, others describe evaluation metrics and strategies. For instance, Wu et al. [2012] describe 11 metrics divided into based on recommender algorithms and depending on recommender algorithms (further divided into system's angle and users'

Table 2.2: Recommendation Summary.

Characteristics		Publications
Features	**Topological**	Cano et al. [2006], Liben-Nowell and Kleinberg [2007]
	Semantic	Chen et al. [2015], Yu et al. [2014] and Xia et al. [2014]
	Both	Brandão et al. [2013], Lopes et al. [2010] and Schall [2014]
Techniques	**Content-based**	Fujino et al. [2017], Sharma and Yan [2013] and Zhang et al. [2014]
	Collaborative	Schall [2014] and Yang et al. [2014]
	Hybrid	Brandão et al. [2013], Lopes et al. [2010] and Yang et al. [2015]
Evaluation	**Accuracy**	Chen et al. [2015] and Lopes et al. [2010]
	Different Aspects	Belém et al. [2016], Brandão et al. [2013], Fouss and Saerens [2008], Pu et al. [2012], Shani and Gunawardana [2011] and Wu et al. [2012]

angle). Shani and Gunawardana [2011] differentiate offline experiments (using a pre-collected data set to simulate the behavior of users) from online ones (real users interact with the recommender system, which allows to evaluate the quality of the recommendation according to their real behavior). The authors also describe 14 evaluation metrics. Besides presenting metrics, Pu et al. [2012] show how to combine them in order to present better recommendations to a user. Finally, we emphasize that the surveys refer to general recommender systems. Therefore, such metrics can be applied to evaluate recommendation algorithms based on social professional network data as well.

2.4.4 Recommendation Overview on Social Professional Networks

There are different algorithms and methodologies to do recommendation on social professional networks. Despite their differences, their goal is one: to improve professional productivity, quickness and agility. For example, the algorithms may recommend a person to improve the quality of publication reviews or an item (for instance, a software project) to help a professional to find in which to work on.

Moreover, evaluation metrics of recommendation algorithms on general social networks can also be applied to evaluate algorithms on social professional networks. Nevertheless, the choice of such metric depends on the goal of the recommendation. For instance, if the main goal is to recommend different items, a diversity metric should be used on the evaluation.

Finally, Table 2.2 summarizes the publications that cover recommendations techniques on SPNs. We also summarize the evaluation strategies regarding the use of accuracy metrics and other different metrics (e.g., novelty and diversity).

2.5 Ranking applied to Clustering and Recommendation

The main goal of ranking is to define importance weights to distinct objects in order to make finding the relevant ones easier. Hence, different research areas have extensively investigated ranking functions. In general, those functions consider ranking models, such as vector space, probabilistic information retrieval, statistical language, graph-based or set-oriented [Chaudhuri et al., 2004; Dom et al., 2003; Harman, 1992]. Figure 2.7 summarizes the stages of a ranking algorithm development. Once again, here, we focus on ranking functions applied to clustering and recommendation algorithms in social professional networks.

Ranking on clustering approaches. Combining ranking algorithms (or techniques) with clustering techniques aims to improve the quality of the resulting clusters [Ahmed et al., 2014; Baumes et al., 2005; Delis et al., 2016; Sun et al., 2009; Zhong et al., 2017]. For instance, Sun et al. [2009] propose an algorithm called RankClus to cluster and rank conferences and authors. The ranking function considers numbers of papers accepted by a conference or published by an author and is used in the clustering process. Furthermore, Ahmed et al. [2014] improve the quality of a clustering algorithm by defining a ranking function that uses quantitative constraints to represent the area of expertise of LinkedIn users.

Ranking on recommendation algorithms. The main goal of ranking is to sort the resulting recommendations by relevance [Brandão et al., 2013; Fouss and Saerens, 2008; Liben-Nowell and Kleinberg, 2007; Lopes et al., 2010; Pu et al., 2012; Schall, 2014; Shani and Gunawardana, 2011; Sharma and Yan, 2013; Song et al., 2011; Tang et al., 2016; Xia et al., 2014; Yang et al., 2015; Zhang et al., 2014]. In general, a score is attributed to each recommendation as defined by a ranking function. For instance, Liben-Nowell and Kleinberg [2007] define a ranking function for each topological feature used as a predictor of pairs of collaborators that should work together. In order to recommend relevant projects to developers, Zhang et al. [2014] rank the projects by using cosine distance to measure the similarity between two projects.

2.6 Future Directions

In this section, we provide insights into future directions and trends in techniques based on social professional networks.

Considering semantic and multiple features. Here, the main idea is to treat individuals, research and expertise areas, research and working groups, publications and codes by considering their particular properties. For example, features that are important to one research group, such as number of publications in conferences, may not be relevant to other group in which publishing in journals is the ultimate goal. In this context, a way to distinguish groups is using multiple features and attributing different weights to such features according to the groups properties. By doing so, applications and methods may better represent the reality. In this thesis, we follow this direction by analyzing how different features can be used to measure tie strength in Chapter 4 and then proposing a new metric that results from a combination of two features in Chapter 5.

Grouping people. In most social networks, people tend to form groups, which also happens in social professional networks. However, most of recommendation and ranking approaches are for a single individual in SPN. Indeed, we have found few publications that do *group* recommendation in this context (such as [Sharma and Yan, 2013]). Additionally, Salehi-Abari and Boutilier [2015] propose an approach to do group recommendation and inference that can be applied to social professional networks. Hence, there are still open issues for strategies to recommend groups to people or people/items to groups and ranking groups.

Identifying data veracity. A problem of using the features of social professional networks is identifying if they indicate the truth or not. The development of methods to do such distinction is important, especially for online social professional networks. The data available on such networks have been used to develop productivity indicators, rank researchers and reviewers, evaluate graduate programs and conferences, and so on. Thus, such data have to be reliable.

Capturing nonprofessional social process. In social professional networks, another challenge is to infer if a user behavior is based on nonprofessional information (feelings and emotions) or not. A possible solution is to combine the SPN data with the data from other source, for instance a friendship social network. Then, more information will be available to better infer a user's feelings and behavior. For instance, consider applications that recommend people for professional activities (evaluating papers, hiring committees, reviewing code, etc). A person may not be a good recommendation solely based on the feelings of the others involved in the task.

Temporal social professional networks. Temporal networks allow to understand the dynamics of the properties from nodes and relationships over time [Atzmueller

et al., 2016; Kostakos, 2009; Wang et al., 2016]. According to Nicosia et al. [2013], the precise temporal ordering of the edges essentially influences the notion of node adjacency and reachability in such networks. Hence, concepts and metrics designed and applied to analyzing static social networks have to be adapted and extended to time-varying networks. For instance, Casalnuovo et al. [2015] address temporal modeling on GitHub for analyzing the socialization between developers as a predecessor to enrolling a project. Also, they evaluate how the expertises of past experience and social aspects of prior interactions to integrants of a project influence productivity at the start and in the long term. In this context, there are many open aspects that can be investigated in temporal social networks. In Chapter 6, we help to fill this gap by proposing a new algorithm to measure tie strength in temporal co-authorship social networks.

2.7 Concluding Remarks

A social professional network is an important environment that can reveal patterns from professional interactions and behavior. Such patterns may be used to improve the performance of developers, authors, reviewers, and so on. Also, they may help professionals find relevant information. Furthermore, the patterns may be applied to evaluate the quality of research groups, open source projects, relationship between co-workers and how it influences in their production, project reviews, and so on. Therefore, the data available in SPNs provide valuable information and have many uses and practical applications.

In this chapter, we presented a survey regarding social professional networks. We described a general taxonomy to social networks considering the tasks (the use of social networks to solve problems) and the issues (problems that emerge when dealing with social networks) as a first-level classification. Also, we defined the types of social professional networks and further exemplified them. Next, we focused on works that cover clustering and recommendation algorithms, and ranking functions applied to those two problems. We have also identified relationships among clustering, recommendation and ranking approaches, which reveal the importance of one to each other. Then, we concluded by presenting future directions on open problems.

Finally, important challenges to solve SPNs tasks are related to the *data*: how different features properly represent entities (e.g., users, communities, research areas, software projects), how data veracity can be measured and identified, and how data from different sources can be combined. Therefore, research on data management is still crucial for the success of social networks and their applications.

Chapter 3

Background

The previous chapter presented existing work on social professional network, including a general taxonomy to social networks. Specially, we focused on works that address clustering, recommendation and ranking on social professional networks. Now, we first explain the statistical approaches applied in this work. Finally, we detail previous work on tie strength specially in non-temporal and temporal social networks.

3.1 Basic Concepts for the Experimental Analyses and Evaluations

Here, we focus on describing the methodology to develop our research. Hence, we detail the co-authorship social networks (Section 3.1.1) and the methods that help to analyze and measure the strength of ties (Section 3.1.2).

3.1.1 Co-authorship Social Networks

In order to analyze the strength of ties, we consider co-authorship social networks, which represent relationships extracted from publications. We have initially built three social networks using publications available on Lattes[1] (further described in Chapter 4). Each network represents a different research area: medicine, computer science and sociology. Then, we built large academic social networks from three different areas of expertise. The areas and their datasets are: (*i*) Computer Science given by DBLP[2] (collected in September 2015); (*ii*) Medicine by PubMed[3] (April 2016); and (*iii*)

[1]Lattes: `http://lattes.cnpq.br`
[2]DBLP: `http://dblp.uni-trier.de`
[3]PubMed: `http://www.ncbi.nlm.nih.gov`

Table 3.1: Datasets and their basic statistics and information.

Dataset	Number of nodes	Number of edges	Period
DBLP Articles	837,583	2,935,590	2000 to 2015
DBLP Inproceedings	945,297	3,760,247	2000 to 2015
PubMed	443,784	5,550,294	2000 to 2016
APS	180,718	821,870	2000 to 2013

Physics by APS[4] (March 2016). For DBLP, we split it in two datasets: DBLP Articles and DBLP Inproceedings. For PubMed (a US national library of Medicine National Institute of Health that comprises biomedical publications), we consider publications from the top-20 journals classified by h-index. For APS (American Physical Society), we consider a sample dataset with its journal publications. Then, we build a co-authorship SN for each dataset with features shown in Table 3.1.

These three large co-authorship social networks are used in the experiments performed in Chapter 5 and 6. We emphasize that in Chapter 6, we consider the time of the co-authorships in order to catch the temporal aspect of the social networks. The datasets supporting the analyses of these chapters are publicly available at `http://www.dcc.ufmg.br/~mirella/projs/apoena/`.

3.1.2 Approaches to Tie Strength Analyses and Measures

Here, our goal is analyzing how the strength of ties can be measured by using different metrics and algorithms. Also, we study how topological properties affect the strength of ties measured by neighborhood overlap and absolute frequency of interaction (co-authorship frequency) and how tie strength dynamism over time is. At that end, we use regression and correlation analysis, quartiles and other statistical methods. In this section, we only present general concepts to such statistical techniques. Further details are presented in Chapters 4, 5 and 6.

Regression analysis. The goal of this statistical process is to identify the relationship between one (or more) independent variable (its value is not changed by other variables also considered in the analysis) and a dependent variable (the opposite meaning of independent variables) [Jain, 1991]. Also, there are simple regression models and multiple regression models. The former describes the relationship between a single dependent variable and a single independent variable. The latter predicts the value of a dependent variable from the value of two (or more) independent variables. Given the relationship between variables, a model is defined as a hypothesis, and estimations of

[4]APS: `http://www.aps.org`

the model parameter values are considered to define an estimated *regression equation*. Then, various tests can be applied to determine how satisfactory the model is. If the model is satisfactory and given values for the independent variables, the value of the dependent variable can be predicted by the estimated regression equation.

Correlation analysis. This statistical technique relates to regression analysis in the sense that both handle the relationship among variables. In general, correlation coefficients are a measure of a monotone (linear or not) association between variables [Chok, 2010]. Values of such coefficient vary from -1 to +1. A correlation coefficient of +1 represents a perfectly relationship in a positive monotone sense, whereas -1 indicates an ideally relationship in a negative monotone way, and 0 means the absence of monotone relationship between variables [Cohen, 1988]. Furthermore, the most commonly used correlation coefficients are Pearson (it measures the linear relationship between two variables), Spearman's rank and Kendall's tau (both are appropriated to a non-linear relationship between two variables) [Chok, 2010].

Arithmetic mean. It is a central tendency measure (a single value that summarizes a distribution) that represents the average of a sequence of numbers. All values x in a sequence are added up and then divided by the total number of observations n: $\bar{x} = \sum x/n$. According to Manikandan [2011], the advantages are: it is a good representation of the data (because it uses every value in the data), it keeps out the variations between different distributions when compared to others central tendency measures and it is very related to standard deviation (detailed in this section). On the other hand, it disadvantages include the sensitivity to extreme values or outliers (specially when a sequence of numbers is small and skewed), absence of a meaningful value for nominal or nonnominal ordinal data, and similar values to distributions with many repeated or approximate numbers.

Median. It is also a central tendency measure given by a number that is in the middle position of an ascending or descending sorted distribution of the input set. When a sequence of values with n numbers is odd, the median is the number in the $n/2^{th}$ position. If it is even, it is given by the mean of $n/2^{th}$ and $(n/2+1)^{th}$ value. Following Manikandan [2011], advantages of median are: it is easy to compute; outliers and skewed data do not distort its value; and ratio, intervals and ordinal scales can be represented by median. Disadvantages include the representation of observations is not precise and does not consider all values in the data, applying it in further mathematical calculation is hard, and joining two distributions looses the ability to represent the individuals medians of the joined group.

Quartiles. It is a kind of quantile[5], i.e. cutpoints in which a sequence of numbers is divided into parts with equal size. Specifically, quartiles are the three points that divide a ranked distribution in four equal regions. The first quartile is the middle number between the smallest one and the median of a distribution. The second one is the median of a distribution. Then, the third quartile represents the middle value between the median and the highest value in a sequence of numbers [Brase and Brase, 2012]. According to Sharma [2012], the advantages of quartiles are: the simplicity to compute, independence of extreme values or outliers, an appropriate measure of variation for a distribution, and the adequateness in case of skewed distributions. The disadvantages are: the value being based on the 50% observed data makes a not good measure when not considering all data (because it considers only 50% of the data), it presents high variance of values for different samples, and its value is not affected by intermediate numbers in range of the middle 50% of the distribution. Even with such problems, quartiles still allow to analyze the distribution (for example, minimum and maximum value) of strength of ties.

Variance. It is a measure of how a data is distributed about the mean or expected value [Sharma, 2012]. Here, we use the sample variance to estimate the population variance because the data being used is sample data. Thus, given a distribution with x elements of size n, the variance is $s^2 = (\sum x^2/n - 1) - ((\sum x)^2/n(n-1))$. The advantages of variance are: it considers all data of the distribution in its calculation and it is a prerequisite for the computation of many stable measures (for example, standard deviation). However, the disadvantages are: its interpretation is hard, the value is squared, and extreme values can influence the resulting value. Overall, variance allows to analyze how distance the strength of each tie are from the mean.

Standard deviation. It is a solution to interpret the squared unit of variance [Sharma, 2012]. The standard deviation is a positive square root of variance: $s = \sqrt{s^2}$. According to Sharma [2012], the advantages are: the value is based on every value on the distribution, its value is less affected by variations of different samples than other measures, it allows to calculate the combined standard deviation of two or more distributions, and it is useful in further statistical calculation (for instance, comparing skewness or correlations). On the other hand, disadvantages are: the calculation is harder when compared to other measures of variation, and more weight is given to extreme values and less to those close to the mean. Despite its disadvantages, it still reveals important characteristics about variation of the strength of ties.

[5]Quantile R project: `http://stat.ethz.ch/R-manual/R-devel/library/stats/html/quantile.html`

Additionally, we have tried to apply factorial design methods to understand how topological properties explain variations in neighborhood overlap and absolute frequency of interaction. However, we could not identify an appropriate model because: *(i)* each topological property is numeric and defining representative levels to the factors of the factorial design is hard; *(ii)* when we have tried to define levels, not all properties had values for a specific combination of all levels; and *(iii)* the degree of freedom is very high for error.

3.2 Related Work over Tie Strength

In this section, we first overview the strength of ties and how to calculate it (Section 3.2.1). Then, we focus on tie strength in temporal social networks (Section 3.2.2).

3.2.1 Tie Strength Overview

Many studies address tie strength in social networks [Brandão and Moro, 2015; Bruggeman, 2016; Castilho et al., 2017; Goulas et al., 2015; Granovetter, 1973; Wiese et al., 2015]. Following Granovetter [1973]'s theory, ties are weak when they serve as bridges in the network by connecting individuals from different groups, and strong when they link individuals in the same group. Moreover, the strength of a tie has been studied in different domains with diverse goals. For instance, measuring the strength of co-authorship ties [Brandão and Moro, 2015], work ties [Castilho et al., 2017], weak ties [Granovetter, 1973], friendship ties [Seo et al., 2017; Zignani et al., 2016] and contact (through calls and SMS) ties [Wiese et al., 2015]. Such studies contextualize the importance of measuring tie strength in an appropriate way. Overall, distinct relationships play different roles in social networks and should be distinctly qualified as well through (for example) their strength. Indeed, studies show that the strength of ties have (for example) large impact at micro-macro levels in the network, depending on their weight, and influence the patterns of communications [Brandão and Moro, 2015; Granovetter, 1973; Zignani et al., 2016].

Tie strength can be calculated by considering topological and/or semantic properties in the social network. First, topological properties capture structural characteristics of the graph that constitutes the social network [Zaki and Meira Jr, 2014]. For instance, Brandão and Moro [2015] use neighborhood overlap to measure tie strength in co-authorship networks. Second, semantic properties catch non-structural characteristics of nodes and edges in the social network. For example, Gilbert and Karahalios

Table 3.2: Given two nodes i and j, there are different metrics that can be used to measure the strength of ties.

Description	Equation	Publications
Adamic-Adar coefficient	$\sum_{k \in N(i) \cap N(j)} \frac{1}{log N(k)}$, where $N(i)$ refers to the neighbors of a node i.	[Kahanda and Neville, 2009], [Zignani et al., 2016]
Clustering coefficient	$\frac{2e_i}{(k_i(k_i-1))}$, where e_i is the number of edges between all neighbors of i and k_i is the number of neighbors of i.	[Brandão and Moro, 2015], [Zignani et al., 2016]
Collaboration weight	$\sum_p \frac{\delta_i^p \delta_j^p}{n_p - 1}$, where δ_i^p is 1 if node i collaborates in a work p and zero otherwise, n_p is the number of collaborators in a work p and all single-collaborated work are excluded.	[Newman, 2001a], [Pan and Saramäki, 2012]
Frequency or interaction intensity	$w_{i,j}$ represents the absolute number of interaction between i and j.	[Onnela et al., 2007]
Neighborhood overlap or Jaccard Index or Topological Overlap	$\frac{\lvert X_{c_i} \cap X_{c_j} \rvert}{(\lvert X_{c_i} \cup X_{c_j} \rvert - (i,j \ themselves))}$, where X_{c_i} represents the neighbors of node i, and X_{c_2} the neighbors of j.	[Brandão and Moro, 2015], [Easley and Kleinberg, 2010], [Vaz de Melo et al., 2015], [Onnela et al., 2007], [Pan and Saramäki, 2012]
Normalized direct social weight	$\frac{\sum_{\forall \lambda \in \Lambda_{i,j}} \omega(i,j,\lambda)}{\sum_{\forall k \in Ni} \sum_{\forall \lambda \in \Lambda_{i,k}} \omega(i,k,\lambda)}$, where $\lambda \in \Lambda$ represents all types of interactions (e.g., number of co-authored papers or shared projects) between i and j.	[Zuo et al., 2016]

[2009] define a relationship as weak or strong on Facebook by considering features available on interaction history, such as days since first or last communication time, and inbox messages exchanged. Third, combining both is also possible. For example, Zignani et al. [2016] use interaction-graph properties (topological) and temporal features (semantic) to predict link strength. Likewise, Kahanda and Neville [2009] measure tie strength on Facebook by mapping semantic features (picture postings and groups, the interaction among users, users' gender and interests) and topological properties (node degree and number of shared neighbors). Also, Seo et al. [2017] measure the friendship strength considering communication information between users, personal similarity and group similarity; all three calculated by using semantic and topological properties.

Each property type provides advantages and disadvantages over each other. Generally, topological properties can be calculated in any social network. Also, they usually have relatively low computational time cost. However, in some situations, considering only topological properties to measure the strength of ties may not be very accurate, because relationships may be influenced by aspects not only related to network structure. Hence, semantic properties may improve the accuracy of approaches that measure tie strength since they consider aspects related to the context of nodes and edges in the

social network (e.g., the content of messages exchanged between users). Nevertheless, semantic properties are not valid for any social network and could be hard to obtain, as selecting them depends on what the nodes and edges represent in the social network. Specifically, for academic social networks, the data available comes from collaboration between authors and/or publications [Cheng et al., 2014; Digiampietri and Maruyama, 2014]. Not having the over-used social interaction, data requires new and better topological features. Hence, Table 3.2 shows different topological properties that have been used to measure tie strength on such context. Note that neighborhood overlap is the most common here because the overlap captured by such metric increases when tie strength increases.

3.2.2 Tie Strength in Temporal Networks

The temporal variation in the social networks topology usually challenges traditional methods applied in static networks. For example, Jiang et al. [2016] propose a new technique to identify rumor sources in time-varying SNs. Also, Jin et al. [2012] develop a system to predict company performance based on how inter-company networks change over time. Further, Kang et al. [2014] present a framework that extends traditional SNs data management with spatial, temporal and uncertain aspects.

Although the many research efforts in investigating social networks, the combination of tie strength and temporal aspects has not been largely explored yet. For example, Dasgupta et al. [2008] use tie strength associated with time to demonstrate its influence in operators network. Likewise, Karsai et al. [2014] use tie strength to characterize the impact of time-varying and heterogeneous interactions on rumor spreading. Both aforementioned studies consider the temporal evolution of the strength of ties, but they do not propose a new way to measure such property by including time. On the other hand, Kostakos [2009] and Nicosia et al. [2013] propose a set of network properties that consider the temporal aspect in their computation. They showed many of such properties need to be calculated differently from the static networks.

A related problem is how to define what strong and weak ties are in temporal networks. For instance, Laurent et al. [2015] define strong ties as frequent interactions that connect nodes intra-communities and model the network structure locally, whereas weak ties are infrequent interactions situated inter-communities and maintain the network structure globally connected. Karsai et al. [2014] consider both the amount of interactions and the time of the interactions to define the strength of ties. Then, strong ties are time repeated and frequent interactions among pairs of individuals, whereas weak ties occur only occasionally. In a different manner, Nicosia et al. [2013] define

two nodes i and j as strongly connected if they are in a not symmetric relation (i is temporally connected to j but not vice-versa), whereas they are weakly connected if in a symmetric relation (both i is temporally connected to j, and j is temporally connected to i).

In this thesis, we consider the concept of strong and weak ties for temporal SNs based on Karsai et al. [2014]'s idea, i.e., a strong tie persists over time, and a weak tie occurs sporadically. However, Karsai et al. [2014] characterized the strength of ties based on a *single* time window of the network. Here we experimentally verify if the time window is a factor for characterizing the strength of tie by analyzing the persistence and transformation of ties over time. We show that, in fact, the strength of ties is very sensitive to the time window used to compute it.

3.3 Concluding Remarks

In this chapter, we overviewed concepts adopted to analyze and measure the strength of ties in co-authorship social networks. We also presented the datasets that we use to build academic social networks.

Moreover, we overviewed studies that cover the strength of ties in non-temporal and temporal networks. Many studies address tie strength in non-temporal social networks [Brandão and Moro, 2015; Bruggeman, 2016; Castilho et al., 2017; Goulas et al., 2015; Granovetter, 1973; Wiese et al., 2015]. Following Granovetter [1973]'s theory, ties are weak when they serve as bridges in the social network by connecting users from different groups, and strong when they link individuals in the same group. In this context, we propose a new topological feature that helps to measure tie strength in co-authorship social networks.

Regarding temporal networks, in this work, we define the concept of strong and weak ties for temporal social networks based on Karsai et al. [2014]'s idea (although they have not experimentally verify it, and we have). Specifically, we consider that a strong tie persists over time and a weak tie occurs sporadically. Also, we analyze the persistence and transformation of the ties over time. To do so, we propose a new algorithm and compare with an existing ones.

Chapter 4

A Preliminary Study on the Strength of Co-authorship Ties

Topological properties capture the characteristics of the graph that represents a social network [Zaki and Meira Jr, 2014]. In this chapter, we consider topological properties that have been applied for analyzing the importance of researchers [Barabasi et al., 2001; Gonçalves et al., 2014; Yan et al., 2012], the distance among researchers [Burt, 2004; Newman, 2001b] and the density of connections in the network [Barabasi et al., 2001]. One special property is the strength (or weakness) of the ties (edges, relationships or links) in the network. Now, we propose to analyze how some topological properties relate to the strength of ties in co-authorship social networks.

Specially, considering the aforementioned scenario of evaluating research through bibliometry and social networks (Section 1.1), we discuss how to improve such evaluation by analyzing how what interferes with tie strength. Therefore, we build (non-temporal) co-authorship social networks considering real datasets of publications from three different areas (computer science, medicine and sociology), which are also quantitatively compared (Section 4.1). We then characterize the strength of ties measured by neighborhood overlap and define a nominal scale to classify the ties as weak or strong. Also, we verify if the Granovetter's theory governs the three networks (Section 4.2). Afterwards, we analyze how nine topological properties impact on the strength of ties. Initially, we study the correlation between each property and neighborhood overlap. Then, our analysis takes one step forward and considers a regression model to quantify how the combination of each property to neighborhood overlap may improve even further the evaluation results (Section 4.3). Finally, we analyze the strength of ties intra and inter communities by using neighborhood overlap and weight in different clustering algorithms (Section 4.4), then showing a practical use for tie strength metrics.

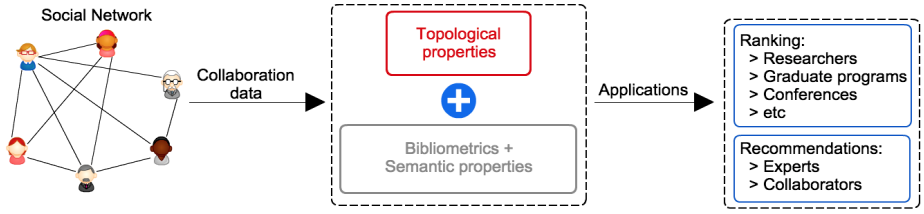


Figure 4.1: Architecture of a general research evaluation-oriented system.

4.1 Datasets Main Features

We consider systems and database applications that measure research productivity by evaluating the social aspects of researchers. Examples of such applications include (but are not limited to) ranking researchers, graduate programs and conferences, and recommending experts and collaborators, as presented in [Brandão et al., 2013; Chan et al., 2016; Ductor, 2015; Lima et al., 2013; Lopes et al., 2010; Lopes et al., 2011; Ribas et al., 2015; Silva et al., 2014; Yu et al., 2016a]. Figure 4.1 illustrates a general architecture of such systems: from a social network (e.g., co-authorship network), the collaboration data is extracted; then the system applies bibliometrics and analyzes semantic properties, whose results are sent to the applications. Here, we show the importance of analyzing the topological properties as well, as also done in [Barabasi et al., 2001].

For characterizing the importance of topological properties, we build three co-authorship social networks using the *CiênciaBrasil* datasets[1]. The publications available in *CiênciaBrasil* are from Brazilian researchers and have been collected from Lattes, an online platform for archiving researchers' curriculum vitae, in November 2013. Each network represents the co-authorships among researchers from three areas: computer science, medicine and sociology.

Figure 4.2 presents the distributions of the number of co-authors for the three areas. We consider these areas because there are clearly three different degrees of collaboration: low for sociology (up to three co-authors), medium for medicine (up to seven co-authors) and high for computer science (up to 15 co-authors). Note that for the authors in computer science and medicine, publishing together is a common practice, which is not for sociology (as presented by Simon [1974]). For instance, from 7,195 publications in sociology, 83.96% have only one author. Also, although the three networks are from Brazilian researchers, there are other studies that corroborate such behavior on different datasets [Acedo et al., 2006; Glänzel and Schubert, 2005; Huang and Huang, 2006]. Finally, Table 4.1 summarizes the datasets with the number

[1]Datasets available at `http://www.dcc.ufmg.br/~mirella/Tools/DEXA2015/`

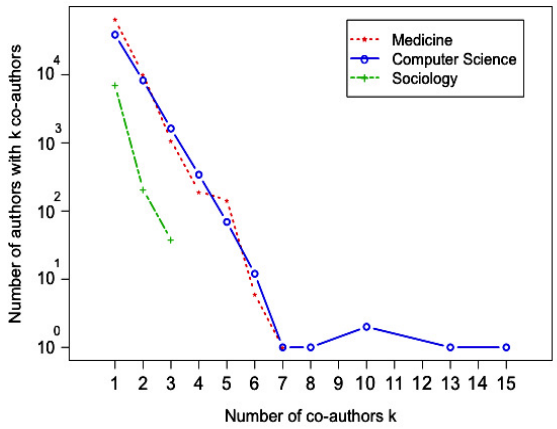


Figure 4.2: Distribution of numbers of co-authors for researchers in each area.

Table 4.1: Description of the datasets for building social networks.

Area	#Insts	#Res	#Publs	AvgPubA	#Pairs (#dist)	#SubA
CS	111	543	48,706	89.69	16,312 (1,563)	884
Med	114	368	75,553	205.30	16,089 (778)	664
Soc	43	96	7,195	74.95	322 (39)	68

Note: CS = Computer Science, Med = Medicine, Soc = Sociology

of institutions, number of researchers (authors of papers), number of publications, average number of publications per author, number of pairs of co-authors (and number of distinct pairs of co-authors) and number of subareas.

4.2 Characterizing the Strength of Ties

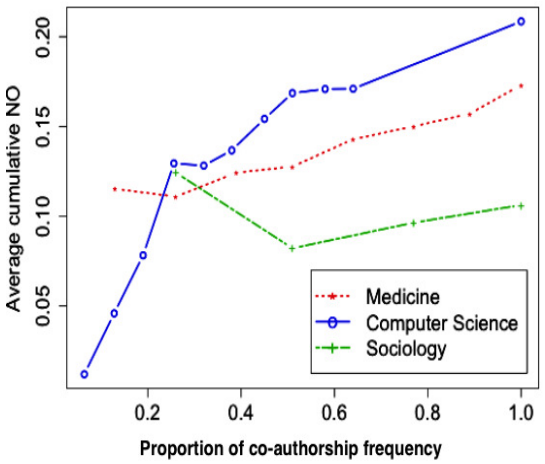
In this section, we measure the strength of co-authorship ties using the metric called *neighborhood overlap*. We now describe and characterize such metric in the three non-temporal co-authorship SNs (Section 4.2.1). Also, as we are studying the strength of ties, we have to verify if the ties follow Granovetter's theory (weak ties tend to connect nodes from different communities), i.e., if such theory governs the co-authorship social networks when the strength of ties is measured by neighborhood overlap. If the theory is not valid in the studied co-authorship social networks, we should consider other theories to characterize the ties as weak or strong. Hence, we also analyze the topological properties when weak and strong ties are removed (Section 4.2.2).

4.2.1 Neighborhood Overlap Characterization

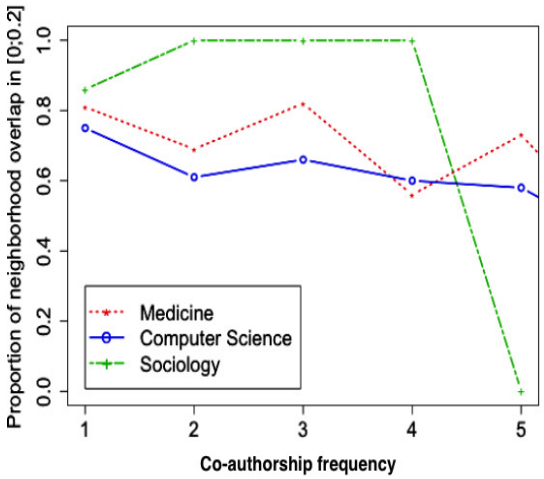
We consider the topological properties of three co-authorship social networks to investigate their impact to the ties strength. Here, the strength of a tie (relationship between a pair of researchers) is estimated by the neighborhood overlap metric of an edge connecting researchers v_i and v_j [Easley and Kleinberg, 2010]. The metric is given by the equation: $\frac{|\mathcal{N}(v_i) \cap \mathcal{N}(v_j)|}{|\mathcal{N}(v_i) \cup \mathcal{N}(v_j)| - \{v_i, v_j\}}$, where $\mathcal{N}(v_i)$ represents the co-authors of researcher v_i, and $\mathcal{N}(v_j)$ the co-authors of v_j.

Using neighborhood overlap to measure the strength of co-authorship ties is not well defined in the literature yet. Hence, such investigation is important to discover whether such metric captures the real importance of the tie to a researcher. We emphasize that neighborhood overlap does not consider the semantic of the relationship between pairs of researchers (for example, the period or the asymmetric importance of the co-authorship to a researcher) because it is not the focus of this work. However, neighborhood overlap captures the *density of co-authorship ties among researchers*. Such density is important because it measures the strength of the ties considering the *neighborhood effect*, which means the larger the number of common co-authors that a pair of researchers has, the more such pair tends to initiate a collaboration. This indicates that such tie is stronger if considering the neighborhood overlap. In the academic context, such effect is a variable that explains the tendency of a researcher to collaborate with others based on the relational effects of the researcher working in the neighborhood. For instance, given a pair of researchers v_i and v_j with a tie in the network where each researcher has six co-authors with only one in common. Such tie is weak, because there is only one co-author in common from 10 possibilities. In such case, the tie being weak means that the neighbors do not have much effect on the co-authorship, i.e. the intensity of co-authorship is small.

Conceptually, the tie strength grows as the neighborhood overlap increases. A tie is considered *weak* when the neighborhood overlap is very small [Easley and Kleinberg, 2010], and the problem becomes what "very small" means (0.01? 0.1? 0.2?). Note that ties are strong when their neighborhood overlap has opposite values to those defined for weak ties. According to Granovetter [1973], the tie strength is based on properties associated to the individuals' relationship (e.g. the intensity or the age). Hence, in order to define a nominal scale for tie strength, we compare two properties: neighborhood overlap and edge weight (i.e., an edge is the link between two researchers, and the weight is the number of their co-authored publications). Here, the edge weight represents the absolute frequency of interaction in co-authorship social networks. Thus, we call edge weight as *co-authorship frequency*.



(a) The average cumulative neighborhood overlap increases, when larger values for co-authorship frequency are included.

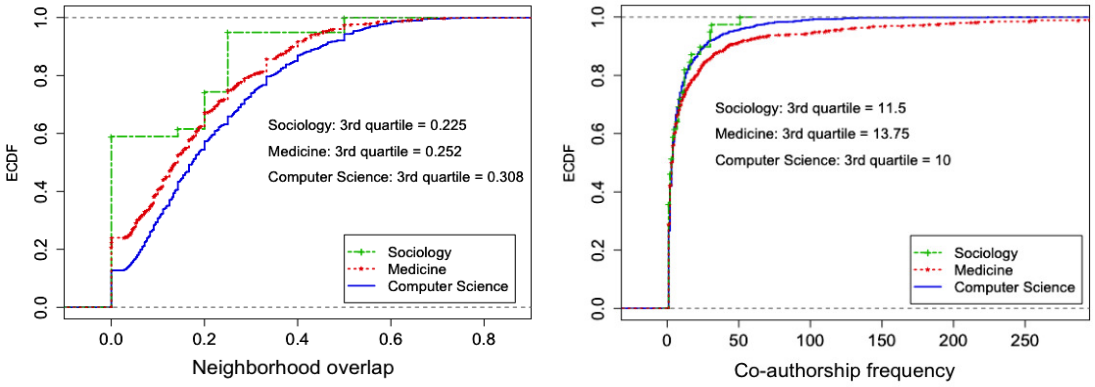


(b) Proportion of weak ties in co-authorships of same co-authorship frequency.

Figure 4.3: Analyzing the neighborhood overlap versus co-authorship frequency.

Considering the three areas, Figures 4.3a and 4.3b show the relationship between their neighborhood overlap and co-authorship frequency. Figure 4.3a shows the average cumulative neighborhood overlap for a fraction of co-authorship frequency. To compute such average, we sort all edges in increasing order by co-authorship frequency. Then, we take the top k ($0 \leq k \leq 1$) fraction of edges from the sorted list and calculate the average neighborhood overlap for those edges. Observe that when edges with larger co-authorship frequency are included, the average cumulative neighborhood overlap increases in the three areas on average. Also, including all co-authorship frequencies, the average cumulative neighborhood overlap does not reach much more than 0.2 (0.21 to computer science, 0.17 to medicine and 0.11 to sociology). Such average represents the typical value of neighborhood overlap in each network.

Then, Figure 4.3b presents the proportion of ties with co-authorship frequency varying from one to five when the neighborhood overlap ranges from zero to 0.2. In computer science and medicine, 55% of the ties with co-authorship frequency in the range [1;5] have neighborhood overlap in the range [0;0.2]. In sociology, most ties with small co-authorship frequency have neighborhood overlap in the range [0;0.2]. Such analysis suggests that ties with small co-authorship frequency also have high proportion of ties with neighborhood overlap between [0;0.2]. Therefore, we define that *a tie is weak* when the neighborhood overlap is within [0;0.2] as well. We also note that, in practice, the values of the nominal scale vary slightly depending on the research area, but to simplify the analysis, we have standardized the values in only one scale, which



(a) (b)

Figure 4.4: Empirical CDF of neighborhood overlap and co-authorship frequency computed by the co-authorship between pairs of researchers.

does not result in loss of information.

Now, the goal is to analyze the distribution of neighborhood overlap and co-authorship frequency in the three networks and compare them. Such analysis contributes to understanding the presence of weak and strong ties in the networks. Figure 4.4 presents the ECDF[2] (Empirical Cumulative Distribution Function) of neighborhood overlap and co-authorship frequency for the three co-authorship networks. The third quartile values in the graphic indicate that 75% of the data is less than that number in each area. It shows that only 25% of the co-authorships have neighborhood overlap equal or higher than 0.308 in computer science, 0.252 in medicine and 0.225 in sociology. The number of co-authorships among researchers (co-authorship frequency) is also small for the three networks; i.e., 25% of the pairs of researchers have co-authorship frequency equal or higher than 10 in computer science, 13.75 in medicine and 11.5 in sociology. Hence a direct conclusion is that weak ties are strongly present in co-authorship networks independently from the research area.

Also, the analysis of neighborhood overlap and co-authorship frequency distributions indicates that computer science has more ties with co-authors in common than medicine and sociology. Hence, the neighborhood overlap does indeed capture the real co-authorship among researchers in the three networks, because the results concur with the distribution of co-authors from Figure 4.2.

[2]ECDF assigns a probability of $1/n$ to each value of neighborhood overlap and co-authorship frequency, sorts the data in increasing order, and calculates the sum of the assigned probabilities up to and including each value.

4.2.2 Granovetter's Theory Analysis

Granovetter's theory [Granovetter, 1973] raises the hypothesis about the importance of weak ties in same situations. Following such theory in the academic context, the weak ties connect researchers from different communities, for instance, different research groups or teams. Considering the case where the ties are strong between two individuals, such theory suggests the existence of a triad claiming that if A and B are connected, and A and C are connected, then B and C will probably be connected. In other words, the strong ties link researchers *within* the same groups and teams. In order to better understand the strength of ties behavior when measured by neighborhood overlap and verify whether Granovetter's theory governs the networks, we present how removing the weak ties and the strong ties affects the topological properties.

Table 4.2 presents the properties of the social networks with all ties, when removing ties with neighborhood overlap equal to zero, and in the ranges [0;0.1] and [0;0.2]. As expected, all three networks are topologically affected by removing weak ties. In general, the average degree, diameter, density and total number of triangles decrease, whereas the total number of communities (the communities are detected by Louvain Method [Blondel et al., 2008]), total number of connected components and average clustering coefficient increase. Such increases indicate the weak ties really connect researchers from different communities and validate Granovetter' theory. Furthermore, the diameter decreases when ties are removed, because the nodes in the social network get disconnected and then the size of the main connected component also decreases. Note that changes in the topological properties when removing ties are common. However, if such ties are *not* important and/or influential in the network, such changes are *not* significant.

We now discuss the values of topological properties when strong ties are removed. Table 4.3 shows the social networks properties values with all ties, when removing ties with neighborhood overlap in the ranges [1;0.8], [1;0.5] and [1;0.2[. The ties are strong when their neighborhood overlap is higher than 0.2. We remove the strong ties starting from the range 1.0 to 0.8, then 1.0 to 0.5, and 1.0 to 0.2 (not including 0.2). Most properties change their value more (i.e. have greater impact) when weak ties are removed than strong ties. For example, when weak ties are removed, the average degree varies more than when strong ties are removed. In computer science, for instance, removing weak ties changes the average degree from 3.44 to 1.47, whereas removing strong ties changes from 3.44 to 1.97. Second, the properties for modularity, average clustering coefficient, average neighborhood overlap and average path length change differently when weak and strong ties are removed. For instance, the average clustering coefficient

Table 4.2: Co-authorship social networks properties when removing weak ties.

Topological Property	All Ties			Equal 0			[0;0.1]			[0;0.2]		
	CS	**Med**	**Soc**	**CS**	**Med**	**Soc**	**CS**	**Med**	**Soc**	**CS**	**Med**	**Soc**
# removed ties	0	0	0	199	168	23	481	302	24	896	516	29
Avg. degree	3.44	2.48	0.87	3.00	1.94	0.36	2.38	1.52	0.33	1.47	0.83	0.22
Diameter	14	11	6	12	9	2	14	20	2	13	6	2
Density	0.02	0.02	0.04	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.01	0.01
# communities	20	23	12	108	118	35	137	128	35	215	205	38
Modularity	0.85	0.69	0.85	0.75	0.72	0.73	0.83	0.8	0.72	0.89	0.87	0.62
C. component	8	11	12	98	110	35	127	119	35	212	201	38
Avg. clust. coef.	0.47	0.42	0.43	0.63	0.69	0.91	0.68	0.66	0.81	0.71	0.78	0.79
# triangles	2125	641	6	2128	641	6	1813	506	5	1190	268	3
Avg. path length	4.75	3.31	2.12	4.56	4.21	1.11	5.20	7.14	1.17	4.49	2.36	1.17
Avg. NO	0.21	0.17	0.11	0.26	0.25	0.31	0.33	0.31	0.27	0.43	0.39	0.25

Note: Avg = average, CS = Computer Science, Med = Medicine, Soc = Sociology, C. component = connected components, NO = neighborhood overlap

Table 4.3: Co-authorship social networks properties when removing strong ties.

Topological Property	[1;0.8]			[1;0.5]			[1;0.2[		
	CS	**Soc**	**Med**	**CS**	**Soc**	**Med**	**CS**	**Soc**	**Med**
# removed ties	1	0	0	124	2	31	667	8	262
Avg. degree	3.44	—	—	3.167	0.867	2.379	1.97	0.64	1.64
Diameter	14	—	—	14	6	11	18	5	11
Density	0.015	—	—	0.014	0.037	0.015	0.009	0.029	0.011
# communities	32	—	—	31	13	22	56	17	39
Modularity	0.85	—	—	0.836	0.85	0.675	0.804	0.84	0.645
C. component	8	—	—	8	12	11	39	17	28
Avg. clust. coef.	0.47	—	—	0.347	0.143	0.383	0.089	0.103	0.11
# triangles	2117	—	—	1233	2	487	130	1	73
Avg. path length	4.75	—	—	4.77	2.14	4.4	5.07	2.08	4.65
Avg. NO	0.207	—	—	0.133	0.034	0.133	0.028	0.019	0.031

increases 50% for computer science when removing all weak ties and decreases 81.22% when removing strong ties. These results agree with Granovetter's theory, because the clustering coefficient measures the trend of nodes in a network to form clusters. By definition, the greater the clustering coefficient, the greater the number of closed triads. Additionally, the giant component breaks more rapidly when a critical number of weak ties is removed, since the number of connected components varies from 8 to 212 in computer science, 11 to 201 in medicine, and 12 to 38 in sociology. Furthermore, when strong ties are removed, such value varies from 8 to 39 in computer science, 11 to 17 in medicine, and 12 to 28 in sociology. All these results suggest Granovetter's theory is *valid for the three networks*: weak ties link researchers from different groups or teams, whereas strong ties connect researchers from the same ones.

Table 4.4: Social network topological properties (see [Easley and Kleinberg, 2010] for formal definitions).

Notation	Description
eBetweenness	edge betweenness of each edge in the network
Co-authorship frequency	frequency of researchers published a work together
ClosenessA1/ClosenessA2	closeness of each researcher in a pair of researchers
AvgCloseness	average closeness of each pair of researchers
EccentricityA1/EccentricityA2	eccentricity of each researcher in a pair of researchers
AvgEccentricity	average eccentricity of each pair of researchers
ClusterCoefA1/ClusterCoefA2	clustering coefficient of each researcher in a pair of researchers
AvgClusterCoef	average clustering coefficient of each pair of researchers
nTrianglesA1/nTrianglesA2	number of triangles of each researcher in a pair of researchers
AvgNTriangles	average number of triangles of each pair of researchers
wDegreeA1/wDegreeA2	weight degree of each researcher in a pair of researchers
AvgWDegree	average weight degree of each pair of researchers
EigenvecA1/EigenvecA2	eigenvector value of each researcher in a pair of researchers
AvgEigenvec	average eigenvector of each pair of researchers
PageRankA1/PageRankA2	page rank of each researcher in a pair of researchers
AvgPageRank	average page rank of each pair of researchers

4.3 The Impact of the Properties on Tie Strength

So far, we have measured tie strength by the neighborhood overlap of two researchers. Indeed, considering co-authorship social networks, the actual strength of a tie may depend on characteristics of the network graph per se. Therefore, we relate the strength of ties to topological social network properties that capture: *(i)* the importance of researchers (weight degree, eigenvector and pageRank) and pairs of researchers (edge betweenness and co-authorship frequency) within a network; *(ii)* the distance of a researcher from the furthest other (closeness and eccentricity); and *(iii)* the degree to which researchers tend to cluster together (clustering coefficient and number of triangles), as defined in Table 4.4.

Analyzing the importance of each property to neighborhood overlap and quantifying their strength may reveal knowledge to improve systems that combine bibliometry and social network analysis (e.g. algorithms for ranking graduate programs [Lopes et al., 2011]). Also, relative measures (consider and compare more than one aspect) can better represent the reality than absolute counts (for only one aspect) [Pendlebury, 2009]. Hence, the combination of neighborhood overlap with other metrics can potentially generate better results. Here, we analyze the correlations between each property (individually) and neighborhood overlap (Section 4.3.1), and use a regression model to quantify the importance of properties to neighborhood overlap as well (Section 4.3.2).

Table 4.5: Pearson correlation coefficients between topological properties and neighborhood overlap. Values lower than 0.1 are insubstantial.

Properties	Computer Science	Medicine	Sociology
AvgClusterCoef	0.61	0.73	0.88
ClusterCoefA1/	0.47	0.62	0.81
ClusterCoefA2	0.53	0.56	0.79
AvgNTriangles	0.38	0.26	0.93
nTrianglesA1/	0.37	0.32	0.86
nTrianglesA2	0.28	0.15	0.82
Co-authorship frequency	0.31	0.3	0.34
EigenvecA1/	0.3	0.2	0.65
EigenvecA2	0.23	0.11	0.62
AvgEigenvec	0.3	0.17	0.68
AvgWDegree	0.27	0.11	0.6
wDegreeA1/	0.24	0.17	0.54
wDegreeA2	0.17	0.03	0.34
PageRankA1/	0.2	-0.04	-0.17
PageRankA2	0.09	-0.02	-0.1
AvgPageRank	0.19	-0.03	-0.26
AvgEccentricity	0.13	0.21	-0.034
EccentricityA1/	0.11	0.2	0.086
EccentricityA2	0.13	0.2	0.23
ClosenessA1/	-0.06	-0.14	0.013
ClosenessA2	-0.1	-0.15	-0.2
AvgCloseness	-0.09	-0.15	-0.067
eBetweenness	-0.4	-0.5	-0.26

4.3.1 Correlation Analyses

We quantify the correlations between each property and the strength of ties by using the Pearson linear correlation coefficient and the Spearman's rank correlation coefficient [Jain, 1991]. The Pearson coefficient measures the linear relationship between two variables. When such relationship is not linear, the Spearman's rank correlation coefficient is more appropriate. The Pearson coefficient is presented in Table 4.5, whereas Spearman results are very similar and thus omitted.

We follow the conventions to interpret correlation coefficient from [Cohen, 1988]: greater than 0.7 is *very large*, within [0.5;0.7) is *large*, within [0.3;0.5) is *moderate*; within [0.1;0.3) is *small*, and anything smaller than 0.1 is *insubstantial* (note that the same ranges are valid for negative correlation). Table 4.5 shows that, for most properties, the correlations with neighborhood overlap tend to be *small* or *moderate*. In computer science and medicine, the neighborhood overlap varies slightly when the other topological properties change. However, in the sociology network, these correlations are balanced (there are 11 large correlations against 12 *small* and *moderate* ones). Such

situation may be explained by the smaller size of the sociology co-authorship network and/or limited number of co-authorships (the maximum number of co-authors in a paper is three for sociology network).

As for specific differences among the networks, number of triangles (nTrianglesA1, nTrianglesA2 and AvgNTriangles), weighted degree (wDegreeA1, wDegreeA2 and AvgWDegree) and eigenvector (EigenvecA1,EigenvecA2 and AvgEigenvec) have small or moderate correlations in computer science and medicine, but large in sociology. Such large linear correlation in sociology is expected, because, conceptually, the existence of more triangles indicates more neighbors in common. A direct question is: why the number of triangles is not linearly correlated with neighborhood overlap in computer science and medicine? We may only speculate that such correlation is non linear or there is no correlation due to factors not considered in this work. Furthermore, the large linear correlation between neighborhood overlap and metrics that capture the importance of a researcher in a network (weighted degree and eigenvector) indicates that in sociology, important researchers in the network have co-authorship ties stronger than others. Up to now, we cannot claim such behavior for computer science and medicine.

Moreover, for computer science and medicine, clustering coefficient is the property most linearly correlated with neighborhood overlap: the correlation between AvgClusterCoef and NO reaches 0.61 in computer science and 0.73 in medicine. For sociology, the most correlated property is number of triangles (0.93), although the correlation between clustering coefficient and neighborhood overlap is also large (0.88).

These correlations provide evidence of properties that are strongly related to the strength of ties in co-authorship networks, and thus can help to explain such strength. They may also provide insights to improve methods that consider the strength of ties concept. For instance, based on the observed patterns and the research area, the design of an assessment system for conferences or teams evaluation that considers the strength of ties may also consider clustering coefficient, number of triangles, edge betweenness, weighted degree and/or eigenvector to improve its accuracy and overall quality. In addition, the correlated topological properties can be used to answer the question: why a certain tie has a particular strength? Finally, topological properties may also be combined to better explain the strength of ties or have a non linear correlation (greater than linear correlation) with neighborhood overlap. We investigate these issues next.

4.3.2 Regression Analyses

Another way to further assess the relative importance of each topological property to the neighborhood overlap is to use a *regression model*. Such model is obtained

from a statistical process called regression analysis that estimates the relationships among variables [Jain, 1991]. The quality of the regression model is estimated by the coefficient of determination R^2, which represents the fraction of the variation in the response variable y that is explained by other variables. Here, the response variable y is neighborhood overlap, and the variables are the nine topological properties described in Table 4.4. Overall, the goals in this section are: to *identify* which of the topological properties are necessary to build a model that efficiently characterizes the neighborhood overlap in each research area and to *quantify* the relative importance of each property to neighborhood overlap.

To define an appropriate regression model, we apply *linear regression models* without and with logarithm and exponential transformations. First, we use simple linear regression model (without and with transformations) considering each topological property as the variable (factor) and the neighborhood overlap as the response variable (or estimated variable). However, for most topological properties, the quality of regression is poor. For instance, the R^2 value for the linear regression between edge betweenness and neighborhood overlap is 0.161 in computer science (0.068 in sociology and 0.25 in medicine) and with logarithm transformation is 0.329 (-0.22 in sociology and -0.96 in medicine). Nonetheless, using *simple exponential regression model* has improved the quality of regression. For example, the R^2 value for the exponential regression between edge betweennness and neighborhood overlap is 0.966 in computer science (0.746 in sociology and 0.971 in medicine). Also, the results show that most properties are statistically significant (non-zero), with 95% confidence level, for all three areas.

Therefore, we apply a *multiple exponential regression model* to estimate a response variable y as exponential function of k variables (i.e., topological properties) $x_1, x_2, ..., x_n$ using the following equation: $y = \beta_0 * \beta_1^{x_1} * \beta_2^{x_2} * ... * \beta_k^{x_k}$, or $ln(y) = ln(\beta_0) + ln(\beta_1)x_1 + ln(\beta_2)x_2 + ... + ln(\beta_k)x_k$. We build one model for each social network by determining parameters $\beta_0, \beta_1, ..., \beta_k$ in order to minimize the least squared error for all researchers in the co-authorship network.

The first line of Table 4.6 shows the R^2 values for the models built considering *all* topological properties ($k = 9$). We have only considered the average values of node properties in the multiple regression models (e.g., for ClosenessA1, ClosenessA2 and AvgCloseness, we use AvgCloseness, where average represents the property value of two researchers in the co-authorship). Then, each subsequent line presents the cumulative results after removing the properties – i.e., the second line is after removing Closeness, the third is after removing Eccentricity from the previous model without Closeness, and so on. The properties are removed from the least to the most important. Specifically,

Table 4.6: Results with all properties and removing one property at a time.

Regression Model	**Model Quality(R^2)**		
	CS	**Med**	**Soc**
All properties	0.953	0.917	-0.103
AvgCloseness (-)	0.954	0.823	-0.105
AvgEccentricity (-)	0.954	0.884	-0.09
AvgClusterCoef (-)	0.969	0.935	-0.04
AvgNTriangles (-)	0.964	0.942	-0.01
AvgWDegree (-)	0.964	0.959	0.009
AvgEigenvec (-)	0.963	0.967	0.453
AvgPageRank (-)	0.967	0.967	0.546
Co-authorship frequency (-)	0.965	0.971	0.747

Note: The negative R^2 indicates that the model does not follow the trend of the data, i.e., the regression model is a worse fit [Cameron and Windmeijer, 1997].

the first two to be removed are closeness and eccentricity because their definitions consider the smallest path in the graph, which in theory has no relation to neighborhood overlap. Then, clustering coefficient and number of triangles come next because they are metrics that consider the neighbors information. The last ones to be removed are the most important because they consider the number of relationships of a node, which is directly associated to weak ties. After removing the co-authorship frequency, the only variable in the regression model is edge betweenness.

Considering the results, the models can well explain the strength of ties between researchers in computer science and medicine, with R^2 reaching 0.96 for computer science and 0.97 for medicine. For sociology, the R^2 is smaller: -0.105, which indicates a completely inappropriate model. However, after removing some topological properties from the model, the R^2 values start to increase and reach 0.747, which indicates a reasonably good model. The worst R^2 value for each co-authorship network can be explained by applying an exponential regression model. As shown in Table 4.5, there is a large linear correlation between some topological properties and the neighborhood overlap. Also, there is a noticeable increase of the models accuracy when the clustering coefficient (large correlation with NO in the three networks) is removed.

Analyzing the removal of each property enables to identify which ones are important to the quality of the regression model. Specifically, for computer science, removing clustering coefficient and pageRank increases the R^2 value, which indicates that such properties can *not* be used to explain variations of neighborhood overlap values. Likewise, removing other properties (such as number of triangles, eigenvector and co-authorship frequency) decreases the quality of the model, but not significantly (around -0.5%, -0.1% and -0.2%, respectively). For medicine, removing closeness and eccentricity reduces the regression model's quality, which reinforces the importance of

such properties to the model. Thus, the two metrics that measure the distance of a researcher from the furthest other (closeness and eccentricity) non linearly explain variations in the strength of the ties. Likewise, removing other properties increases the R^2 value, showing that they are not very important to the model. Then for sociology, keeping only eBetweenness is enough to have a good model. Also, for the three networks, the quality of the model is good when keeping only edge betweenness.

Such results show the relevance of a metric that represents the importance of a tie between researchers to explain the non linear variation in neighborhood overlap. Note that the computation of neighborhood overlap uses the betweenness concept. Thus, the correlation between the two metrics is expected. The novelty here is that such correlation is non linear and is a pattern for the three areas. At the end, such results indicate that each area network has the strength of ties related to other topological properties in *different* ways.

Finally, regarding the statistical significance of each model parameter, we set up a series of hypothesis tests, one for each parameter β_i, specified by a null hypothesis $H_0 : \beta_i = 0$. We found that most parameters are statistically significant (i.e., non-zero), with 99% confidence level for the models with the highest R^2 value. The only exceptions (p-value > 0.05) are the coefficients associated with the average eigenvector and the average page rank for computer science.

4.4 A Comparative Analysis of the Strength of Co-authorship Ties in Clusters

Clustering algorithms represent a classical problem of data mining and has many applications over a plethora of domains. Then, identifying which algorithm is proper to one such domain is a challenge per se. Likewise, evaluating the quality of the created clusters is hard due to its problem-driven nature, as a good clustering algorithm for a problem may not be as good for another [Almeida et al., 2011].

In the context of social networks (SN), clustering algorithms are useful for detecting (finding) communities. Examples of studies include to explore regional innovation systems, clustering effect in scientific communities and concentration of developers in a country [Brandão and Moro, 2017a]. Specially in academic SN, detecting clusters helps to discovery patterns that may increase the researchers' productivity, reveal the impact in research policy and understand group formation [Kshitij et al., 2015]. However, once again, the problem is how to verify the quality of the created clusters.

Here, we focus on applying clustering techniques in co-authorship SN, an aca-

demic network in which the nodes are researchers and there are edges between them if they have published together. Specifically, we focus on co-authorship SN from three different research areas: computer science, medicine and sociology. Characterizing these different areas by applying clustering algorithms helps to understand their profiles. For example, for ranking purposes, research areas with few collaboration patterns need evaluation criterion different from the ones with predominance of collaboration.

Clustering techniques have been applied to different types of networks, for example, directed networks [Malliaros and Vazirgiannis, 2013], social professional networks [Brandão and Moro, 2017a] and mobile SN [Kim and Kim, 2014]. From these techniques, we have chosen three that are commonly applied to undirected graphs. The first is the Louvain method that is one of the most used clustering algorithm based on modularity maximization (it measures how nodes in a cluster are better connected as opposity of a random connection) [Blondel et al., 2008]. The second one is the Clique Percolation method that is able to detect overlap communities, i.e., nodes can belong to more than one community [Palla et al., 2005]. The third is Markov Cluster algorithm, which forms clusters by alternating two Markov processes: expansion and inflation [Van Dongen, 2000]. According to Mishra et al. [2007], modularity, overlap and Markov chain are examples of strategies commonly used to detect communities in social networks.

There are many clustering techniques and identifying which one is the best for co-authorship social network is not an easy task due to the many aspects that can be evaluated. Also, the quality of a cluster is problem-driven as a "good" clustering algorithm for a problem is not necessarily good for another [Almeida et al., 2011]. In this sense, we evaluate clustering algorithms by using the strength of co-authorship ties (a type of social ties) metrics. The study of social ties has been used to build rigorous models that reveal the evolution of SN and the dynamics of information exchange [Aiello et al., 2014].

A social network cluster is a collection of individuals with dense interactions patterns internally and sparse interactions externally [Mishra et al., 2007]. In other words, when the strength of ties is defined by metrics that consider the neighborhood of nodes, the strength of ties intra cluster should be higher than inter clusters. Thus, we measure such strength by using neighborhood overlap and co_authorship frequency.

Also, there are different ways to measure clustering quality as described in Chapter 2, such as BetaCV, C-index and modularity [Brandão and Moro, 2017a]. However, identifying whether such metrics give the expected answer for a graph is very difficult [Almeida et al., 2011]. Moreover, most of these metrics are biased and unreliable in larger real graphs. Indeed, in this work, we investigate whether tie strength metrics

can be used to evaluate clustering quality. This study represents a new direction in the evaluation of clustering algorithms and may help to fill this gap in the state-of-the-art.

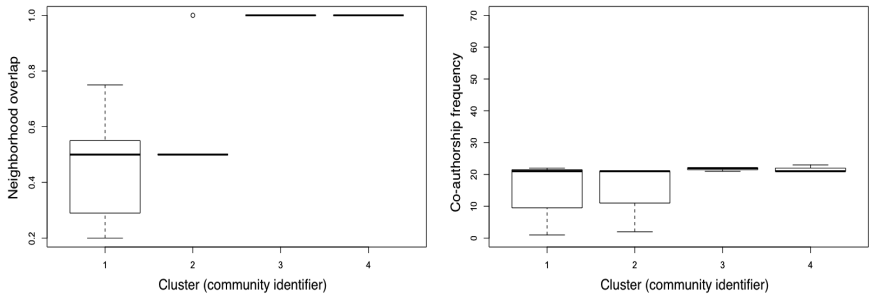
Overall, the contributions here are the analysis of the distribution for strong and weak ties intra and inter clusters, and the dynamism of the strength of ties through clustering algorithms. The results reveal whether tie strength metrics do indeed evaluate clusters quality based on the clustering techniques definition that the ties intra a community should be strong and inter clusters should be weak. Also, this study is important to get insights on the strength of the co-authorships among researchers intra and inter clusters. Such insights can help the design of methods for assessing research quality and productivity. For example, methods that consider the weak ties and communities concepts as Burt [2004] and Silva et al. [2014] may attribute different weights to the importance of weak ties depending on their community size.

Specifically, we analyze the results of three clustering methods: Louvain method (Section 4.4.2.1), clique percolation method (Section 4.4.2.2) and Markov cluster algorithm (Section 4.4.2.3). Then, we compare the results of these methods (Section 4.4.3). We have chosen these three algorithms because they are important to detect core groups on SN [Kim and Kim, 2014]. Note that in the social network context, clusters are also called as communities [Girvan and Newman, 2002]. Thus, we use both terms interchangeably and maintain the nomenclature of the clustering algorithms' authors (Louvain method called as LM, clique percolation method as CPM and Markov cluster algorithm as MCL).

4.4.1 Analyses Setup

Considering the datasets presented in Table 4.1, we apply three clustering algorithms: Louvain method (LM), clique percolation method (CPM) and Markov cluster algorithm (MCL). Hence, we measure the strength of ties for each pair of researchers in each cluster detected by the algorithms. Such strength is measured by using neighborhood overlap and co-authorship frequency. As presented in Section 4.2, we consider that a tie is weak when the neighborhood overlap is in the range $[0; 0.2]$ and strong otherwise. Likewise, a tie is weak when the co-authorship frequency is in the range $[1; 5]$ and strong otherwise [Brandão and Moro, 2015].

Indeed, we analyze the tie strength dynamism through different clusters formed by each algorithm. Such analyses provide insights whether the strength of ties metrics can be used to evaluate clustering quality. By clusters definition [Blondel et al., 2008; Palla et al., 2005; Van Dongen, 2000], ties intra-clusters should be strong and ties inter-clusters should be weak. Therefore, a cluster should have most pairs of researchers (ties)



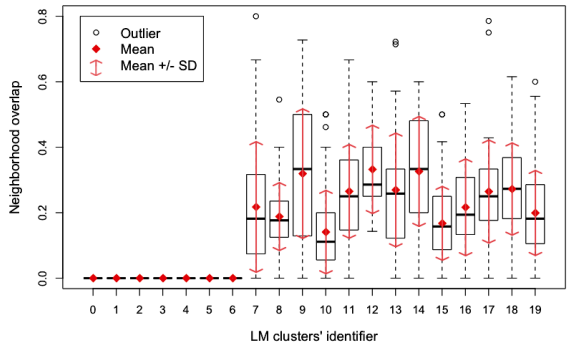
(a) Neighborhood overlap (b) Co-authorship frequency

Figure 4.5: The strength of ties intra-communities in a perfect clustering. In each box plot, the central rectangle spans the first to the third quartiles, the segment inside is the median, traits above and below the rectangle represent the minimum and maximum values. The clusters' identifiers order the box plots.

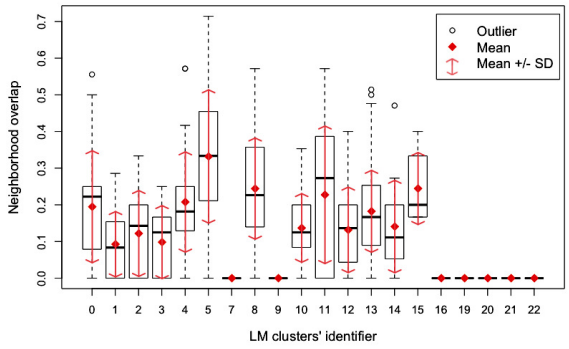
classified as strong and most ties that connect different clusters as weak.

One of the problems in evaluating clustering quality is the absence of a ground truth for comparison [Almeida et al., 2011]. Thus, we verify the strength of ties in a synthetic data that represents a situation with perfect clustering. According to Harman et al. [2005], a perfect clustering has a perfect modularization, i.e., all modules in a cluster are connected to all other modules and there are no inter-cluster connections. Thus, we build a graph with 17 (a random choose number) nodes and 23 edges. We link the nodes in a way to form four clusters and there are no connections among nodes from different clusters. Cluster #1 is the largest one (7 nodes and 12 edges), cluster #2 is the second largest (4 nodes and 5 edges), clusters #3 and #4 have the same size (3 nodes and 3 edges). Figures 4.5a and 4.5b present the neighborhood overlap and co-authorship frequency of a perfect clustering, respectively.

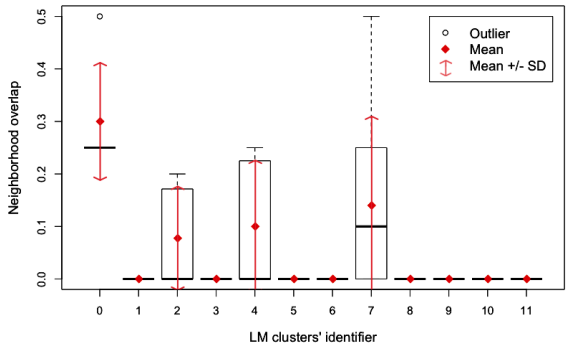
Note that the minimum value of neighborhood overlap is 0.2, i.e., most communities in these networks are composed by strong ties. The smallest clusters have neighborhood overlap equal to 1 (i.e., all ties are strongly connected), because all nodes are connected to each other, but in a real social network this hardly happens. We emphasize that a high neighborhood overlap indicates that pairs of researchers are more connected to each other intra a cluster. Also, the co-authorship frequency of all clusters has the median high than 20. This is a property strictly related to the frequency of nodes interactions. Thus, this may not be find in real networks. However, co-authorship SN with a high degree of collaboration tend to have a high co-authorship frequency [Brandão and Moro, 2015]. Therefore, most detected clusters should have more strong ties than weak ones.



(a) Computer science



(b) Medicine



(c) Sociology

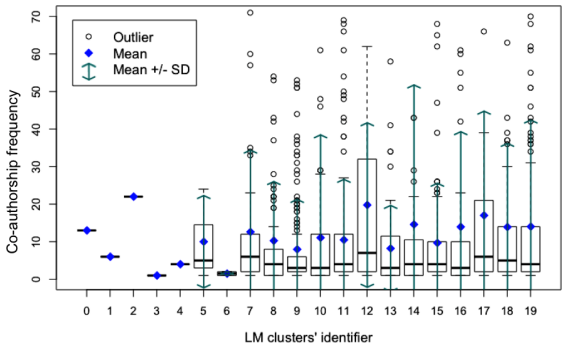
Figure 4.6: LM – The strength of ties intra-communities measured by neighborhood overlap.

4.4.2 Evaluated Clustering Techniques

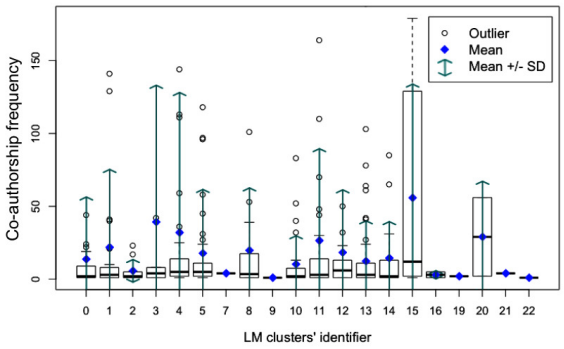
In this section, we present the analyzes of the three clustering techniques: Louvain method (Section 4.4.2.1), Clique Percolation method (Section 4.4.2.2) and Markov Cluster algorithm (Section 4.4.2.3).

4.4.2.1 Community Detection Using Louvain Method

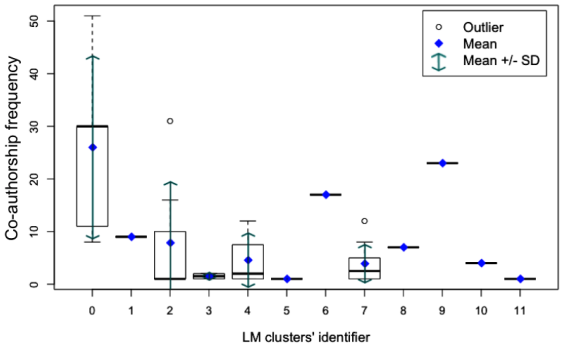
The Louvain method (LM) [Blondel et al., 2008] is a simple, efficient and one of the most common methods for detecting communities in large networks. In summary, this method makes greedy seeks to optimize the modularity of a partition of the network, where modularity is a topological property of a network and designed to measure the density of links intra communities [Blondel et al., 2008]. It is important to emphasize that the LM considers the SN as unweighted. Hence, it allows to study the dynamism of neighborhood overlap and co-authorship frequency of links between researchers in



(a) Computer science



(b) Medicine

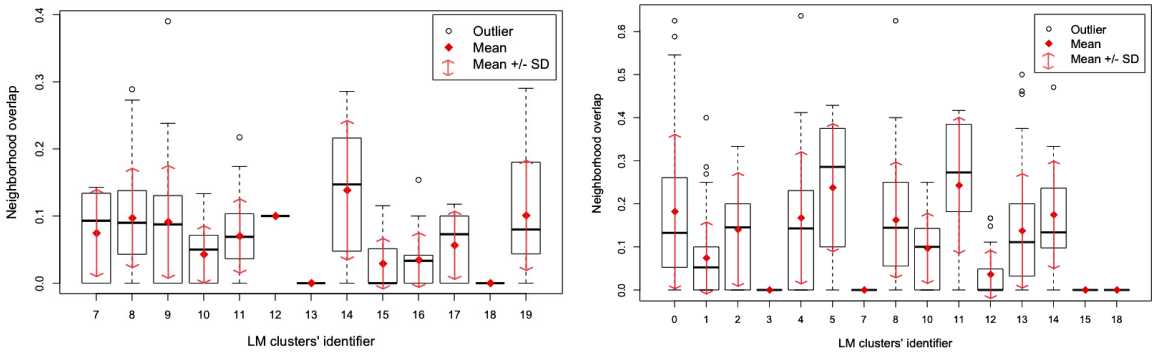


(c) Sociology

Figure 4.7: LM – The co-authorship frequency as a measure of the strength of ties intra-communities (computer science and medicine box plots present many outliers and some are omitted make the graph clearer).

clusters that are formed by the modularity and the network topology. The study of the influence of neighborhood overlap and co-authorship frequency in the formation of clusters is in Section 4.4.2.3 (because the cluster technique explored in such section depends on the weights attributed to the edges, i.e., neighborhood overlap and co-authorship frequency are used as weight of the social network, not only as a metric to measure tie strength).

Figures 4.6 to 4.9 summarize (in box plots) the neighborhood overlap and co-authorship frequency of ties intra and inter different communities detected by the Louvain method. Considering intra communities perspective, Figure 4.6 shows that there are communities with neighborhood overlap equal to zero, few edges compose all such communities in the three SN (the number varies from one edge to four), and all edges have one node in common (for example, all edges connect to a researcher A). These smaller communities are detected by LM because the researchers are not



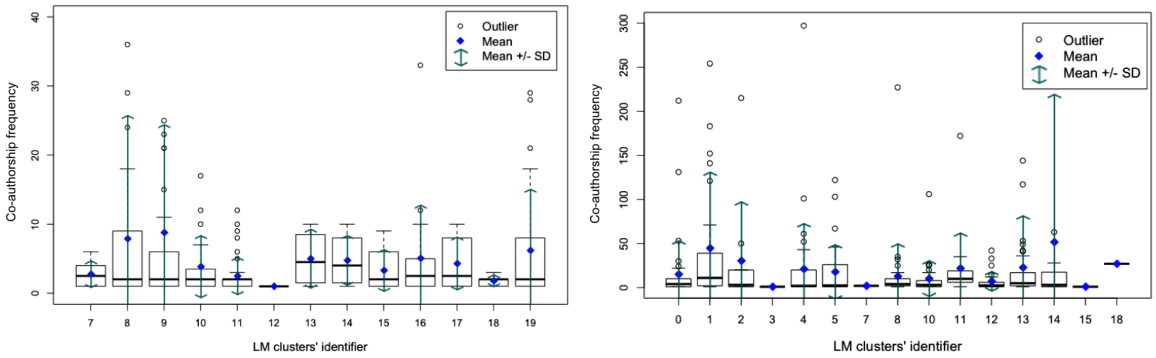
(a) Computer Science (b) Medicine

Figure 4.8: LM – The strength of ties inter-communities measured by neighborhood overlap.

densely connected to other communities. Also, most communities present weak ties (neighborhood overlap less or equal than 0.2). There are 12 communities with the mean and median values smaller or equal to 0.2 in computer science, 11 in sociology and 14 in medicine. This shows that despite the communities being densely connected, the presence of weak ties is strong. Furthermore, the analysis of the arrows in the box plot shows that in computer science, community #8 has the least disperse data, i.e., community #8 has interactions between researchers with uniformity in the value of neighborhood overlap. This uniformity may reveal a pattern in researchers interaction (we leave for future work the investigation of such a pattern). In sociology and medicine networks, all communities have a high dispersion data, which indicates the lack of uniformity or homogeneity in the co-authorships.

Figure 4.7 shows the same communities from Figure 4.6, but now, focusing on the co-authorship frequency of interactions between researchers. In computer science, most communities have the median value of co-authorship frequency less than 5, which indicates absence of very strong ties between researchers. Moreover, the mean and median among communities are too close, which indicates a pattern of high interaction between researchers. Regarding dispersion, it is high in communities with more than four edges (non-zero neighborhood overlap). In sociology, most communities have the means and median of co-authorship frequency smaller than 10. Thus, most ties are classified as weak. In medicine, the co-authorship frequency in communities #3, #4 and #15 reach a value greater than 100. Also, the mean and median among the communities are close.

Figures 4.8 and 4.9 compare the neighborhood overlap and co-authorship fre-

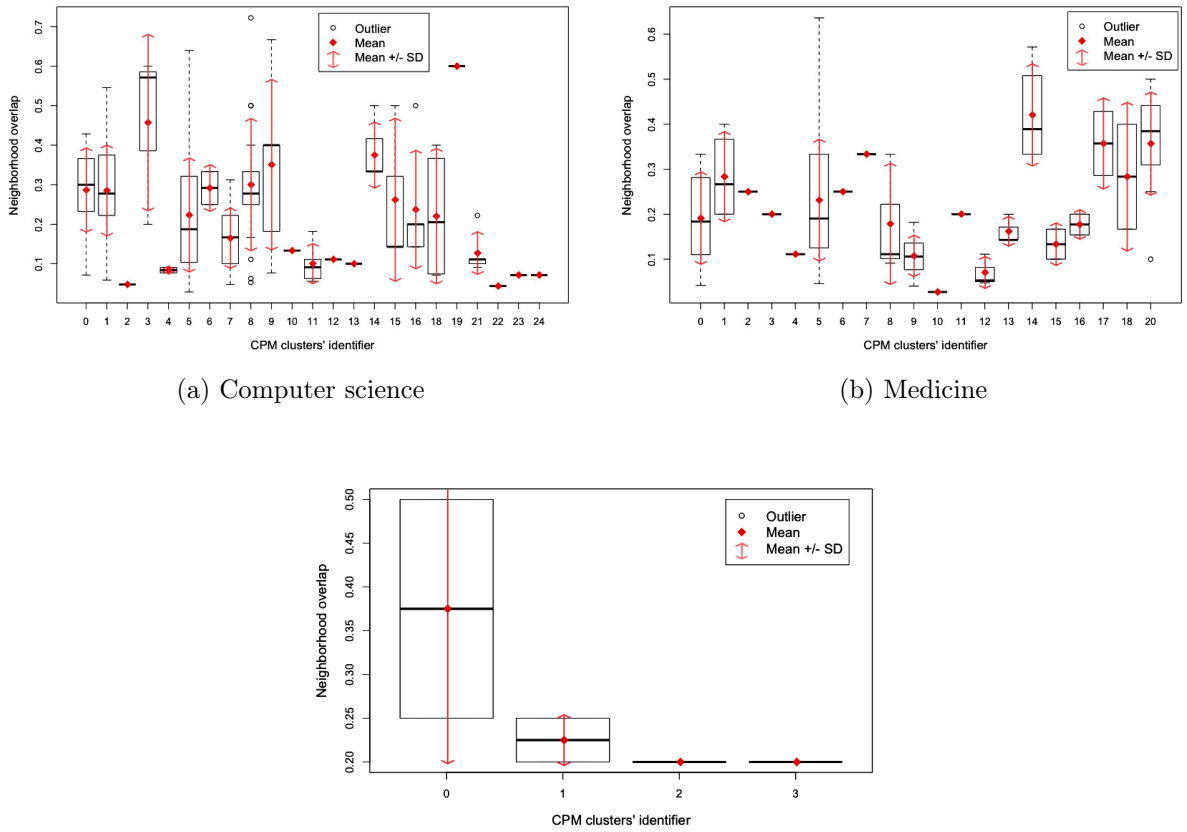


(a) Computer Science (b) Medicine

Figure 4.9: LM – The strength of ties inter-communities measured by co-authorship frequency.

quency of ties inter communities. For example, considering community #1 of the boxplots, the neighborhood overlap/co-authorship frequency values are from interactions of its researchers with those from other communities. This study is important to show how strong/weak the co-authorship between researchers from different communities is. Note that in sociology, there are no interactions between researchers from different communities, i.e. researchers from sociology publish within their communities only. Figure 4.8 shows that the ties inter communities are weak in most cases. All communities in computer science and 13 in medicine have the mean and median less than 0.2. However, the high dispersion also indicates the presence of strong ties in most of inter communities interaction. Regarding the co-authorship frequency, Figure 4.9 shows that the mean and median of the communities are close in computer science and medicine, suggesting a pattern in the co-authorship frequency of ties inter communities. In other words, the average co-authorship frequency of co-authorships inter communities tend to be similar independent of the community.

The results in this section show the difference between neighborhood overlap and co-authorship frequency as measures of the strength of ties. We note that medicine has more communities with smaller mean and median neighborhood overlap values than computer science, but the co-authorship frequency of such communities are higher than computer science. Also, in both areas, the communities with highest neighborhood overlap do not indicate communities with highest co-authorship frequency. In sociology, the neighborhood overlap and co-authorship frequency of researchers in each community is small, but communities with the highest neighborhood overlap do not have the highest co-authorship frequency. Such aspects suggest that the strength of the



(a) Computer science

(b) Medicine

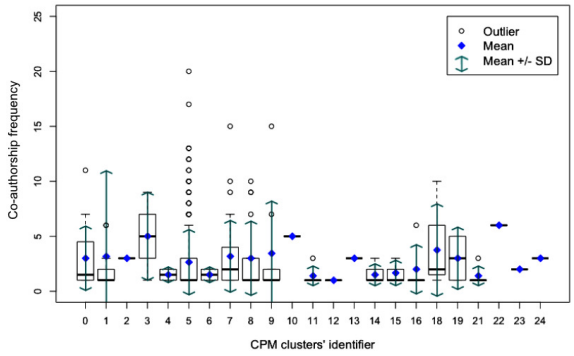
(c) Sociology

Figure 4.10: CPM – Neighborhood overlap as a measure of the strength of tie intra-communities (note the small number of outliers).

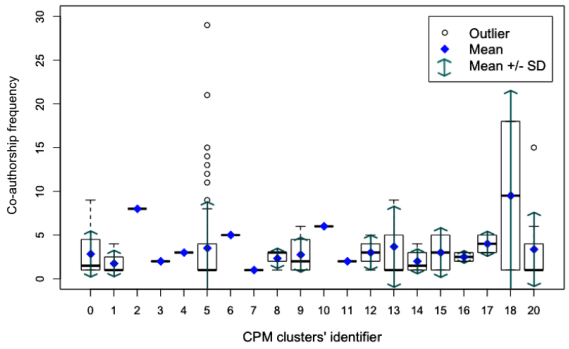
intensity of co-authorships among researchers measured by the co-authorship frequency does not always correspond to the strength of the interactions among researchers' neighbors measured by neighborhood overlap. Also, considering the outliers, the communities have less outliers to neighborhood overlap than to the co-authorship frequency. For future work, the study of such outliers might reveal interesting properties in the co-authorships among researchers. Lastly, considering that the ties intra-communities should be strong, whereas the ties inter-communities should be weak, Louvain method is not an appropriate method to detect communities in co-authorship SNs.

4.4.2.2 Uncovering Communities with Clique Percolation Method

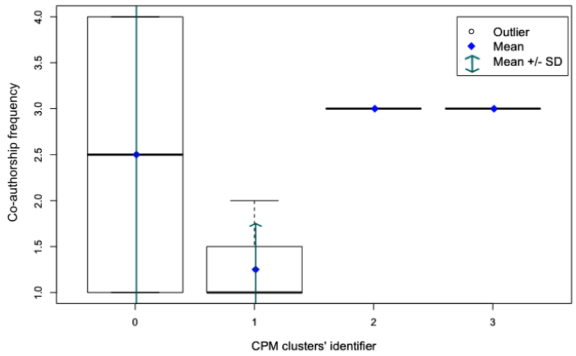
The clique percolation method (CPM) locates the k-clique communities of networks and considers that a typical member in a community is linked to many other members, but not necessarily to all other nodes in the community [Palla et al., 2005]. Overall,



(a) Computer science



(b) Medicine



(c) Sociology

Figure 4.11: CPM – The co-authorship frequency as a measure of the strength of ties intra-communities (we omit some outliers for computer science and medicine for better presentation; the number of outliers is smaller than in Figure 4.7).

a community is a union of smaller fully connected subgraphs that share nodes. Such complete subgraphs are called *k-cliques*, where k refers to the number of nodes in the subgraph. Then, *k-clique-community* is defined as the union of all k-cliques that can be reached from each other through adjacent k-cliques [Palla et al., 2005].

We apply CPM by using the algorithm implemented in CFinder[3] and k=3. By definition, a community is actually a connected graph when k=2 and a set of disconnected nodes without any edge when k=1. The parameter k is an important factor that determines the nature of the communities. Using different values for k reveals the nature of the communities [Deb et al., 2009]. We have chosen k=3 in order to discover triangles and because such a value is also used in most general cases [Palla et al., 2005]. Finally, the CPM allows overlap, i.e., a node can be a member of different communities

[3]CFinder: `http://www.cfinder.org/`

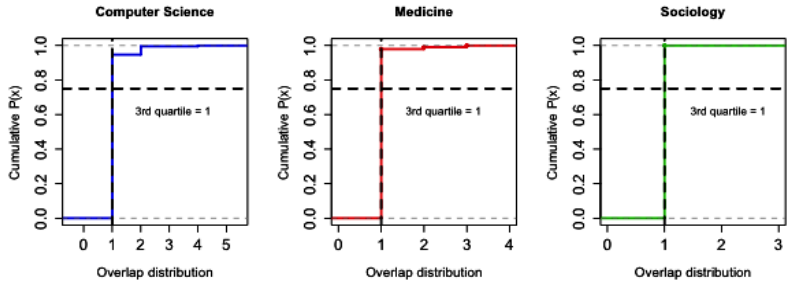


Figure 4.12: Empirical CDF of overlaps among communities detected by CPM.

at the same time, and communities overlap with each other by sharing nodes.

Figures 4.10 and 4.11 show the communities uncovered by the CPM (k=3) applied in the three co-authorship SN. Note that the clustering algorithm considers the network as unweight (as already explained in Section 4.4.2.1). The analysis of neighborhood overlap values in Figure 4.10 reveals that although the communities are formed by cliques, some of them have only weak ties (i.e., pairs of researchers weakly connected regarding neighborhood overlap): ten in computer science, two in sociology and seven in medicine. In other words, cliques formed by co-authorship of researchers do not have only strong interactions. Additionally, each community may have ties linking different cliques, and such ties are also weak in communities with only weak ties. Other communities have only strong ties: four in computer science, two in sociology and eight in medicine. It is also interesting to investigate these communities in order to identify patterns in the high cooperativeness. In the remaining communities, there is a mix of strong and weak ties. Furthermore, most communities have the median and mean different from the others, meaning that researchers have distinct behavior of co-authorship in each community.

Regarding the co-authorship frequency as a measure of the strength of tie, Figure 4.11 shows that most communities are composed by ties between researchers with small co-authorship frequency. In computer science and medicine, only one community has tie with co-authorship frequency greater than 10. In sociology, the co-authorship frequency does not reach five. Although the cliques compose such communities, the high connectivity among researchers groups does not indicate a strong intensity of co-authorship.

According to CPM definition, one researcher may be in more than one community. The number of overlaps is small in the three networks: in computer science, only one researcher is in four communities; in sociology, there is no overlap; and in medicine, one researcher is in three communities. Figure 4.12 presents the ECDF (Empirical Cumulative Distribution Function [Lewis and McKenzie, 1988]) of overlaps between

Table 4.7: Computer Science: Variation of neighborhood overlap and co-authorship frequency between pairs of researchers in different communities.

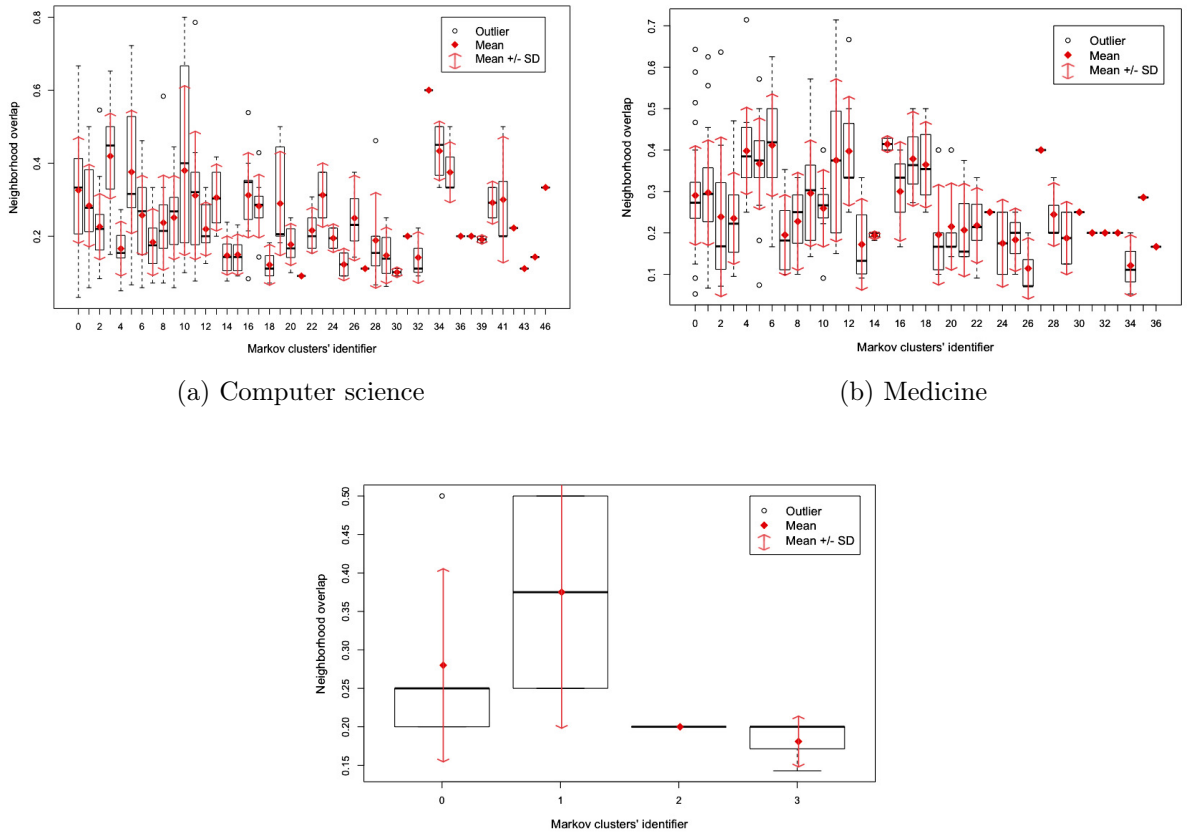
Researchers	Community	#Ties	MeanNO	VarianceNO	MeanW	VarianceW
Researcher A	5	2	0.0715	1.30E-05	1.5	0.5
	21	2	0.106	6.17E-005	1	0
Researcher B	5	2	0.0403	2.07E-05	3.5	12.5
	7	1	0.0556	0	4	0
Researcher C	4	2	0.0839	9.78E-05	1.5	0.5
	10	1	0.0999	0	5	0
	11	3	0.0999	0.005	1	0
	18	2	0.0742	1.51E-05	1.5	0.5
Researcher D	5	4	0.539	0.0021	1.75	2.25
	22	1	0.0435	0	6	0
Researcher E	12	1	0.111	0	1	0
	23	1	0.0714	0	2	0
Researcher F	8	1	0.2727	0	2	0
	13	1	0.1	0	3	0
Researcher G	5	1	0	0.0769	2	0
	21	2	0.1667	0.0062	2	2
Researcher H	14	3	0.333	0	1	0
	15	2	0.143	0	1	0
Researcher I	11	1	0.1111	0	1	0
	24	1	0.0714	0	3	0
Researcher J	5	1	0.0435	0	2	0
	7	1	0.0526	0	3	0

Table 4.8: Medicine: Variation of neighborhood overlap and co-authorship frequency between pairs of researchers in different communities.

Researchers	Community	#Ties	MeanNO	VarianceNO	MeanW	VarianceW
Researcher A	5	2	0.0997	0.0059	2	2
	8	2	0.2121	0.029	2	2
Researcher B	5	2	0.0943	0.0002	1.5	0.5
	9	1	0.04	0	4	0
	16	1	0.154	0	2	0

communities. We observe that 75% (third quartile) of the overlaps are lower or equal to one in the three networks, in which one overlap means the existence of overlap between two communities.

In order to analyze the strength of ties among researchers that are in more than one community and to identify differences of co-authorships with distinct groups, we measure the mean and the variance of neighborhood overlap and co-authorship frequency of the researcher's co-authorship intra each community. Tables 4.7 and 4.8 summarize the following information of researchers intra each community: community label, number of ties, mean and variance of neighborhood overlap of the researcher's ties, and the mean and variance of co-authorship frequency. Note that the researchers' names are not revealed. Analyzing the two tables does not allow to identify in which community a researcher publishes more. For instance, Researcher C has more ties (co-authorships) in community #11, but the mean neighborhood overlap is equal to community #10 and the mean co-authorship frequency is lower than community #10.



(a) Computer science

(b) Medicine

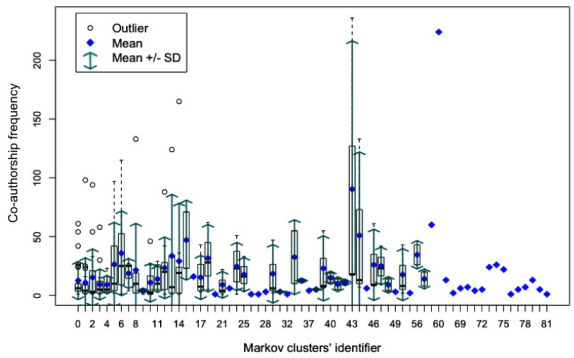
(c) Sociology

Figure 4.13: MCL – The strength of ties intra-communities measured by neighborhood overlap (clusters' identifiers in x axis are ordered by the size of communities).

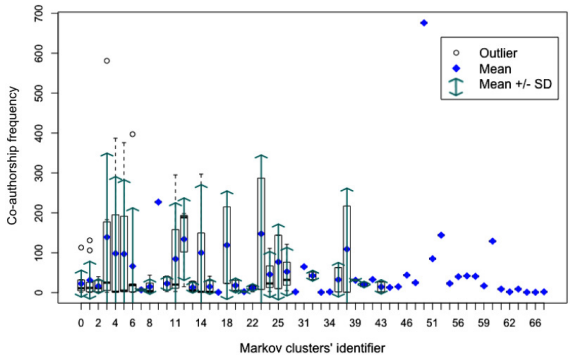
Moreover, the neighborhood overlap, the co-authorship frequency and the variance are low for most of researchers. This result suggests that researchers in overlaps of communities may have weak ties with other researchers and work as a bridge. It is interesting to investigate the properties of these researchers more thoroughly, which is left for future work.

4.4.2.3 Clustering with MCL Algorithm

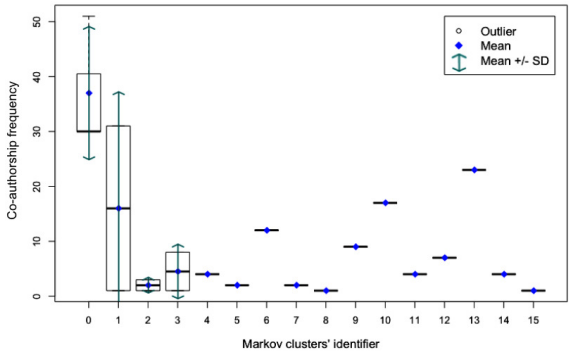
The Markov Cluster Algorithm (MCL) is an unsupervised clustering algorithm for graphs based on simulation of stochastic flow in graphs (known as network) [Van Dongen, 2000]. MCL deterministically finds cluster structures by computing the probability of random walks though the network. This process uses two operators called *expansion* and *inflation*, which are responsible for transforming one set of probabilities into



(a) Computer science



(b) Medicine



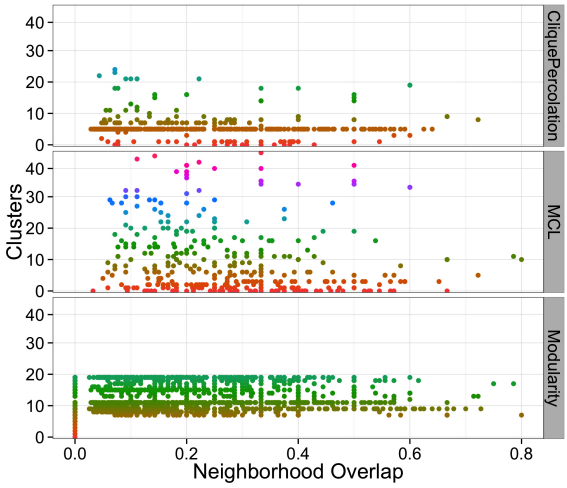
(c) Sociology

Figure 4.14: CPM – The strength of ties intra-communities measured by co-authorship frequency (x axis ordered by the size of communities; all outliers are present).

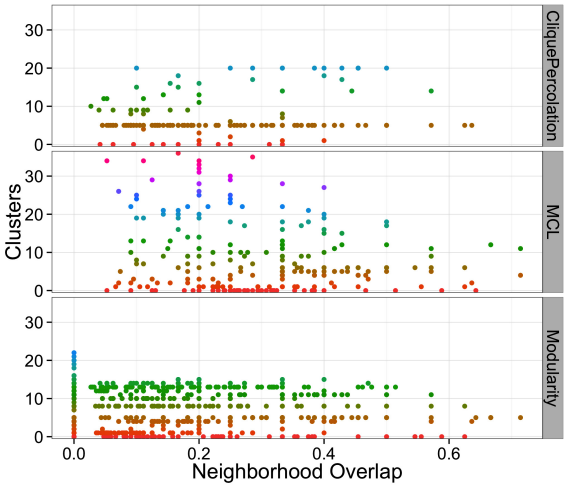
another. We use the algorithm available in Micans[4] and apply it to the computer science, sociology and medicine SN. Also, we keep the default values of the expansion and inflation parameters.

One type of input to MCL is a file describing the edges of a graph: the source and target nodes, and a co-authorship frequency as weight of the edges. The MCL interprets the co-authorship frequency of the edges as similarity to cluster the nodes. In order to understand how the neighborhood overlap and co-authorship frequency influence on clustering formation, we run the algorithm two times changing the value of the edge weight (one time the weights are equal to neighborhood overlap and another to co-authorship frequency). Using neighborhood overlap, the MCL has found 140 clusters in computer science, 35 in sociology and 139 in medicine. On the other hand, having the co-authorship frequency between researchers as edge weight, the MCL has

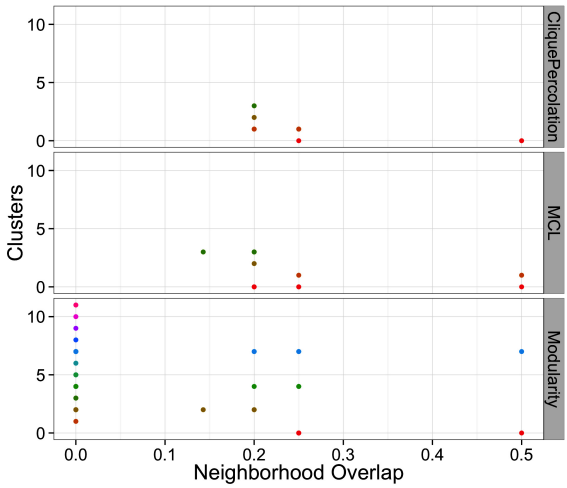
[4]Micans: http://micans.org/mcl/index.html



(a) Computer science



(b) Medicine



(c) Sociology

Figure 4.15: Comparing the results of the clustering methods and using neighborhood overlap to measure the strength of the ties. Each cluster is represented by a color, and each edge by a point.

detected 82 clusters in computer science, 16 in sociology and 68 in medicine. Some clusters are composed of only one node, and they are more present in clusters formed with neighborhood overlap as edge weight. This result indicates that the similarity among researchers is lower considering the neighborhood overlap than the co-authorship frequency.

Figures 4.13 and 4.14 show the results ordered by the size of communities when the neighborhood overlap and co-authorship frequency between researchers are considered as edge weight, respectively. Both figures do not include clusters with only one node for clarity. There are communities formed only by weak ties and also only by strong

ties, for example, the communities #25 and #34 in computer science, respectively. However, most communities include both types of tie. Considering the clusters size, the biggest communities are in the beginning of each graphic. For example, clusters #0 and #1 are the largest in sociology. The number of nodes in the largest clusters for neighborhood overlap and co-authorship frequency as edges weight is respectively 30 and 27 for computer science, four nodes (two communities of the same size) and four nodes (four communities of the same size) for sociology, 22 and 17 for medicine. Figure 4.13 also presents that the largest clusters are more formed by strong ties than weak ties, because the first quartile of these clusters is higher or equal to 0.2. Also, Figure 4.14 shows that the largest communities have too high co-authorship frequency, because the third quartile pass 30 in the three areas.

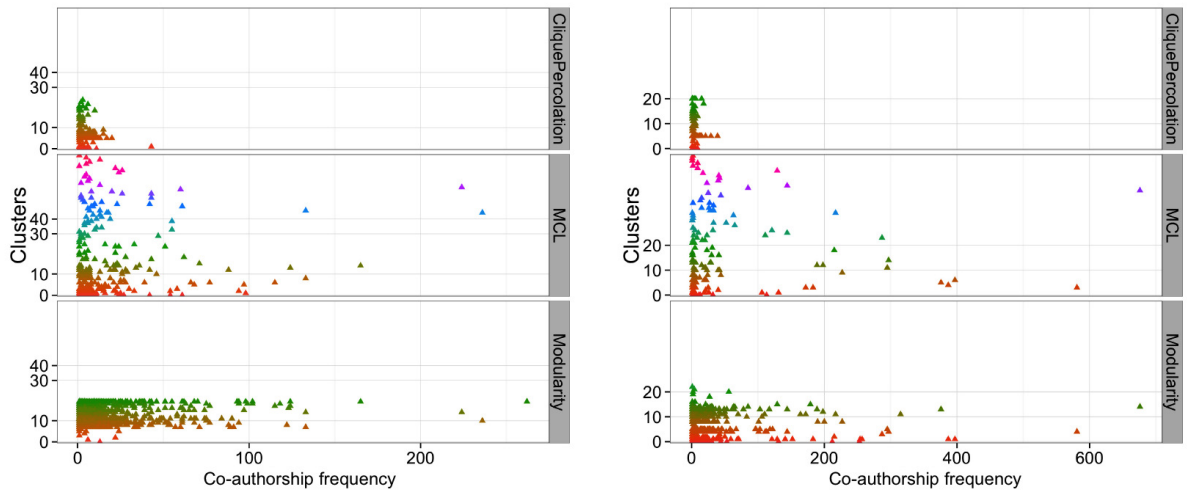
Lastly, MCL does not find ties connecting researchers from different communities for the three co-authorships social networks. This reveals that MCL provides a good clustering result since clustering algorithms minimizes inter-cluster edges [Malliaros and Vazirgiannis, 2013].

4.4.3 Comparative Analyses

This section compares the results of the methods used for finding communities (Louvain Method, Clique Percolation Method and Markov Cluster Algorithm). Figures 4.15 and 4.16 contrast the clusters (known as communities in LM and CPM methods) of each method regarding the neighborhood overlap and co-authorship frequency, respectively. We observe that LM tends to find less and larger clusters than the other two methods. Also, MCL detects a huge number of clusters, and some of them are singleton clusters[5]. In the co-authorship social network context, although CPM allows community overlaps, for it is common a researcher publishing with researchers from others communities, MCL provides the best clusters, because most of the detected communities are composed by strong ties.

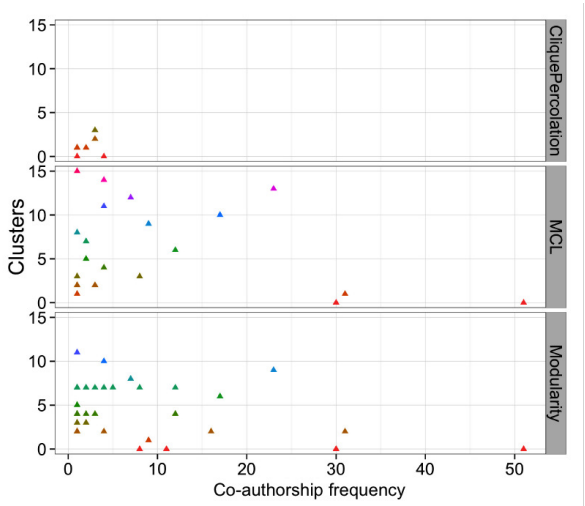
Figures 4.15 and 4.16 present the values of neighborhood overlap/co-authorship frequency of each pair of researchers (represented by each point) in each cluster for the three clustering algorithms. Moreover, Figure 4.15 shows a high concentration of edges until neighborhood overlap reaches 0.6 in computer science and medicine. In sociology, the maximum value of neighborhood overlap is 0.5, and there is more concentration of edges between 0.2 to 0.3. Also, note that CPM and MCL exclude edges with neighborhood overlap equal to 0. Some strong ties are also removed in

[5]There are some improvements of the MCL algorithm as proposed by Satuluri et al. [2010] and Satuluri and Parthasarathy [2009], but we did not use them in this work and left for future work.



(a) Computer science

(b) Medicine



(c) Sociology

Figure 4.16: Comparing the results of the clustering methods and measuring the strength of the ties with co-authorship frequency. Each cluster is represented by a color, and each edge by a triangle.

CPM, probably because these edges are in a 2-clique (we choose k=3 for CPM). Here, we emphasize that MCL is able of better differentiate the relationships putting them in distinct clusters. On the other hand, the concentration of points in CPM and LM do not show the same for these methods. Additionally, Figure 4.16 shows a high concentration of edges for co-authorship frequency less than 100 in computer science and medicine, and less than 10 in sociology. We also note that co-authorship frequencies equal to zero are not removed in any clustering method.

Overall, the three clustering algorithms form clusters with weak and strong ties. However, MCL is the best to detect clusters with more strong ties than weak ones.

Table 4.9: Properties correlation to the strength of ties.

Property	**CS**	**Med**	**Soc**
Clustering coefficient	SLC	SLC	SLC
Edge Betweeness	EC	EC	EC
Number of Triangles			SLC
Eigenvector			SLC
Closeness		EC	
Eccentricity		EC	

SLC = Strong linearly correlated, EC = Exponentially correlated

4.5 Concluding Remarks

In this chapter, we built non-temporal co-authorship networks from three areas to quantify the impact of properties on the strength of ties (neighborhood overlap). The characterization of neighborhood overlap in the three networks shows that the average value of this metric is around 0.2, i.e. the networks are formed more by weak ties. Additionally, our results showed that the Granovetter's theory governs the three networks and how topological properties are affected by removing weak and strong ties. Also, the correlation among topological properties and neighborhood overlap was different in each research area, as summarized by Table 4.9.

We have also evaluated each property for increasing the quality of the regression model. Out of them, the clustering coefficient and edge betweenness were related to neighborhood overlap in the three networks. Such result is trivial, because of the definition of neighborhood overlap. However, the most important contribution is the discovery of other properties related and non-related to the strength of ties, and whether the relations are linear or not. Such study can help to improve the quality of systems whose design considers the strength of ties, and to better understand the reasons for a tie being strong or weak.

Regarding the dynamism of tie strength intra and inter clusters, we applied three clustering algorithms in the three co-authorship SN from Lattes. We have applied the unweighted version of LM, and its evaluation results showed that such method identifies less clusters than the others. By applying CPM, we note that there was a small number of overlaps between communities, and researchers in the overlaps form weak ties (they work as bridges). Regarding MCL, we have applied the algorithm twice in each algorithm: one with NO as weight and another with W. MCL identified a large number of clusters than the other methods. Furthermore, the tie strength inter-communities tends to be weak for LM and CPM; whereas MCL algorithm does not find edges inter-communities. A main conclusion of these initial analyses of using neighborhood overlap and co-authorship frequency in clustering evaluation is: MCL is

the best clustering algorithm to be applied in co-authorship SN when compared to LM and CPM. Nevertheless, we also conclude that considering only the strength of ties metrics is not enough to define clustering qualities. Therefore, in the next steps, we plan to apply internal measures (like BetaCV, C-index, and so on) to compare with the results generated by the tie strength metrics.

Finally, in the three networks, the neighborhood overlap may be used to measure their strength of ties. In this chapter, we observe that neighborhood overlap captures not only the neighbors of a tie, but also the real intensity of co-authorship between pairs of researchers. Additionally, such metric is easy to compute because it requires only the topological structure of the networks. However, next chapter presents a few problems of using only neighborhood overlap or co-authorship frequency to measure the strength of ties.

Chapter 5

Tie Strength over Non-temporal Co-authorship Social Networks

In this chapter, we first identify problems in neighborhood overlap and co-authorship frequency metrics that complicate their sole use to measure the strength of co-authorship ties (e.g., presenting extreme values that do not represent reality). The existence of such problems suggests the metrics should be considered together or with other SN properties to better measure the strength of ties in non-temporal social networks. Therefore, we also propose a new metric, called *tieness* [Brandão et al., 2016; Brandão and Moro, 2017b], that helps to define a tie as weak or strong. Note the goal of tieness is not to replace neighborhood overlap and absolute frequency of interaction, but to be an *additional* feature that may allow deeper and complementary analyses.

In summary, tieness is an easy-computing metric that considers the neighbors and the intensity of co-authorships between researchers to measure tie strength. It differs from the existing ones by combining relevant aspects from the social network. Moreover, tieness can solve problems present in neighborhood overlap and weight (a simpler way to call absolute frequency of interaction), which have been largely used to measure tie strength [Easley and Kleinberg, 2010; Onnela et al., 2007]. It may also be applied to different social networks, not only co-authorship social networks, e.g., a movie producing network such as the one in [Viana et al., 2016].

After discussing methods (Section 5.1), we present the contributions of this chapter, summarized as follows:

– We discuss four case studies in which neighborhood overlap and absolute frequency of interaction alone have problems to measure the strength of ties. Also, we show the relationship between both metrics in three real datasets built from

digital libraries of distinct fields – Computer Science with DBLP, Medicine with PubMed and Physics with APS (Section 5.2).

– We propose a new metric called *tieness* that is a combination between a modification in neighborhood overlap and absolute frequency of interaction. It is easy to calculate and better differentiate tie strength in different levels. We also introduce a nominal scale to tieness based on the values of a modified neighborhood overlap and absolute frequency of interaction. Such nominal scale allows to identify when a tie is weak or strong and if it links researchers from different communities or not (Section 5.3).

– We validate tieness and its nominal scale according to Granovetter's theory by removing weak and strong ties (Section 5.4).

5.1 Methods Overview

The main goal of this chapter is to propose a new metric to measure the strength of co-authorship ties. In order to do so, we empirically evaluate four cases in which existing metrics commonly used to measure tie strength (neighborhood overlap and absolute frequency of interaction) present problems. Then, we propose our new metric called *tieness* focusing on solving these problems.

Next, we analyze the linear and non-linear correlation between neighborhood overlap (NO) and absolute frequency of interaction (W). The result of such correlation helps to identify whether both metrics are independent, i.e., whether they add or multiply when taken together. We do so by analyzing the relationship between both metrics on academic social networks from three different areas of expertise. The areas and their datasets are: (i) Computer Science given by DBLP; (ii) Medicine by PubMed; and (iii) Physics by APS. Then, we build a co-authorship SN for each dataset with features shown in Table 3.1 of Chapter 3.

Considering the four problem cases and correlation results, we propose tieness by combining a modification in neighborhood overlap and the absolute frequency of interaction. As neighborhood overlap is a normalized metric and absolute frequency of interaction is not, we have to normalize the latter before combining with a modification in neighborhood overlap. Thus, we guarantee that tieness is in the range $[0; 1]$.

In the following, we propose a nominal scale to tieness by analyzing the ECDFs (Empirical Cumulative Distribution Function [Lewis and McKenzie, 1988]) of neighborhood overlap, absolute frequency of interaction, modified neighborhood overlap and

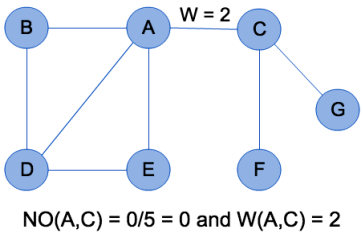


Figure 5.1: Case 1, no common co-author.

tieness for each social network. ECDF is a graph used to evaluate the data distribution, estimate percentiles and compare distinct distributions. The analysis of such graph reveals the percentile of data that falls below a specific value.

Finally, we validate such nominal scale by following Granovetter's theory, which claims that weak ties connect nodes from different communities (groups, clusters), whereas the strong ones link nodes from the same community. In other words, weak ties are acquaintances and provide access to novel information, while strong ties represent relationships with people whose social circles overlap. In order to follow this theory, we remove weak and strong ties at a time, and analyze the effect of such removals in the co-authorship social networks.

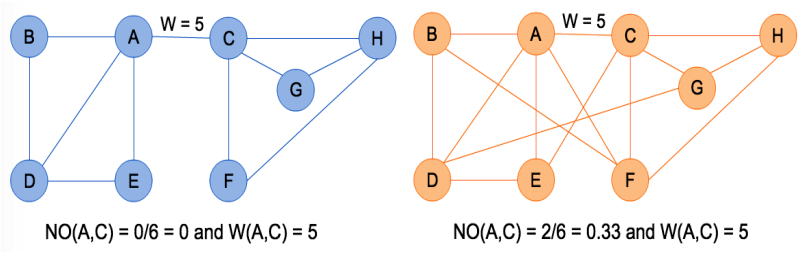
5.2 Neighborhood Overlap and Absolute Frequency of Interaction

In this section, we first present four cases in which neighborhood overlap and absolute frequency of interaction *cannot* be solely used to measure tie strength. Then, we empirically show their relationship on three different networks.

5.2.1 Four Motivating Cases

We have empirically studied different co-authorship social networks and identified four cases in which existing metrics cannot be solely used to measure tie strength.

Case 1: A pair of collaborators without any common neighbor. One of the problems of using only NO to measure the strength of ties is when an author has a high frequency of collaboration with another author, but they do not have any *common neighbor*. In this case, the NO is zero, which does not represent reality. Figure 5.1 exemplifies this case. Another problem here is that NO and W present contradictory results: analyzing NO, the pair AC is a bridge as the strength of co-authorship is very



(a) (b)

Figure 5.2: Case 2, no community information.

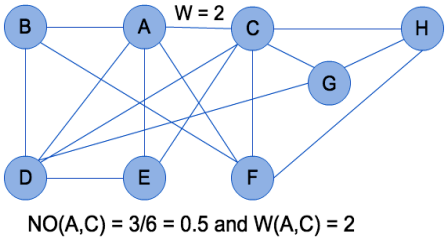


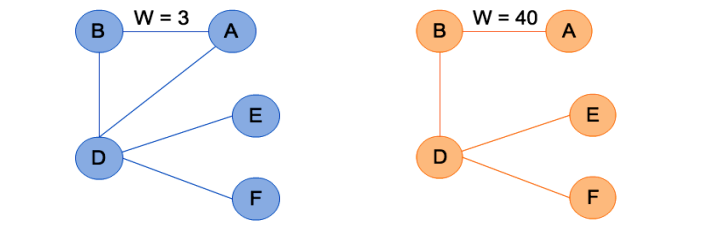
Figure 5.3: Case 3, many common co-authors.

weak; at the same time, W may indicate that such tie is not very weak. Therefore, considering both metrics is better to analyze how strong a tie is.

Case 2: Determining if two collaborators are from the same community or not. One problem in measuring the strength of ties using only W is that such metric provides a simple vision of the relationship. It is not possible to know if the relationship is intra a community or not. This case is exemplified by Figure 5.2, in which AC connects different communities (i.e. inter-community) in Figure 5.2a and AC are intra a community in Figure 5.2b, but in both contexts AC has co-authorship frequency equal to 5. Since ties with low W may be intra a community and ties with high W may be inter communities, using only W is not enough to asses how weak/strong a tie is (i.e., it does not allow to properly verify Granovetter [1973]'s theory, in which weak ties serve as bridges in the social network).

Case 3: Little collaboration between a pair of collaborators and plenty of common neighbors. In this case, NO and W give values with opposite meaning, i.e. high NO and low W. Such results make hard to define tie strength. Certainly, it depends on the analysis of the context. However, following Granovetter's theory, such tie should be strong. Figure 5.3 gives an example of this case for the edge AC.

Case 4: Results with extreme values. Here, the problem is when NO or W has extreme values that may not represent the reality. Figure 5.4a shows a maximum



(a) (b)

Figure 5.4: Case 4, results too small/high.

Table 5.1: The correlation coefficients between neighborhood overlap and co-authorship frequency. All p-values are smaller than 2.2e-16.

Dataset	Kendall	Pearson	Spearman
DBLP Articles	-0.049	-0.074	-0.062
DBLP Inproceedings	-0.023	-0.068	-0.029
PubMed	-0.032	-0.062	-0.03
APS	0.013	0.0003	0.016

value to NO, because the edge AB is part of a triad. Nevertheless, the value of W for the same edge is very small, which means that the tie is not necessarily very strong. Figure 5.4b presents a similar situation, but when W is very high and NO has the minimum value (zero) for the edge AB. In this case, defining a tie as weak or strong based on only one of the metrics may provide a misleading interpretation.

Based on these four cases, we claim that developing a new metric for tie strength is necessary. Then, after experimentally analyzing both metrics in Section 5.2.2., we introduce a new one in Section 5.3.

5.2.2 Analysis of NO and W over Different Networks

We now analyze the relationship between neighborhood and absolute frequency of interaction on DBLP, PubMed and APS. As we consider co-authorship social networks, we call absolute frequency of interaction as *co-authorship frequency*, which measures the amount of publications that a pair of researchers has co-authored. Table 5.1 presents the correlation between both metrics for each dataset considering three coefficients: Kendall measures the degree of non-linear dependence between two variables; Pearson evaluates the linear relationship between two variables; and Spearman measures the non-linear association between two variables [Abdi, 2007; Jain, 1991].

Overall, the correlation between neighborhood overlap and co-authorship frequency is small for the three coefficients. Therefore, they are monotonically and linearly *independent* in the three datasets. In other words, both metrics are important to measure the strength of ties as they capture *different* characteristics of the social network, as empirically discussed in Section 5.1.

5.3 Tieness: a New Metric for the Strength of Ties

Motivated by the problems generated by using neighborhood overlap and co-authorship frequency ($coAfrequency$ – a short name to the absolute frequency of interaction in the co-authorship social networks context) alone to measure tie strength, we now introduce a new metric called *tieness*. Specifically, tieness results from a combination between a modification in neighborhood overlap (entitled *modified neighborhood overlap*), which captures the social circle of nodes involved in a tie, and co-authorship frequency, which represents the absolute number of publications common to a pair of researchers, as shown by Equation 5.1.

$$tieness_{i,j} = \frac{|\mathcal{N}(v_i) \cap \mathcal{N}(v_j)| + 1}{1 + |\mathcal{N}(v_i) \cup \mathcal{N}(v_j) - \{v_i, v_j\}|} \; coAfrequency_{i,j} \tag{5.1}$$

where $\mathcal{N}(v_i)$ represents the co-authors (neighbors) of researcher v_i, and $\mathcal{N}(v_j)$ the co-authors of v_j. Note that we sum one at the numerator of neighborhood overlap to indicate that there is a link between v_i and v_j. This solves the problem when a pair of authors does not have any co-author in common. Then, we sum one at the denominator to give the right proportion to the equation. Also, for unweighted social networks, tieness value is the same as the modified neighborhood overlap.

Regarding computation time cost of tieness, the operations with the highest time cost are intersection ($O(\mathcal{N}(v_i)+\mathcal{N}(v_j))$) and union ($O(\min(\mathcal{N}(v_i), \mathcal{N}(v_j)))$) using hash tables. Thus, the time complexity of tieness is $O(\max(\mathcal{N}(v_i), \mathcal{N}(v_j)))$ — Big O notation property: $O(\min(\mathcal{N}(v_i), \mathcal{N}(v_j))) + O(\mathcal{N}(v_i)+\mathcal{N}(v_j)) = O(\min(\mathcal{N}(v_i), \mathcal{N}(v_j)) + \mathcal{N}(v_i) + \mathcal{N}(v_j)) = O(\max(\min(\mathcal{N}(v_i), \mathcal{N}(v_j)), \mathcal{N}(v_i), \mathcal{N}(v_j)))$ [Cormen et al., 2009].

A problem of Equation 5.1 is that $coAfrequency$ is a non-normalized metric, i.e., the set of weights of the datasets is not in the range 0 to 1. In order to solve this problem, we try to normalize $coAfrequency$ by using two methods: the norm (equal to the Eucliadian distance) of the set of weights that can be seen as a vector [Abdi and Williams, 2010], and the unity-based normalization[1]. However, the first

[1]Etzkorn, B. "Data normalization and standardization." BE BLOG [Online]. Available: `http://www.benetzkorn.com/2011/11/data-normalization-and-standardization` (2011).

method is not appropriate, because the norm of the $coAfrequency$ vector is very high, which reduces most of the weights to the magnitude of 10^4. Regarding the second method, it means to fit the data within unity (1), so all data will be in the range 0 to 1. However, sometimes it is important to choose a different range to the data. The unity-based normalization allows to normalize the data within a selected range. Thus, let the co-authorship frequency of all edges in the social network be defined as a vector $coAfrequency$ that represents each data point k (i.e., value of the edge). Then the unity-based normalization is computed by

$$||coAfrequency_{i,j}|| = a + \frac{(coAfrequency_k - \min(coAfrequency))(b-a)}{\max(coAfrequency) - \min(coAfrequency)} \tag{5.2}$$

where $coAfrequency_k$ is the k-value in the vector $coAfrequency$, $\min(coAfrequency)$ is the minimum value among all the set of co-authorship frequency in the social network (i.e. the minimum value in $coAfrequency$), and $\max(coAfrequency)$ is the maximum value among all the set of co-authorship frequency (i.e. the maximum value in $coAfrequency$). Moreover, a and b define the range of values for the co-authorship frequency, i.e the data will be normalized in that range. Here, we select $a = 1$ and $b = 2$, because considering the range $[0, 1]$ makes the value of neighborhood overlap be annulled when co-authorship frequency is 1 without the normalization. Thus, the range $[1, 2]$ guarantees that the co-authorship frequency can indeed contribute to increase the value of tieness.

Such improvement is presented in Equation 5.3, where $tieness_{i,j}$ is in the range $[0; 4]$. Then, we divide the equation by 4 to put $tieness_{i,j}$ in the range $[0; 1]$.

$$tieness_{i,j} = \frac{\frac{|\mathcal{N}(v_i) \cap \mathcal{N}(v_j)|+1}{1+|\mathcal{N}(v_i) \cup \mathcal{N}(v_j) - \{v_i, v_j\}|} \; ||coAfrequency_{i,j}||}{2} \tag{5.3}$$

where $||coAfrequency_{i,j}||$ is co-authorship frequency of a pair of researchers v_i and v_j as unity-based normalized by Equation 5.2.

Tieness is calculated for each edge (pair of nodes) in the social network. Let $tieness$ be a vector that contains $tieness_{i,j}$ for each edge k in the social network. Thus, the overall level of tieness in a social network is measured by the average of the tieness values of all edges:

$$\overline{tieness} = \frac{1}{|E|} \sum_{k=1}^{|E|} tieness_k \tag{5.4}$$

where $tieness_k$ is the value of tieness for each edge in the social network, and $|E|$ is

Table 5.2: Tieness for each case study and an extra case study representing the situation when NO and *coAfrequency* are in accordance. Note that *coAfrequency* is normalized considering only the values in the table to compute tieness. Thus, $min(coAfrequency) = 2$ and $min(coAfrequency) = 40$.

Case	Image	NO	coAfrequency	Tieness
Case 1: A pair of researchers without any common neighbor	Figure 5.1	0	2	0.085
Case 2: Determining if two researchers are from the same community or not	Figure 5.2a	0	5	0.075
Case 2: Determining if two researchers are from the same community or not	Figure 5.2b	0.33	5	0.23
Case 3: Little collaboration between a pair of researchers and a plenty of common neighbors	Figure 5.3	0.5	2	0.285
Case 4: Results with extreme values	Figure 5.4a	1	3	0.513
Case 4: Results with extreme values	Figure 5.4b	0	40	0.5
Regular Case: NO and coAfrequency in agreement	Figure 5.3 with $w = 12$	0.5	12	0.36

the number of edges in the social network. Also, the time complexity of the algorithm to measure the overall tieness is $O(|E|\ max(\mathcal{N}(v_i), \mathcal{N}(v_j)))$.

In order to understand how tieness represents ties in SN, Table 5.2 shows tieness' values for each case study. In Case 1, tieness gives a small value that indicates the presence of interactions (opposite of neighborhood overlap). For Cases 1, 2 and 3, tieness enables to infer if a pair of researchers is intra a community or not. Then regarding Case 4, tieness gives a high result when only w or only NO has an inflated value. In the Regular Case, when they are in accordance indicating that a tie is strong, tieness also provides a high value that may represent a strong tie.

Indeed, an advantage of using our new metric is the values of the strength of co-authorship ties are more distinct, then allowing to better differentiate the strength of a tie and establish different levels of tie strength. Moreover, we can consider the value of the modified neighborhood overlap and co-authorship frequency separately to evaluate the final result of tieness. Thus, the definition of a nominal scale is necessary to identify when a tie is weak or strong.

We define a nominal scale to tieness by comparing the modified neighborhood overlap and co-authorship frequency. In doing so, we follow concepts discussed by

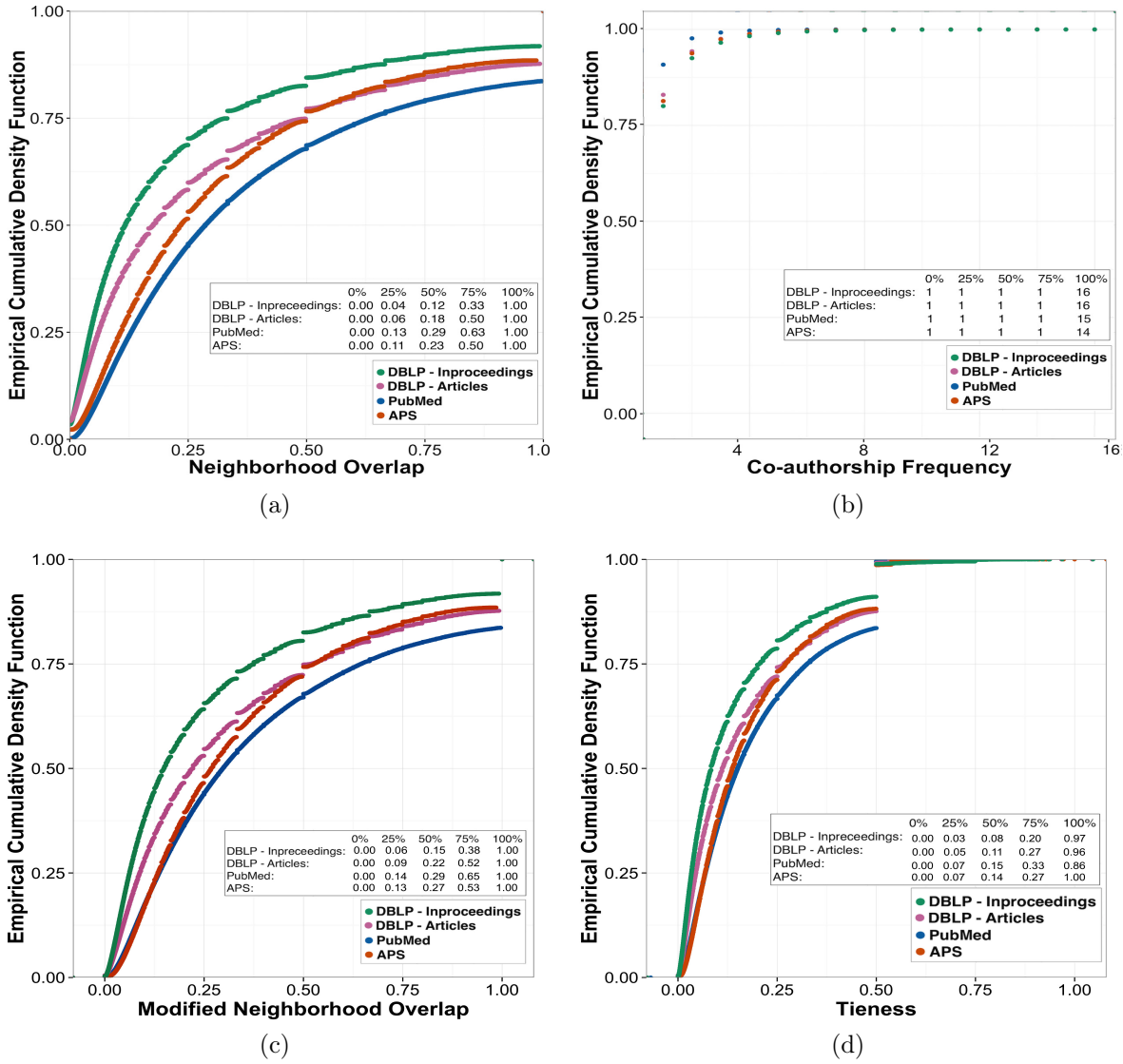


Figure 5.5: ECDF of each metric. In this scenario, modified neighborhood overlap and tieness metrics have more distinct values through the quartiles.

Easley and Kleinberg [2010]: a weak tie has a small neighborhood overlap and a strong tie has a large one.

Therefore, Figure 5.5 shows the ECDFs and quartiles for neighborhood overlap, co-authorship frequency, modified neighborhood overlap and tieness. The analysis of ECDFs shows that co-authorship frequency provides many repeated results to the strength of co-authorship ties, as 50% of data are equal to 1. On the other hand, the neighborhood overlap, modified neighborhood overlap and tieness provide different results for each quartile. Furthermore, considering the neighborhood overlap's ECDFs of each dataset, they are very different from each other. For example, the values of APS's ECDF are different from PubMed's ECDF. However, modified neighborhood

Table 5.3: DBLP Articles: Number of connected components when weak and strong ties are removed from the social network.

State of the SN	# edges	% edges	# c. components	# changes
Original	2,935,590	–	35,253	–
Tieness – weak ties removed	2,150,917	73.27	40,659	1.15
Modified NO – weak ties removed	2,259,563	76.97	37,848	1.07
Tieness – strong ties removed	2,276,784	77.56	22,301	0.63
Modified NO – strong ties removed	2,205,849	75.14	22,067	0.626

Note: c. components is connected components.

overlap and tieness ECDFs have similar values through different datasets. This result may indicate that tieness is less sensible to the dataset and better distinguishes the relationship between nodes.

Having studied such distributions, we may now consider the values of quartiles to define a nominal scale. In other words, the quartiles distributions help to identify when a tie is weak or strong, and if it connects different communities or not. Equation 5.5 shows the nominal scale to tieness based on the quartiles. Note for an unweighted social network, such scale is also valid because modified neighborhood overlap has the same value as tieness to the second and third quartile.

$$\begin{cases} \textbf{weak}, tieness \leqslant 0.10 \\ \textbf{moderate}, 0.10 < tieness < 0.43 \\ \textbf{strong}, 0.43 \leqslant tieness \end{cases} \tag{5.5}$$

5.4 Results and Discussion

In order to validate the proposed nominal scale, we verify if Granovetter's theory governs the social network and the strength of ties with such values. Given that weak ties are bridges that connect different parts of the network, his theory claims the network tends to be more disconnected when weak ties are removed (i.e., the number of connected components tends to increase). Hence, we analyze the number of connected components in the social network after removing weak and strong ties.

Tables 5.3-5.6 present the number of edges and connected components after removing weak and strong ties in each dataset. Also, we show results when the strength of ties is measured by tieness (weighted SN) and modified neighborhood overlap (considering the SN as unweighted). According to these tables, when weak ties are removed, the number of connected components is higher than when removing strong ties. Also, there are differences between the result for modified neighborhood overlap and tieness, which is caused by the co-authorship frequency of interaction. Moreover, the number

Table 5.4: DBLP Inproceedings: Number of connected components when weak and strong ties are removed from the social network.

State of the SN	# edges	% edges	# c. components	# changes
Original	3,760,247	–	28,168	–
Tieness – weak ties removed	2,227,898	59.24	44,334	1.57
Modified NO – weak ties removed	2,396,706	63.74	41,012	1.46
Tieness – strong ties removed	3,128,445	83.19	17,281	0.61
Modified NO – strong ties removed	3,061,258	81.41	17,125	0.61

Note: c. components is connected components.

Table 5.5: PubMed: Number of connected components when weak and strong ties are removed from the social network.

State of the SN	# edges	% edges	# c. components	# changes
Original	5,550,294	–	8,926	–
Tieness – weak ties removed	4,485,605	80.82	10,600	1.19
Modified NO – weak ties removed	4,583,204	82.58	10,517	1.18
Tieness – strong ties removed	3,577,424	64.45	3,453	0.39
Modified NO – strong ties removed	3,481,747	62.73	3,447	0.39

Note: c. components is connected components.

Table 5.6: APS: Number of connected components when weak and strong ties are removed from the social network.

State of the SN	# edges	% edges	# c. components	# changes
Original	821,870	–	4,957	–
Tieness – weak ties removed	676,768	82.34	5,846	1.18
Modified NO – weak ties removed	705,020	85.78	5,442	1.1
Tieness – strong ties removed	611,732	74.43	2,931	0.59
Modified NO – strong ties removed	580,663	70.65	2,869	0.579

Note: c. components is connected components.

of removed edges is larger when weak ties are removed. Indeed, the larger number of connected components may be explained by the larger removal of bridging edges.

We now compare the proportion of the number of connected components by the number of edges for tieness and modified neighborhood overlap when weak and strong ties are removed from the social network. Table 5.7 presents these proportions. The analysis of such proportions shows that the number of connected components per edge is larger when weak ties are removed. Thus, the nominal scale is valid. Moreover, as the removal of weak ties (defined according to the nominal scale) breaks the connected components of the social network, tieness is indeed able to identify when a tie connects different communities or not.

Furthermore, we note that the different research areas considered (computer science, medicine and physics) present similar behavior. The presence of weak ties is bigger than the strong ones when they are measured by tieness. This is a result from a network with nodes not very well clustered (regarding their neighbors). In order to

Table 5.7: Proportion between the number of connected components and the number of edges in the social networks when weak and strong ties are removed.

	Tieness		**Modified neighborhood overlap**	
Datasets	**#cc/#NW ties**	**#cc/#NS ties**	**#cc/#NW ties**	**#cc/#NS ties**
DBLP Articles	40,659/2,150,917 =0.02	22,301/2,276,784 =0.01	37,848/2,259,563 =0.02	22,067/2,205,849 =0.01
DBLP Inproceedings	44,334/2,227,898 =0.02	17,281/3,128,445 =0.005	41,012/2,396,706 =0.02	17,125/3,061,258 =0.005
PubMed	10,600/4,485,605 =0.002	3,453/3,577,424 =0.001	10,517/4,583,204 =0.002	3,447/3,481,747 =0.001
APS	5,846/676,768 =0.01	2,931/611,732 =0.005	5,442/705,020 =0.01	2,869/580,663 =0.005

Note: cc is connected components, NW and NS are all non weak and non strong ties, respectively.

verify it, we analyze the clustering coefficient[2] from the four co-authorship social networks. The results show that the highest clustering coefficient is from PubMed (equal to 0.357), and the smallest one is from DBLP Inproceedings (equal to 0.16). Thus, the clustering coefficient from the four networks are very small, which justifies the low tieness for the pairs of researchers.

Although tieness is able to better differentiate the strength of ties when compared to neighborhood overlap and co-authorship frequency, there are limitations. One of them is that tieness classifies a tie as strong when the modified neighborhood overlap and weight are very high. Thus, few ties are classified as strong. A solution to this is changing the nominal scale, but it requires to make more analyses over the social networks. Another limitation is applying tieness in co-authorship social networks from research areas in which collaborations among researchers are *not* a common practice. For example, in sociology area, the level of collaboration is low [Brandão and Moro, 2015]. Nonetheless, this is a limitation intrinsic to the definition of co-authorship networks, which should contain a good number of connections for any proper analysis.

Moreover, defining a nominal scale is very hard, because it requires to consider different parameters from the data. Here, the nominal scale of tieness has a simplifying assumption: to consider only the values of the ECDFs and percentiles. Another possibility is to define the nominal scale by combining different properties from the ties in the social networks with tieness in a math model. Then, the nominal scale would be more complete, but more complex s well.

5.5 Concluding Remarks

In the context of academic social networks, we identified problems with using solely a modification in neighborhood overlap and absolute frequency of interaction to measure

[2]Clustering coefficient measures the proportion of nodes neighbors that can be reached by other neighbors [Easley and Kleinberg, 2010], i.e. it also considers the connectivity among neighbors

the strength of co-authorship ties. Then, we presented a new metric to measure such ties strength, called tieness, which has relatively low computational cost and can be applied to other social networks types (since tieness is a topological feature). Also, the definition of tieness comes with a nominal scale that allows to identify when a tie is weak or strong and if it links researchers from different communities or not. The main limitation to such a new metric is that the social network must have nodes collaborating with each other.

We have performed empirical studies by considering the networks from three different areas of expertise (Computer Science, Medicine and Physics). Overall, our analyses showed that tieness provides more distinct values through the ties than neighborhood overlap and absolute frequency of interaction. Such distinction is important to better compare how strong (weak) a tie is regarding another one. We also observed similar behavior through the three different research area.

Furthermore, all the four co-authorship social networks are dominated by the presence of weak ties. This is so, because most pairs of researchers have low amount of shared neighbors and small co-authorship frequency of interaction. Therefore, tieness is able to classify as strong ties only pairs of researchers with very high neighborhood overlap and co-authorship frequency.

Chapter 6

Tie Strength over Temporal Co-authorship Social Networks

Social networks represent relationships and interactions among individuals. Studying their models and patterns allows to solve different problems, such as evaluating enterprise security [Abraham, 2016], predicting the potential location of faults in softwares [Chen et al., 2008], ranking graduate programs [Lopes et al., 2011] and identifying user reputation [Yasin and Liu, 2016]. In this context, it is essential to understand the relationship between people [de la Maza, 2007], such as analyzing the progress of relationships over time and/or their strength.

Indeed, time is a fundamental factor when characterizing the nature and the strength of relationships. For example, acquaintances might become friends (and vice-versa), acquaintances might turn into co-workers, and so on. Such time-varying relationships may be modeled as a temporal social network (SN), or temporal graph, where each node is a person and there is an edge between two nodes in a given time if they share any particular relationship in that time.

Unfortunately, most studies have focused on static (non-temporal) aggregated graphs [Brandão and Moro, 2015; Castilho et al., 2017; Chen et al., 2008; Granovetter, 1973; Koo, 2016; Onnela et al., 2007; Rana et al., 2014], in which the type and the strength of the edges are invariant and usually constructed from a fixed history of interactions between the nodes. Such (static-based) approaches give the same degree of importance for all previous interactions. However, the most recent ones are usually more representative of the class of the relationship (defined according its strength) than the older ones [Gilbert and Karahalios, 2009]. If in static graphs such temporal aspects are aggregated, and therefore hidden, in temporal graphs they come naturally, serving as an appropriate model for dynamic social networks.

Nonetheless, computing temporal social networks properties and their time-varying behavior is very challenging. For example, the clustering coefficient of a network in time t_1 is not necessarily the same in time t_2, as interactions may appear or perish over time. Also, the precise temporal ordering of the edges essentially influences the notion of node adjacency and reachability in such networks [Nicosia et al., 2013]. Hence, concepts and metrics designed and applied to analyzing static networks must be adapted and extended to time-varying networks. Tie strength (a.k.a. strength of the ties) is one of those concepts, which is originally defined as a merge of the time of relationship, the emotional force, the intimacy, and the reciprocal services that represent a link (tie) between people [Granovetter, 1973].

Indeed, we propose a new algorithm entitled *STACY - Strength of Ties Automatic-Classifier over the Years*. STACY is an algorithm that uses social network features to classify the strength of ties in eight different classes (strong, bridge+, bridge, transient, periodic, bursty, weak and random). Moreover, STACY is based on an existing algorithm entitled RECAST (*Random rElationship ClASsifer sTrategy*) [Vaz de Melo et al., 2015] that have been applied to measure the strength of ties in mobile networks. Thus, RECAST has not been used in co-authorship social networks yet. Here, we improve the performance of such algorithm (called as fast-RECAST) and compare it results with the ones generated by STACY.

In this research, we view the strength of a tie as the likelihood of its (re) appearance in the future. We estimate such likelihood by using three social network edge features related to tie strength (edge persistence, neighborhood overlap and co-authorship frequency). Furthermore, we contrast the results by estimating such likelihood with edge persistence and topological overlap (both are considered in fast-RECAST). These properties capture the regularity of the interaction and the similarity between individuals involved in such interaction.

Thus, our main goal is to verify if current definitions of tie strength hold for temporal social networks. To do so, we analyze the dynamism of tie strength in such networks by observing link persistence and link transformation over time [Gilbert and Karahalios, 2009; Vaz de Melo et al., 2015]. This goal is specified as tasks driven by the following research questions: How is tie strength defined for temporal networks? and How much does the strength of ties vary over time? Such research questions were introduced in Chapter 1 as *RQ3* and *RQ4*.

In order to solve such questions, Figure 6.1 presents a general view of our work.

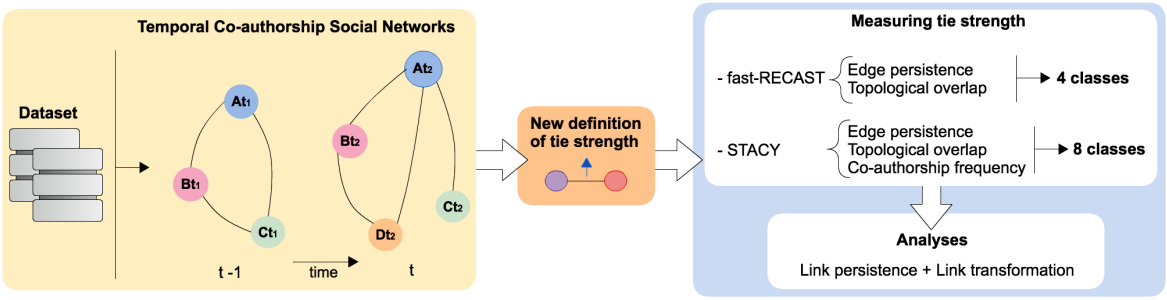


Figure 6.1: Main steps to analyze the link persistence and link transformation through different tie strength classes.

Considering three real datasets (DBLP[1], PubMed[2] and APS[3]), we built four temporal co-authorship social networks (we divide DBLP in two social networks). Note that ties may appear and disappear over time and also may become weak or strong. Focusing on such kind of temporal network, we check if a definition of tie strength is resilient to temporal networks. Moreover, we add multiprocessing features to an existing algorithm to classify the ties as strong, weak, bridges or random in considerably much less time. Also, we propose a new algorithm to classify the ties as strong, bridge+, bridge, transient, periodic, bursty, weak or random. Finally, such classification is also input to the analysis step that considers our proposed definition of tie strength.

Next, Section 6.1 describes the temporal SN models and the original algorithm (called RECAST) to classify the strength of the ties. Section 6.2 presents a definition of tie strength, the new version of the algorithm RECAST (called fast-RECAST), and our new algorithm (called STACY) to measure the strength of ties in large SN. Section 6.3 characterizes STACY classes according to the number of researchers' publications (Section 6.3.2), details our results, and describes how we derive a new computational model (temporal_tieness) from STACY to directly classify tie strength (Section 6.3.4). Finally, Section 6.4 presents the main conclusions of this chapter.

6.1 Fundamental Concepts

Here, we describe the temporal social networks models (Section 6.1.1) and the original RECAST (Section 6.1.2).

[1]DBLP: `dblp.uni-trier.de`
[2]PubMed: `www.nlm.nih.gov/news/medlinedata.html`
[3]APS: `publish.aps.org/datasets`

6.1.1 Temporal Social Networks Models

In order to proceed, we first need to formally define a model for temporal social networks. Instead of proposing a new one, we borrow the ideas from Vaz de Melo et al. [2015], who have modeled a temporal network for studying mobile networks. Therefore, we associate a start time and a duration to each co-authorship. Then, a temporal co-authorship social network is modeled as a graph $G_k(\mathcal{V}_k, \mathcal{E}_k)$ in which time is discretized into steps of duration δ, and k is the time step in which a co-authorship (encounter) occurs. Here, we consider a duration of $\delta = 1$ year, as this is the common granularity for publications (not month or day). Also, if $\delta = 1$ month, there would be no connection between the co-authors in many time steps within one year. The set of nodes $\mathcal{V}_k$ is formed by all network nodes in a co-authorship during the k-th time step, and the set of edges $\mathcal{E}_k$ is composed of co-authorships during the same time step. Thus, there is an edge in $\mathcal{E}_k$ between two nodes i and j with $i, j \in \mathcal{V}_k$, if i and j have co-authored a publication during time step k.

Given an undirected graph $G = (\mathcal{V}, \mathcal{E})$, in which $\mathcal{V} = \{v_1, ..., v_n\}$ is the set of vertices and $\mathcal{E} = \{e_1, ..., e_m\}$ is the set of edges that represent interactions between vertices. A time-varying representation of the co-authorship SNs can be defined by a temporal accumulation graph $G_t(\mathcal{V}_t, \mathcal{E}_t)$ in G that is the aggregation of interactions in each k discrete time step until t. Thus, all vertices interact until t-th time step for a given value of $\mathcal{V}_t$. All edges in the set $\mathcal{E}_t$ represent interactions between vertices (v_i, v_j) during each k time step until t. Since G_t accumulates all co-authorships from the datasets and evolves over time, such aggregate graph contains social and random encounters (relationships).

Also according to Vaz de Melo et al. [2015], a random version G_t^R of the temporal aggregated graph G_t is necessary to analyze the patterns of such network. The random graph must have similar social network topological features as the G_t graph (number of nodes, edges, and empirical degree distribution) and the nodes are connected in a different way from G_t.

For this model to work, it requires two central pieces: a definition of tie strength in temporal networks and an algorithm that implements it. Next, we present its original algorithm (RECAST) in Section 6.1.2. Then, we present a definition for tie strength, a multiprocessing version of RECAST, and our new algorithm (STACY) to classify tie strength in Section 6.2.

6.1.2 The Original RECAST

Following the model description, this section overviews its original implementation algorithm, called RECAST [Vaz de Melo et al., 2015], which was applied in Dynamic Complex Wireless Networks (DCWN). One contribution of our work is to modify it to measure the strength of ties in large temporal SNs. We chose RECAST because it is the only one that attributes different classes to the tie strength in temporal networks. Such algorithm was originally applied in relatively small mobile networks to classify users' wireless interactions differentiating random interactions from the social ones (friends – called as strong, bridges and acquaintances – called as weak).

According to Vaz de Melo et al. [2015], any system is susceptible to random events and irrational decisions called semi-rational decisions. Nevertheless, conscious decisions still govern most of the interactions. Thus, the evolution of social networks (specially, DCWN) is considerably different from the growth of random networks, as Erdos and Rényi networks [Erdös and Rényi, 1959]. Indeed, social networks that model real interactions have edges created from semi-rational decisions (i.e., such edges tend to be regular and repeat over time), whereas random networks have edges with the same probability of connecting any two nodes. In other words, a person may take a social or a random decision. Naturally, if such person has a probability of executing a social decision higher than a probability of taking a random one, the network tends to be a well-structured social network. If the opposite happens, the network tends to be a random network.

Another aspect that differentiates social networks from the random ones is the presence of communities [Vaz de Melo et al., 2015], which cannot be found in random networks. Thus, the clustering coefficient has been largely considered to distinguish random from social networks. Specifically, considering a random network G^R built with the same number of nodes, edges and empirical degree distribution of its social counterpart G, the clustering coefficient of G^R is one order of magnitude smaller than the clustering coefficient of G [Watts and Strogatz, 1998]. Hence, a network G with clustering coefficient significantly higher than that of its random equivalent G^R has individuals that made (part of) non-random decisions.

RECAST considers such concept of social and random networks and implements the model described in Section 6.1.1 by building both G_t and G_t^R. Two algorithms are necessary to generate G_t^R from G_t: `RND` and `T-RND` [Vaz de Melo et al., 2015]. Given a graph $G(\mathcal{V}, \mathcal{E})$, `RND`$(G)$ returns a random graph $G_t(\mathcal{V}^R, \mathcal{E}^R)$ with the same number of nodes, number of edges and degree distribution as G. Then, the only difference between G and G^R is the connection among nodes, which is the focus of

our study. Therefore, RND attributes an edge between nodes i and j with probability $p_{i,j} = (d_i \times d_j)/\sum_{k=1}^{|V|} d_k$, in which the degree distribution is $D = (d_1, d_2, ..., d_n)$ of G with n nodes. The second algorithm T−RND is an extension of RND and generates random graphs for temporal networks G_t. Thus, T−RND$(G_1 \cup G_2 \cup ... \cup G_t)$ receives a set of consecutive event graphs G_t and returns a random temporal graph G_t^R. Such algorithm builds G_t^R by running RND in each event graph G_t and then accumulating it as $G_t^R =$ RND$(G_1) \cup$ RND$(G_2) \cup ... \cup$ RND(G_t). In summary, both RND and T−RND randomly reproduce the total number of co-authors with distinct authors each person had in a snapshot.

RECAST considers two SN features to identify social relationships:

(i) edge persistence maps the regularity of relationships

$$per_t(i,j) = \frac{1}{t}\sum_{k=1}^{t}[(i,j) \in \mathcal{E}_k]([(i,j) \in \mathcal{E}_k] \tag{6.1}$$

where $per_t(i,j)$ is 1 if there is an edge (i,j) in $\mathcal{E}_k$ at time k (0 otherwise) and complementary cumulative distribution function (CCDF) $\overline{F}_{per(i,j)}(x) = P[per_t(i,j) > x]$;

(ii) topological overlap (a.k.a. neighborhood overlap) represents the individuals similarity

$$to_t(i,j) = \frac{|k|(i,k) \in \mathcal{E}_t \cap k|(j,k) \in \mathcal{E}_t|}{|k|(i,k) \in \mathcal{E}_t \cup k|(j,k) \in \mathcal{E}_t|} \tag{6.2}$$

and the CCDF $\overline{F}_{to(i,j)}(x) = P[to_t(i,j) > x]$. Note that this metric is the same as presented in Chapter 4. We just rewrite to be in the context of RECAST.

Furthermore, RECAST has a single parameter p_{rnd} to distinguish *social* (friends, bridges and acquaintances) from *random* values of the SN features. Thus, Vaz de Melo et al. [2015] identify the feature value $\overline{x}$ that represents a threshold, such that feature values greater than $\overline{x}$ happen with a probability lower than p_{rnd} in G_t^R. Also, for small values of p_{rnd}, feature values higher than $\overline{x}$ are very improbable to occur in a random network, happening mostly due social relationships. Also, the parameter p_{rnd} can be interpreted as the expected classification error percentage.

6.2 Measuring Tie Strength

We now revisit the concept of tie strength (Section 6.2.1) and propose fast-RECAST, an extended RECAST with multiprocessing modules to classify ties (Section 6.2.2). Then,

we propose *STACY*, an algorithm that uses edge persistence, neighborhood overlap and co-authorship frequency to classify tie strength (Section 6.2.3).

6.2.1 Revisiting the Concept of Tie Strength

Given a temporal graph $G_k(\mathcal{V}_k, \mathcal{E}_k)$, where k is the time step in which a co-authorship occurs, a tie (i, j) is likely to be strong if it is present in G_k for most values of k. On the other hand, the tie (i, j) is likely to be weak if it is present in G_k for just a few values of k. In other words, strong ties are likely to persist over time, and weak ties probably occur sporadically. Another characteristic of a strong tie (i, j) is that probably i and j have many neighbors in common. As previously discussed, nodes that have many neighbors in common are more likely to persist over time.

Given these two features, we group ties into the four classes of relationship given by fast-RECAST, namely *strong* (friends), *weak* (acquaintances), *bridges* and *random*. Each class gives a level of tie strength: *strong* are ties that persist over time and share many neighbors in common, *weak* do not persist over time, but share many neighbors in common, *bridges* persist over time but share at most a few neighbors in common, and *random* do not persist over time and share at most a few neighbors in common. Hence, using these four classes of relationships, we investigate if the strength of ties are likely to transform over time. With such analysis, we are able to go deeper into temporal social networks and answer questions such as: are *strong ties* more likely to remain *strong* in the future? Are *weak ties* more likely to become *strong ties* or to become *random*?

Moreover, considering a third feature, co-authorship frequency, a strong tie (i, j) is that probably i and j have a high frequency of co-authorship. STACY uses these three features to classify ties into the eight classes of relationship. Each class represents a level of the strength of ties that is better defined in Section 6.2.3.

6.2.2 Multiprocessing RECAST

The construction of G_t^R using `T-RND` increases the complexity of RECAST to $O(t \times (|\mathcal{V}_t| + |\mathcal{E}_t^R|))$. Then, we propose to apply a multiprocessing Pool module from Python (a module based on communicating processes for writing concurrent programs[4]) in such step of RECAST in order to reduce its complexity. We call this novel, multiprocessing algorithm as fast-RECAST.

[4]Multiprocessing with python: `docs.python.org/2/library/multiprocessing.html`

Algorithm 1 Multiprocessing RECAST (fast-RECAST): a parallelized code to classify edges of G_t as random or social – strong, weak or bridge.

Require: $p_{rnd} \geq 0$
1: **return** $class(i,j)\ \forall (i,j) \in U_t E_t$
2: Construct G_t^R and set $\texttt{RND}(G_1), ..., \texttt{RND}(G_t)$ using **T-RND** with **pool.map_async**
3: Get $\overline{F}_{to}(x)$ and $\overline{F}_{per}(x)$ from G_t^R using **pandas dataframe**
4: Get $\overline{x}_{to}|\overline{F}_{to}(\overline{x}_{to})$ and $\overline{x}_{per}|\overline{F}_{per}(\overline{x}_{per}) = p_{rnd}$ with **pool.apply_async**
5: **for all** edges$(i,j) \in E_t$ **do**
6: **if** $per(i,j) > \overline{x}_{per}$ and $to(i,j) > \overline{x}_{to}$ **then**
7: $class(i,j) \longleftarrow Strong$
8: **else if** $per(i,j) > \overline{x}_{per}$ and $to(i,j) \leq \overline{x}_{to}$ **then**
9: $class(i,j) \longleftarrow Bridges$
10: **else if** $per(i,j) \leq \overline{x}_{per}$ and $to(i,j) > \overline{x}_{to}$ **then**
11: $class(i,j) \longleftarrow Weak$
12: **else**
13: $class(i,j) \longleftarrow Random$

The idea is that more than one random event graph G_t^R is built at a time in a multi-core computer. Thus, the new computational cost is $O(\frac{t}{p} \times (|\mathcal{V}_t| + |\mathcal{E}_t^R|))$, where p is the number of processes. After building G_t^R, the complexity of the classification is $O(|E_t^R| \times |\mathcal{V}_t|)$, in which $O(|V_t|)$ is the cost of computing the two SN features of an edge. We also add a multiprocessing Pool module from Python to call the functions to compute the edge persistence and topological overlap from the aggregated graphs. Both features are computed in parallel and asynchronously.

Algorithm 1 summarizes the code for fast-RECAST[5] with multiprocessing Pool module (lines 2 and 4) and an optimization in the memory use by applying pandas dataframe from python to store the graphs before processing them (line 3). Also, as we consider co-authorship social networks, we rename the social edges from friends to strong ties and acquaintances to weak ties.

In order to show that fast-RECAST performs better than RECAST, we measure the execution time of both algorithms in a laptop with 8 GB 1600 MHz DDR3 of memory and 2.5 GHz Intel two Core i5 of processor. The operation system is Mac OS X El Capitan version 10.11.6. Figure 6.2 presents the execution time in seconds of fast-RECAST and RECAST. Note that we present the results only for PubMed dataset, because it is the largest one. The behavior of fast-RECAST and RECAST regarding the execution time is similar for the other datasets (APS, DBLP Articles and DBLP Inproceedings). The percentage represents the amount of data that we consider

[5]Source code available in `http://homepages.dcc.ufmg.br/~mirella/projs/apoena/datasets.html`

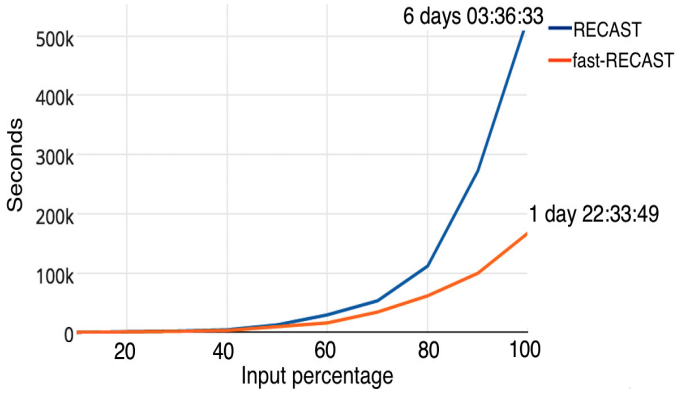


Figure 6.2: The performance of RECAST and fast-RECAST for PubMed dataset (the largest one).

at a time. For example, 10% indicates that we consider only 10% of PubMed dataset to measure the execution time of both algorithms. Then, we take 20% of the dataset and so on. Figure 6.2 reveals that, as the input file grows, fast-RECAST becomes significantly faster than RECAST.

6.2.3 STACY

Now, we propose a new algorithm to automatically classify tie strength called as *STACY - Strength of Ties Automatic-Classifier over the Years*. STACY is the renovation of fast-RECAST by considering the edges weight (co-authorship frequency) different from 1. To distribute the co-authorship frequency in the random graph G_t^R, we use the same algorithm to distribute the edges degree proposed by Miller and Hagberg [2011]. However, instead of attribute edge weight as 1, we randomly attribute the co-authorship frequency from the weighted temporal graph provided as input of STACY. Thus, a weight is attributed to an edge between i and j with probability $p_{ij} = (w_i \times w_j)/\sum_{k=1}^{|V|} w_k$ for a weight distribution $D_w = (w_1, w_2, ..., w_n)$ of G with n nodes [Chung and Lu, 2002; Miller and Hagberg, 2011].

Following the description of G_t in Section 6.1.1, we now define a weighted temporal accumulation graph $G_t^W(\mathcal{V}_t, \mathcal{E}_t)$, where $G_t^W = G_1^W \cup G_2^W \cup ... \cup G_t^W$. Then, $\mathcal{V}_t$ and $\mathcal{E}_t$ are the set of all nodes and weighted edges in the networks, respectively, in the time step 0 to t. Since G_t^W accumulates all co-authorships from the datasets and evolves over time, such aggregate graph contains social and random encounters. Also, we consider a weighted random version $G_t^{R,W}$ of the weighted temporal aggregated graph G_t^W, which is necessary to analyze the patterns of such network. The random graph must have similar topological features as the G_t^W graph (number of nodes, edges, and empirical degree distribution), the nodes are connected in a different way from G_t^W

Table 6.1: STACY relationship classes.

Class	Edge persistence	Neighborhood overlap	Co-authorship frequency
Class1 - strong	social	social	social
Class2 - bridge+	social	random	social
Class3 - transient	random	social	social
Class4 - periodic	social	social	random
Class5 - bursty	random	random	social
Class6 - bridge	social	random	random
Class7 - weak	random	social	random
Class8 - random	random	random	random

and the weight (co-authorship frequency) are randomly distributed through the edges. It is important to emphasize that the co-authorship frequency of each edge in G_t^W is the sum of all co-authorship frequency from each time step.

Our new algorithm classifies the edges in eight different classes: seven social and one random. The eight classes of relationship are described in Table 6.1. A social network property with value equal to "social" indicates an almost zero probability of this value be produced randomly. On the contrary, a social network property value is denominated "random" if there is a high probability of this value be produced randomly. Note that *class1* defines the strongest ties since all properties are social, whereas *class8* represents a completely random relationship. Moreover, *class2* and *class6* denote bridges, i.e., edges that persist over time, but have a small number of common neighbors. *class2* represents bridges with a high co-authorship frequency and *class6* with small one. Also, *class3* denotes a relationship that happens in a strong way (high neighborhood overlap and co-authorship frequency) but only in a specific moment. Thus, we call relationships in *class3* as *transient*. On the other hand, *class4* represents a *periodic* relationship since persists over time and has a high number of common neighbors, but small co-authorship frequency (for example, a co-authorship between colleagues from the same department that happens once a year). Moreover, *class5* defines a relationship with high co-authorship frequency, but does not persist and does not share many neighbors. This relationship tends to be isolated in the network. Finally, *class7* represents a weak tie, because it does not persist over time and the co-authorship frequency is small.

As RECAST, the unique parameter of STACY is p_{rnd} (better explained in Section 6.1.2), which determines when a social network property value is social or random. Indeed, Algorithm 2 presents how ties are classified in STACY. Note that STACY is also parallelized as fast-RECAST.

Algorithm 2 STACY: a parallelized code to classify weighted edges of G_t^W as eight different tie strength classes.

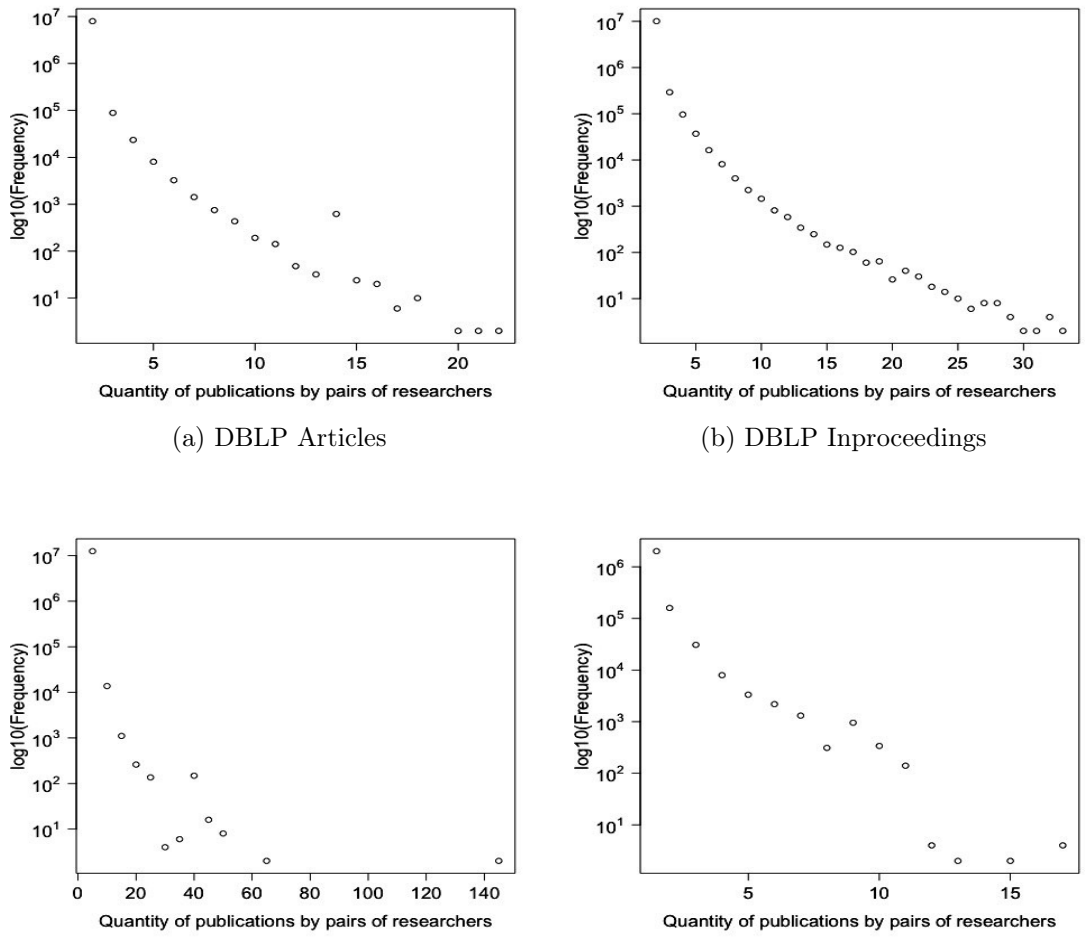
Input: Weighted temporal aggregated graph - G_t^W

Require: $p_{rnd} \geq 0$

1: **return** $class(i,j)\ \forall (i,j) \in U_t E_t$
2: Construct $G_t^{R,W}$ and set $\texttt{RND}(G_1^W), ..., \texttt{RND}(G_t^W)$ using **T-RND** with **pool.map_async**
3: Get $\overline{F}_{to}(x)$ and $\overline{F}_{per}(x)$ and $\overline{F}_{coAfrequency}(x)$ from $G_t^{R,W}$ using **pandas dataframe**
4: Get $\overline{x}_{to}|\overline{F}_{to}(\overline{x}_{to})$ and $\overline{x}_{per}|\overline{F}_{per}(\overline{x}_{per})$ and $\overline{x}_{coAfrequency}|\overline{F}_{coAfrequency}(\overline{x}_{coAfrequency}) = p_{rnd}$ with **pool.apply_async**
5: **for all** edges$(i,j) \in E_t$ **do**
6: **if** $per(i,j) > \overline{x}_{per}$ and $to(i,j) > \overline{x}_{to}$ and $coAfrequency(i,j) > \overline{x}_{coAfrequency}$ **then**
7: $class(i,j) \longleftarrow Class1$
8: **else if** $per(i,j) > \overline{x}_{per}$ and $to(i,j) \leq \overline{x}_{to}$ and $coAfrequency(i,j) > \overline{x}_{coAfrequency}$ **then**
9: $class(i,j) \longleftarrow Class2$
10: **else if** $per(i,j) \leq \overline{x}_{per}$ and $to(i,j) > \overline{x}_{to}$ and $coAfrequency(i,j) > \overline{x}_{coAfrequency}$ **then**
11: $class(i,j) \longleftarrow Class3$
12: **else if** $per(i,j) > \overline{x}_{per}$ and $to(i,j) > \overline{x}_{to}$ and $coAfrequency(i,j) \leq \overline{x}_{coAfrequency}$ **then**
13: $class(i,j) \longleftarrow Class4$
14: **else if** $per(i,j) \leq \overline{x}_{per}$ and $to(i,j) \leq \overline{x}_{to}$ and $coAfrequency(i,j) > \overline{x}_{coAfrequency}$ **then**
15: $class(i,j) \longleftarrow Class5$
16: **else if** $per(i,j) > \overline{x}_{per}$ and $to(i,j) \leq \overline{x}_{to}$ and $coAfrequency(i,j) \leq \overline{x}_{coAfrequency}$ **then**
17: $class(i,j) \longleftarrow Class6$
18: **else if** $per(i,j) \leq \overline{x}_{per}$ and $to(i,j) > \overline{x}_{to}$ and $coAfrequency(i,j) \leq \overline{x}_{coAfrequency}$ **then**
19: $class(i,j) \longleftarrow Class7$
20: **else**
21: $class(i,j) \longleftarrow Class8$

6.3 Experiments and Results

We now describe the experiments to analyze the dynamism of tie strength. We first present the datasets used to build the co-authorship SNs based on digital libraries from distinct areas of knowledge – Computer Science, Medicine and Physics (Section 6.3.1). Then, we apply fast-RECAST and STACY in the full temporal co-authorship SNs to characterize tie strength in these networks and compare their results (Section 6.3.3). Finally, we divide the SNs in two time windows to analyze the ties' dynamism



(a) DBLP Articles

(b) DBLP Inproceedings

(c) PubMed

(d) APS

Figure 6.3: Distribution of quantity of publications by pairs of researchers as counted yearly.

over the years using two different strategies: link persistence (Section 6.3.3.1) and link transformation (Section 6.3.3.2).

6.3.1 Data Description

We consider three publication datasets: DBLP, PubMed and APS, as collected in September 2015, April 2016 and March 2016, respectively. DBLP is a digital library that stores Computer Science publications. We get publications from conference inproceedings and journal articles, and divide them in two datasets: DBLP Inproceedings and DBLP Articles. Pubmed is a US national library of the Medicine National Institute of Health that comprises biomedical publications. We consider publications from the top-20 journals classified by h-index. APS (American Physical Society) is an or-

Table 6.2: Top 10 researchers with most publications and their respectively co-authors with most publications in *strong* class.

DBLP Articles		DBLP Inproceedings		PubMed		APS	
# Pub. R1	# Pub. R2	# Pub. R1	# Pub. R2	# Pub. R1	# Pub. R2	# Pub. R1	# Pub. R2
265	47	665	178	229	150	111	98
246	279	625	461	229	65	115	98
206	47	625	100	229	62	96	84
203	93	625	92	229	55	92	29
203	74	625	90	229	29	86	69
197	64	601	385	162	89	85	11
197	20	600	299	162	45	83	38
189	19	598	93	155	45	83	34
186	26	591	553	155	19	79	22
182	12	582	326	150	94	66	29

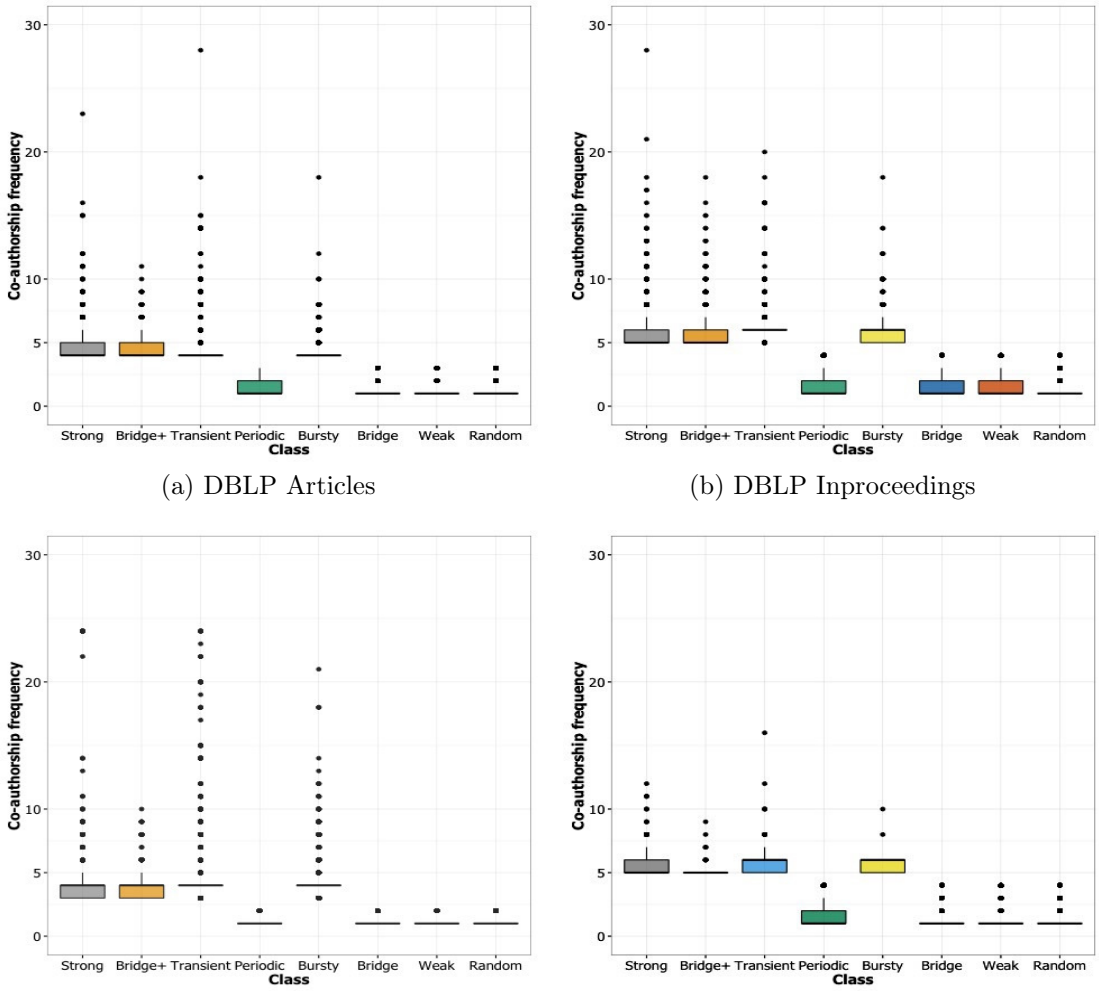
ganization for diffusing and advancing the knowledge of Physics. It provides a sample dataset with its journal publications.

Considering these datasets, we build four co-authorship SNs whose main statistics are in Table 3.1 of Chapter 3. Moreover, Figure 6.3 presents the distribution of pairs of researchers as counted yearly for each dataset. Note that the y-axis represents the frequency in log_{10}. For example, in Figure 6.3a, the number 5 in x-axis and the corresponding number 10^4 in y-axis indicate that the amount of 10^4 (in log_{10} scale) pairs of researchers have 5 publications in common considering all years in DBLP Articles dataset. Observe that the majority of co-authors have a small quantity of publications in a year, and PubMed has the largest number of co-authors in a single publication (a total of 140).

6.3.2 Characterizing STACY Classes

In this section, we characterize the eight classes of STACY according to the number of researchers' publications. Thus, Figure 6.4 presents (in box plots) the number of publications of pairs of researchers for each STACY class. In each box plot, the central rectangle spans the first to the third quartiles and also shows the outliers of the distribution. Note that *strong*, *bridge+*, *transient* and *bursty* classes have pairs of researchers with more number of publications. This is trivial, because these classes have in common the value "social" to co-authorship frequency. However, an interesting result is that the *strong* class (value "social" for the three features) has more ties with high number of publications than the others in the four datasets. The second class that has ties with more publications is *transient* (value "social" also to neighborhood overlap).

Furthermore, Figures 6.5 and 6.6 show the structure of DBLP Articles, DBLP Inproceedings, PubMed and APS co-authorship social networks in each STACY class. These visualizations allow to understand the networks' structure regarding their nodes and edges. For the networks with small number of nodes and edges, we have applied



(a) DBLP Articles (b) DBLP Inproceedings

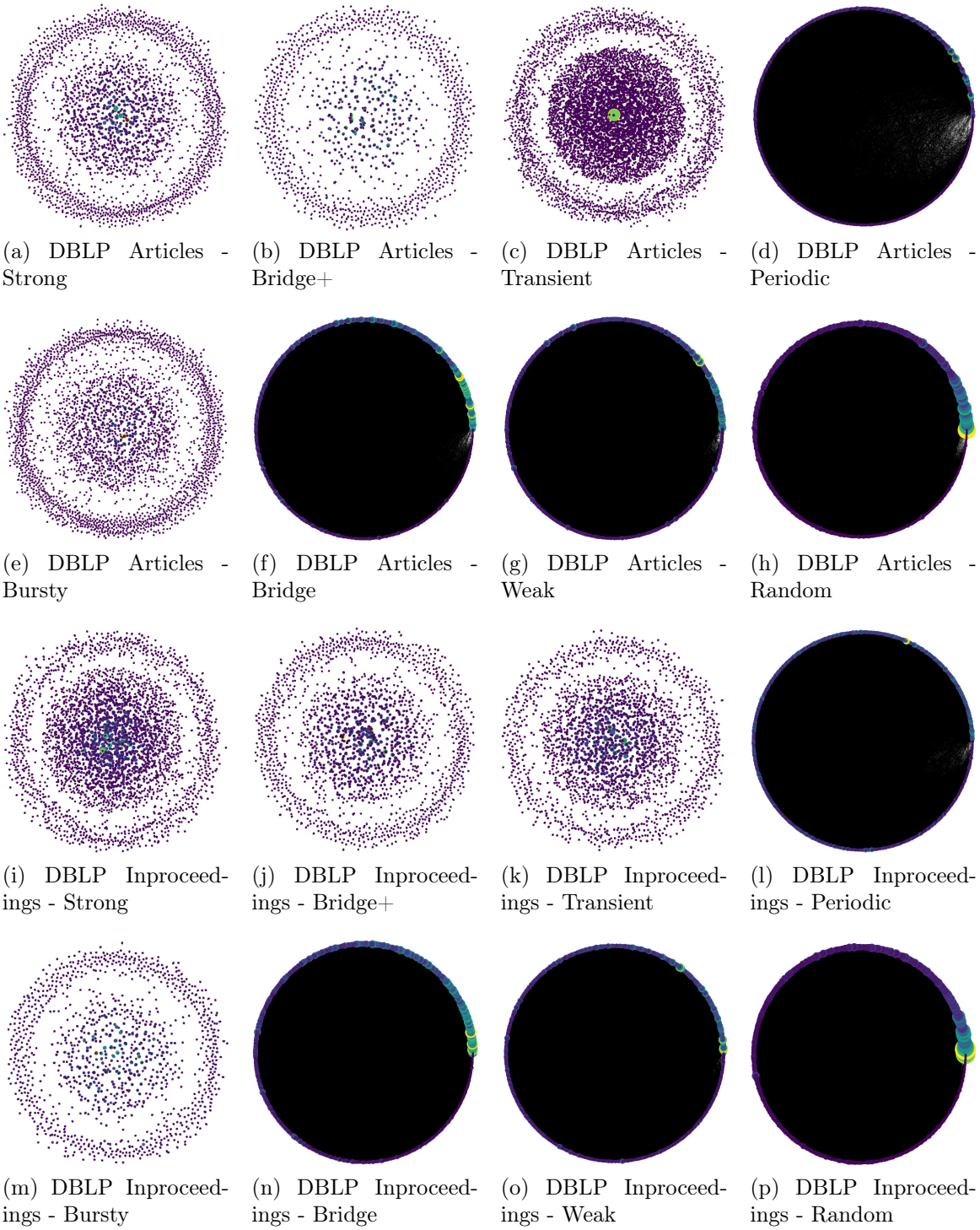
(c) PubMed (d) APS

Figure 6.4: Distribution of quantity of publications by pairs of researchers in each class detected by STACY.

Table 6.3: Top 10 researchers with most publications and their respectively co-authors with most publications in *bridge+* class.

DBLP Articles		**DBLP Inproceedings**		**PubMed**		**APS**	
# Pub. R1	**# Pub. R2**	**# Pub. R1**	**# Pub. R2**	**# Pub. R1**	**# Pub. R2**	**# Pub. R1**	**# Pub. R2**
726	136	1,137	116	229	94	288	35
726	101	1,137	100	229	70	288	8
726	85	1,137	80	229	43	115	34
726	78	1,137	48	229	36	95	17
726	26	1,137	40	229	34	85	28
440	124	1,005	334	229	23	77	11
440	70	1,005	265	229	20	74	21
415	70	1,005	8	229	18	68	28
415	47	928	266	229	16	65	21
415	22	928	196	229	13	65	8

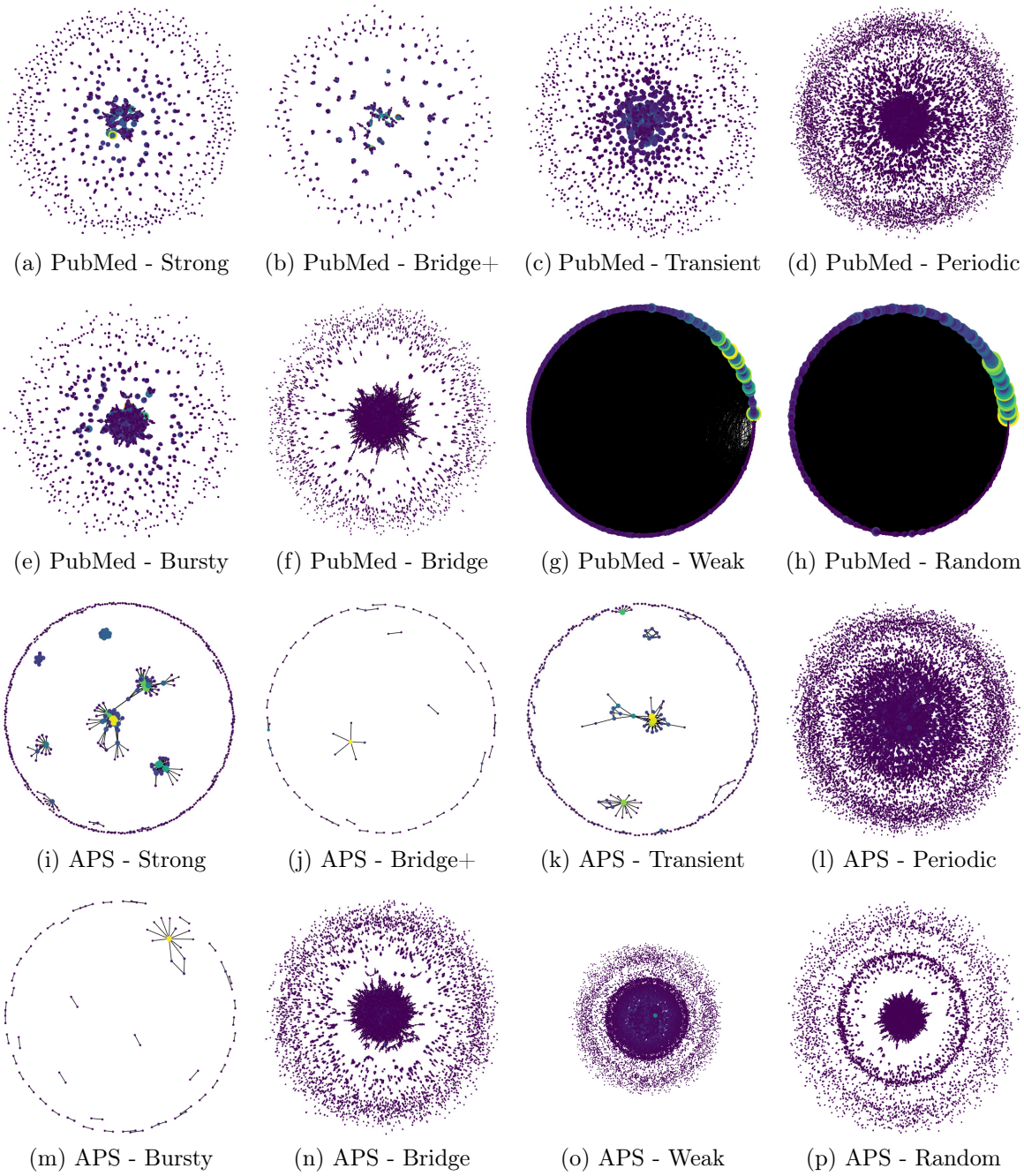
force directed layout to generate the visualizations [Guerra-Gomez et al., 2016; Holten and Van Wijk, 2009; Rahman and Karim, 2016]. However, for the largest ones, we



(a) DBLP Articles - Strong
(b) DBLP Articles - Bridge+
(c) DBLP Articles - Transient
(d) DBLP Articles - Periodic
(e) DBLP Articles - Bursty
(f) DBLP Articles - Bridge
(g) DBLP Articles - Weak
(h) DBLP Articles - Random
(i) DBLP Inproceedings - Strong
(j) DBLP Inproceedings - Bridge+
(k) DBLP Inproceedings - Transient
(l) DBLP Inproceedings - Periodic
(m) DBLP Inproceedings - Bursty
(n) DBLP Inproceedings - Bridge
(o) DBLP Inproceedings - Weak
(p) DBLP Inproceedings - Random

Figure 6.5: Social network for each relationship class from DBLP Articles and DBLP Inproceedings dataset. The size of the nodes varies according to the number of publications of the researchers.

have used *circular layout*, because the force directed layout has not generated a result



(a) PubMed - Strong (b) PubMed - Bridge+ (c) PubMed - Transient (d) PubMed - Periodic

(e) PubMed - Bursty (f) PubMed - Bridge (g) PubMed - Weak (h) PubMed - Random

(i) APS - Strong (j) APS - Bridge+ (k) APS - Transient (l) APS - Periodic

(m) APS - Bursty (n) APS - Bridge (o) APS - Weak (p) APS - Random

Figure 6.6: Social network for each relationship class from PubMed and APS dataset. The size of the nodes varies according to the number of publications of the researchers.

in a "real" time for such size of social networks.

Thus, Figures 6.5 and 6.6 reveal that the network composed by edges classified as *random* has researchers with a larger number of publications in DBLP Articles, DBLP Inproceedings and PubMed. This result shows that researchers with many publications may tend to form random ties. In general, such researchers are senior and

Table 6.4: Top 10 researchers with most publications and their respectively co-authors with most publications in *transient* class.

DBLP Articles		DBLP Inproceedings		PubMed		APS	
# Pub. R1	# Pub. R2	# Pub. R1	# Pub. R2	# Pub. R1	# Pub. R2	# Pub. R1	# Pub. R2
119	37	285	97	133	20	62	22
119	9	285	31	99	22	62	20
119	9	285	14	99	19	62	17
119	8	202	19	99	14	62	15
119	4	197	14	99	11	48	10
119	4	169	18	99	11	48	10
106	24	167	36	99	10	48	8
95	9	130	25	99	9	47	10
95	5	123	5	99	9	47	10
92	8	122	102	99	9	47	8

Table 6.5: Top 10 researchers with most publications and their respectively co-authors with most publications in *periodic* class.

DBLP Articles		DBLP Inproceedings		PubMed		APS	
# Pub. R1	# Pub. R2	# Pub. R1	# Pub. R2	# Pub. R1	# Pub. R2	# Pub. R1	# Pub. R2
400	69	928	243	229	162	369	47
398	18	928	239	229	145	288	67
398	11	928	184	229	105	209	39
398	10	802	276	229	89	175	40
398	8	802	234	229	76	175	35
398	6	802	154	229	68	175	31
318	183	802	77	229	57	175	23
292	137	700	136	229	27	148	40
292	50	700	115	162	84	148	20
276	9	700	77	162	70	144	32

Table 6.6: Top 10 researchers with most publications and their respectively co-authors with most publications in *bursty* class.

DBLP Articles		DBLP Inproceedings		PubMed		APS	
# Pub. R1	# Pub. R2	# Pub. R1	# Pub. R2	# Pub. R1	# Pub. R2	# Pub. R1	# Pub. R2
726	38	1,005	33	229	45	64	5
726	25	742	5	229	29	57	7
726	11	720	742	229	29	48	39
414	92	720	132	229	17	47	39
345	41	720	50	229	15	47	4
317	66	720	6	229	15	44	39
317	11	720	5	229	14	41	5
292	81	703	94	229	10	39	23
292	53	703	22	229	10	39	15
292	5	703	13	229	9	39	15

Table 6.7: Top 10 researchers with most publications and their respectively co-authors with most publications in *bridge* class.

DBLP Articles		DBLP Inproceedings		PubMed		APS	
# Pub. R1	# Pub. R2	# Pub. R1	# Pub. R2	# Pub. R1	# Pub. R2	# Pub. R1	# Pub. R2
726	268	1,137	450	229	90	369	61
726	241	1,137	436	229	86	369	59
726	227	1,137	432	229	84	369	54
726	212	1,137	314	229	83	369	43
726	206	1,137	296	229	82	369	39
726	184	1,137	260	229	81	369	37
726	155	1,137	253	229	70	369	35
726	145	1,137	251	229	70	369	33
726	143	1,137	244	229	65	369	30
726	139	1,137	240	229	65	369	26

Table 6.8: Top 10 researchers with most publications and their respectively co-authors with most publications in *weak* class.

DBLP Articles		DBLP Inproceedings		PubMed		APS	
# Pub. R1	# Pub. R2	# Pub. R1	# Pub. R2	# Pub. R1	# Pub. R2	# Pub. R1	# Pub. R2
182	6	492	31	162	18	67	1
182	4	492	17	139	119	67	1
182	3	492	11	108	16	67	1
182	3	492	7	108	14	67	1
182	1	492	6	108	12	67	1
182	1	492	6	108	11	66	6
182	1	492	5	108	11	66	6
172	10	492	4	108	10	65	5
169	2	492	4	108	9	65	4
169	2	492	4	108	6	65	2

Table 6.9: Top 10 researchers with most publications and their respectively co-authors with most publications in *random* class.

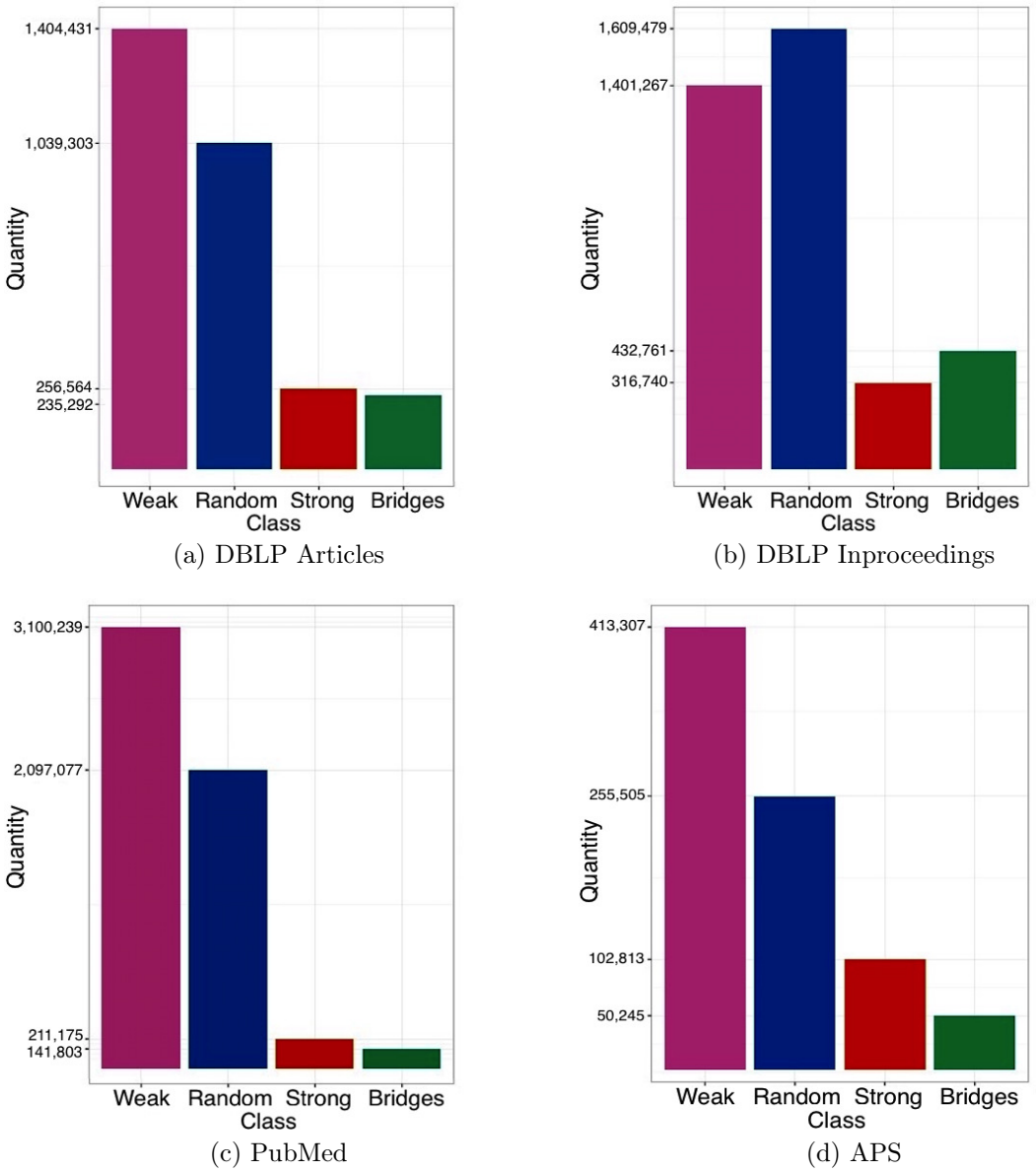
DBLP Articles		DBLP Inproceedings		PubMed		APS	
# Pub. R1	# Pub. R2	# Pub. R1	# Pub. R2	# Pub. R1	# Pub. R2	# Pub. R1	# Pub. R2
726	400	1,137	504	229	82	369	209
726	308	1,137	485	229	75	369	66
726	249	1,137	333	229	61	369	52
726	197	1,137	310	229	61	369	51
726	187	1,137	295	229	52	369	40
726	160	1,137	285	229	50	369	38
726	159	1,137	260	229	50	369	33
726	154	1,137	247	229	47	369	33
726	142	1,137	215	229	44	369	28
726	141	1,137	210	229	41	369	28

have random relationships with junior researchers. Also, the network with edges in the *weak* class also has researchers with many publications in the four social networks. However, such researchers have a number of publications smaller than the researchers in the network with *random* ties. Moreover, the network with edges in *transient* class is the one with researchers with less number of publications. The top 10 researchers with most publications (R1) in each class with their respectively co-authors with most publications (R2) in Tables 6.2 to 6.9 confirm these results.

6.3.3 Comparing fast-RECAST and STACY

As mentioned in Section 6.1.2, RECAST was originally used to classify users' wireless interaction in mobile networks [Vaz de Melo et al., 2015]. The patterns and features of such networks are different from co-authorship social networks. Hence, our goal is to verify whether such algorithm identifies the kind of the relationships (social or random) between co-authors. We also do the same verification for STACY.

Before executing fast-RECAST and STACY, we have to set a value to the parameter p_{rnd} (discussed in Section 6.1.2 and Section 6.2.3, respectively). Vaz de Melo et al. [2015] vary p_{rnd} through four orders of magnitude and observe that the number of edges per class keeps in the same magnitude. Therefore, such algorithm does not need



(a) DBLP Articles (b) DBLP Inproceedings

(c) PubMed (d) APS

Figure 6.7: Amount of pairs of authors in each class generated by fast-RECAST: weak, strong, bridge and random. Common behavior: the four co-authorship social networks have a large number of weak and random ties.

an accurate definition of the parameter to consistently classify the edges. Here, we run fast-RECAST and STACY for $p_{rnd} = 0.01$ and $p_{rnd} = 0$; as we obtain similar results, we show only those for $p_{rnd} = 0$. In summary, when $p_{rnd} = 0$, a given value v of edge persistence (or topological overlap or co-authorship frequency) is considered *social* (or *not random*) when there are no edges in the random graphs with edge persistence (or topological overlap or co-authorship frequency) greater than or equal to v.

Figure 6.7 presents the classification of the co-authorships in each class generated

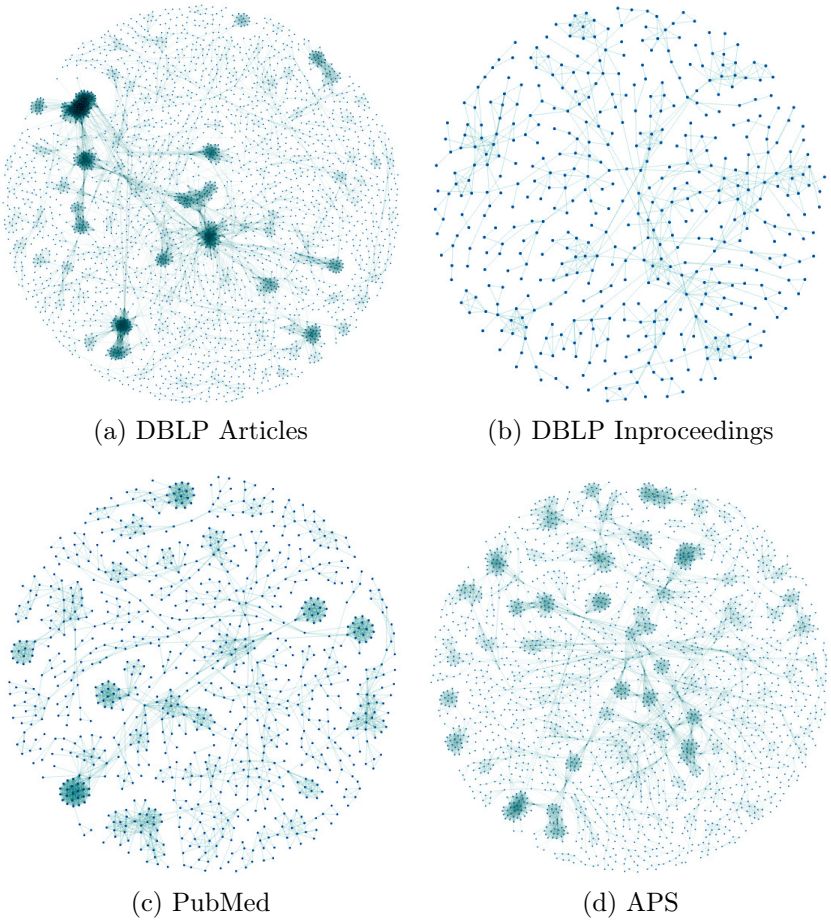
by fast-RECAST for the four co-authorship SNs. In DBLP Articles, PubMed and APS, most co-authorships are classified as weak ties, i.e., edges with small (or *random*) topological overlap and edge persistence. Also, in such networks, more co-authorships are classified as strong ties than as bridge ties. The exception is DBLP Inproceedings, in which most edges are attributed to the random class and more co-authorships are classified as bridges than as strong ties. These results can be explained by the fact that DBLP Articles, PubMed and APS have publications from journals, whereas DBLP Inproceedings has publications from conferences. As discussed in recent studies (such as [Montolio et al., 2013; Silva et al., 2014]), Computer Science has a very peculiar behavior when publishing in journals and conferences. Usually, conferences are for innovative ideas and journals for archival purposes. Hence, journal coauthor networks generally include authors who have already published together, then presenting stronger ties.

Such analysis is confirmed by Figure 6.8, which shows the structure of the co-authorship social networks considering only the edges classified as strong ties. There is a less dense connected component of strong ties for DBLP Inproceedings in comparison to the other networks. Also, PubMed has the largest connected components of strong ties in such way that it is hard o generate a suitable visualization.

Moreover, Figure 6.9 shows how STACY classifies the co-authorship ties in eight different classes for each social network. As fast-RECAST, in DBLP Articles, DBLP Inproceedings, PubMed and APS most ties are classified as *class7* (weak) and *class8* (random). Also, many ties are classified as *class4* (periodic) and *class6* (bridge). The high quantity of ties in *class4* reveals that researchers tend to publish together with small frequency in a year with colleagues from the same community (e.g., team, department, laboratory, etc). Also, the large amount of ties in *class6* indicates that most bridges tend to have a small co-authorship frequency in each time. Note that less ties are classified as *class1*, *class2*, *class3* and *class5*. These four classes have in common the value "social" to the social network property co-authorship frequency (the other four classes have a "random" value to this property). This shows that co-authorship frequency is an important feature to measure tie strength since helps to better differentiate the classes. These results are perceived in the four co-authorship social networks.

6.3.3.1 Link Persistence Analysis

Now, our goal is to investigate whether ties characterized with a given level of tie strength are likely to persist in the future. In a social context, persistence is interpreted as the continuation of a relationship even with the progress of time, geographic distance,



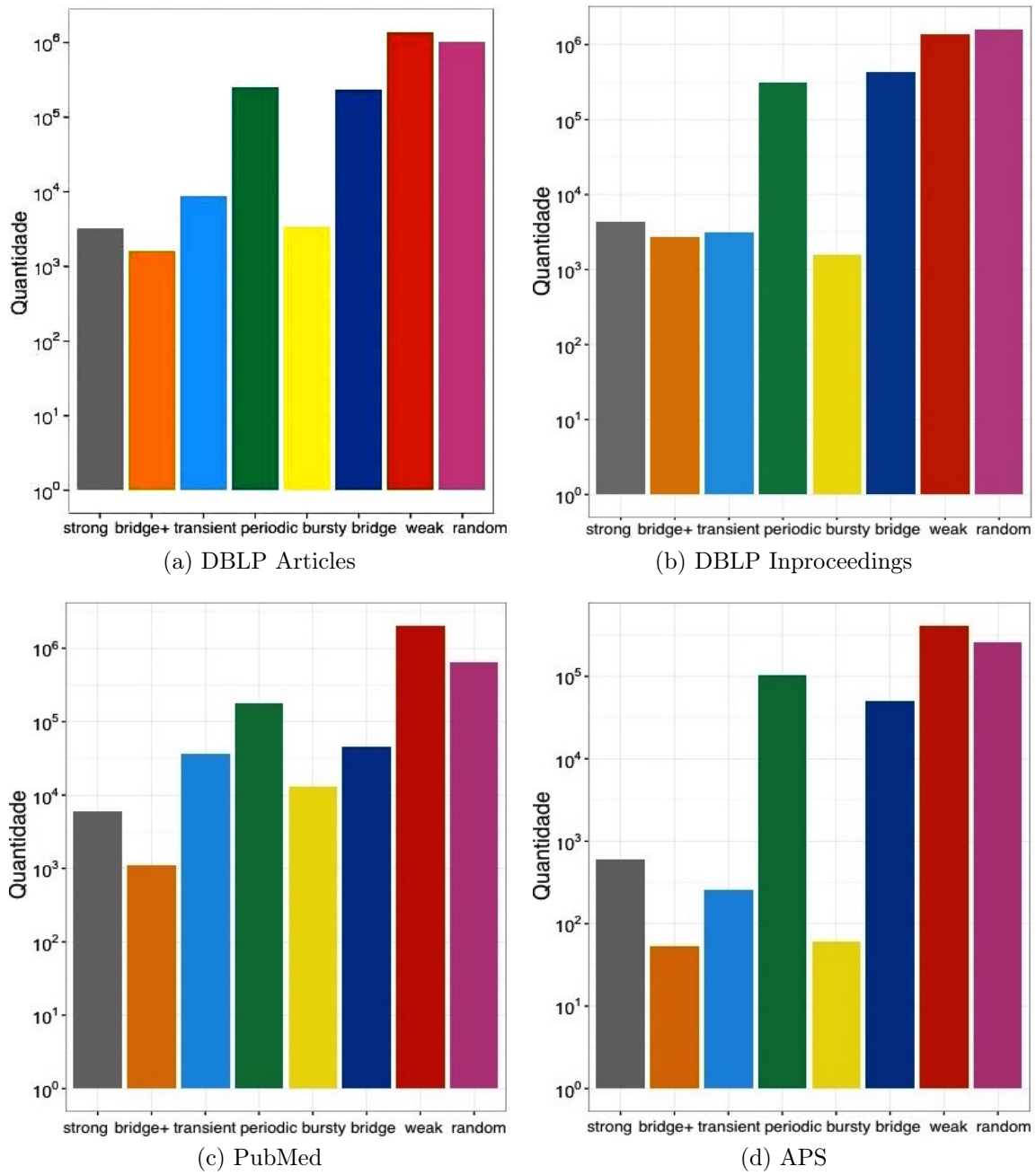
(a) DBLP Articles (b) DBLP Inproceedings

(c) PubMed (d) APS

Figure 6.8: SNs with N nodes and edges classified as strong ties (more visualizations available on `angra.lbd.dcc.ufmg.br/cnare/`): the largest connected components 6.8a - $N = 3,068$, 6.8b - $N = 473$, 6.8d - $N = 2,314$, and the ten greatest connected components (from the 2 to 11 component, inclusive) 6.8c - $N = 976$. We do not plot the largest connected component of PubMed because it is too large with $N = 22,000$.

or occupational mobility [Adams, 1967]. Here, we analyze co-authorship ties persistence over time. Furthermore, a relationship can be: symbiotic, which is based upon common need and is a relation of interdependence; or consensual, which is based upon common value and agreement [Adams, 1967; Gross, 1956]. Gross [1956] asserts that symbiotic ties persist more than consensual ties, but Adams [1967] claims that symbiotic ties with positive concerns (relationship based on obligation and need, when coupled with enduring relationship and continuing interest, evolves into a positive or affectional force) remain connected more than consensual ties. In this context, we consider that co-authorships are symbiotic relationships as they originated from a work involvement.

In order to analyze the persistence over time, we divide the networks into two



(a) DBLP Articles (b) DBLP Inproceedings

(c) PubMed (d) APS

Figure 6.9: Amount of pairs of authors in each class generated by STACY: class1 to class8. Common behavior: most ties are in classes that co-authorship frequency has "random" value.

time windows, which from now on we call *past* and *future*[6]. We apply *fast-RECAST* and STACY in the past and then, verify if the edges of each class (*strong*, *bridge*, *weak* and *random*) continue to be in that same class in the future.

In order to do that, we split the social networks into two time windows and in

[6]One may see the present as the timestamp between these two time windows

Table 6.10: fast-RECAST: 80% represents the past (2000-2012 to DBLP articles and inproceedings, 2000-2013 to PubMed and 2000-2010 to APS) and 20% is the present (2013-2015 to DBLP articles and inproceedings, 2014-2016 to PubMed and 2011-2013 to APS).

Edge type	DBLP Articles		DBLP Inproceedings		PubMed		APS	
	80%	**20%**	**80%**	**20%**	**80%**	**20%**	**80%**	**20%**
Strong	75,128	16,083 (0.21)	136,159	19,608 (0.14)	91,143	19,555 (0.21)	45,020	30,046 (0.67)
Bridge	133,071	28,090 (0.21)	368,177	55,327 (0.15)	50,903	11,239 (0.22)	50,464	31,767 (0.63)
Weak	767,143	28,683 (0.04)	750,837	16,244 (0.02)	1,790,986	67,752 (0.04)	201,978	102,108 (0.51)
Random	931,796	76,298 (0.08)	1,340,167	69,661 (0.05)	1,021,710	63,986 (0.06)	249,711	128,479 (0.51)

Table 6.11: fast-RECAST: 70% represents the past (2000-2010 to DBLP articles and inproceedings, 2000-2011 to PubMed and 2000-2009 to APS) and 30% is the present (2011-2015 to DBLP articles and inproceedings, 2012-2016 to PubMed and 2010-2013 to APS).

Edge type	DBLP Articles		DBLP Inproceedings		PubMed		APS	
	70%	**30%**	**70%**	**30%**	**70%**	**30%**	**70%**	**30%**
Strong	47,647	13,161 (0.28)	89,842	16,615 (0.18)	38,811	10,778 (0.28)	47,440	10,857 (0.23)
Bridge	92,991	25,592 (0.275)	276,785	54,119 (0.2)	21,031	6,400 (0.30)	31,267	6,672 (0.21)
Weak	539,062	27,983 (0.05)	522,980	16,291 (0.03)	1,171,785	71,140 (0.06)	221,660	15,212 (0.07)
Random	681,519	76,260 (0.11)	1,021,110	70,700 (0.07)	584,353	60,105 (0.1)	168,872	16,147 (0.1)

two ways. First, we split the networks into a time window comprising 80% of the initial timestamp (*past*) and a time window comprising 20% of the final timestamp (*future*). Second, we divide the networks into time windows of 70% (*past*) and 30% (*future*). Tables 6.10 and 6.12 present the results for 80% and 20% partition for fast-RECAST and STACY, respectively. The values in the 80% column are the absolute number of edges from the 80% of the publications' years attributed to each class. The values in the 20% column are the number of edges from the past that are also in the future (proportions between parentheses). We observe that strong ties and bridges tend to persist over the years more than weak and random ties. Considering the 70%-30% split, as shown in Tables 6.11 and 6.13 for fast-RECAST and STACY, respectively, the same conclusions can be made.

Considering fast-RECAST results, we emphasize the differences in the results of the APS network in the 80%-20% and 70%-30% partitions. In the first partitioning, the proportion of strong and bridge ties from the past to the present is very high, whereas in the second partitioning such proportion is lower. This result may indicate

Table 6.12: STACY: 80% represents the past (2000-2012 to DBLP articles and inproceedings, 2000-2013 to PubMed and 2000-2010 to APS) and 20% is the present (2013-2015 to DBLP articles and inproceedings, 2014-2016 to PubMed and 2011-2013 to APS).

Edge type	**DBLP Articles**		**DBLP Inproceedings**		**PubMed**		**APS**	
	80%	**20%**	**80%**	**20%**	**80%**	**20%**	**80%**	**20%**
Class1	1,238	485 (0.39)	2,562	674 (0.26)	6,003	2,230 (0.37)	93	17 (0.18)
Class2	886	368 (0.41)	2,498	573 (0.23)	1,113	305 (0.27)	8	2 (0.25)
Class3	0	0	0	0	37,157	2,771 (0.07)	120	95 (0.79)
Class4	1,070,400	64,249 (0.06)	1,149,339	53,445 (0.05)	175,179	34,215 (0.2)	58,663	12,122 (0.21)
Class5	0	0	0	0	12,862	1,372 (0.1)	4	3 (0.75)
Class6	834,614	84,052 (0.1)	1,440,941	106,148 (0.07)	45,419	8,718 (0.19)	36,720	6,840 (0.19)
Class7	0	0	0	0	2,042,114	76,552 (0.04)	256,564	13,908 (0.05)
Class8	0	0	0	0	634,895	36,369 (0.05)	195,001	15,573 (0.08)

Table 6.13: STACY: 70% represents the past (2000-2010 to DBLP articles and inproceedings, 2000-2011 to PubMed and 2000-2009 to APS) and 30% is the present (2011-2015 to DBLP articles and inproceedings, 2012-2016 to PubMed and 2010-2013 to APS).

Edge type	**DBLP Articles**		**DBLP Inproceedings**		**PubMed**		**APS**	
	70%	**30%**	**70%**	**30%**	**70%**	**30%**	**70%**	**30%**
Class1	823	415 (0.5)	1,893	639 (0.34)	2,450	1,226 (0.5)	87	13 (0.15)
Class2	676	312 (0.46)	2,082	591 (0.28)	267	101 (0.38)	9	2 (0.22)
Class3	0	0	0	0	26,947	3,293 (0.12)	4	3 (0.75)
Class4	745,375	58,976 (0.08)	823,799	51,232 (0.06)	85,011	23,799 (0.28)	47,353	10,844 (0.23)
Class5	0	0	0	0	6,624	1,107 (0.17)	13	12 (0.92)
Class6	614,345	83,293 (0.14)	1,082,943	105,263 (0.1)	21,095	5,849 (0.28)	31,258	6,670 (0.21)
Class7	0	0	0	0	1,314,738	77,707 (0.06)	221,656	15,209 (0.07)
Class8	0	0	0	0	358,848	35,341 (0.1)	168,859	16,135 (0.1)

that the co-authorship social network from APS changes more through the years than the other networks. Another possibility is that physics researchers do not change very much the level of co-authorship with their collaborators over time, and this is a pattern of more recent researchers (note that 80% of data consider more recent co-authorships than 70%). We leave for future work further analyses of such claims.

Table 6.14: fast-RECAST: Link transformation results for DBLP Articles.

	Strong	**Bridge**	**Weak**	**Random**	**Disappear**
Strong	43,711 (0.11)	27,134 (0.07)	0	0	312,765 (0.82)
Bridge	14,650 (0.04)	13,874 (0.035)	0	0	361,041 (0.925)
Weak	0	0	0	0	0
Random	0	0	0	0	0

Table 6.15: fast-RECAST: Link transformation results for DBLP Inproceedings.

	Strong	**Bridge**	**Weak**	**Random**	**Disappear**
Strong	34,761 (0.08)	26,411 (0.06)	0	0	351,935 (0.86)
Bridge	13,601 (0.02)	16,298 (0.024)	0	0	659,608 (0.96)
Weak	0	0	0	0	0
Random	0	0	0	0	0

Now, focusing on STACY results, we observe that strong ties tend to persist more than the others in DBLP Articles, DBLP Inproceedings and PubMed in the 80%-20% and 70%-30% partitions. Also, note that STACY is able of better classifying strong ties that persist over time than fast-RECAST. An increase of 0.18 for DBLP Articles, 0.12 for DBLP Inproceedings and for 0.16 PubMed in the 80%-20% partition. For 70%-30% partition, growth is even better 0.22 for DBLP Articles, 0.16 for DBLP Inproceedings and for 0.22 PubMed. The exception is APS, in which most ties in *class3* (transient) and *class5* (bursty) tend to persist over time. This is a unexpected result since both classes have "random" value for edge persistence. Analyzing the main cause for this result, we note that co-authorships in such classes occur from 2009 to 2013, i.e., in the last years of the partitions (in 80%-20%, the 80% includes 2009 and 2010 and in 70%-30%, the 70% includes 2009). Thus, the edge persistence value is small, because the co-authorships occur in the years of the 30% (future). Additionally, no ties are classified as *class3*, *class5*, *class7* and *class8* in DBLP Articles and DBLP Inproceedings in both partitions. This reveals that in such networks transient, bursty, weak and random co-authorships are recent relationships, because they are found in the full version of these SNs (as shown by Figure 6.9). Also, weak and random ties are the ones that less persist over time in PubMed and APS.

This study reveals that most relations of co-authorship are symbiotic without positive concerns, because most of them perish over time. Just a few of them are symbiotic with positive concerns. This pattern is observed in the four co-authorship social networks for fast-RECAST and STACY, and specially considering the temporal data division as 70% and 30% in fast-RECAST.

6.3.3.2 Link Transformation Analysis

We now evaluate the amount of ties from a class in the past that continues in the same class (or changes) in the future. To avoid any kind of bias in the process of classifying

Table 6.16: fast-RECAST: Link transformation results for PubMed.

	Strong	**Bridge**	**Weak**	**Random**	**Disappear**
Strong	349 (0.02)	387 (0.02)	3,267 (0.16)	2,664 (0.13)	17,044 (0.67)
Bridge	66 (0.01)	97 (0.01)	659 (0.07)	667 (0.07)	8,643 (0.84)
Weak	10,532 (0.02)	10,425 (0.02)	94,800 (0.18)	73,039 (0.13)	346,559 (0.65)
Random	1,476 (0.01)	1,792 (0.01)	13,105 (0.06)	11,941 (0.05)	195,803 (0.87)

Table 6.17: fast-RECAST: Link transformation results for APS.

	Strong	**Bridge**	**Weak**	**Random**	**Disappear**
Strong	836 (0.03)	571 (0.02)	2,219 (0.09)	1,691 (0.06)	19,625 (0.8)
Bridge	450 (0.02)	421 (0.02)	918 (0.04)	910 (0.04)	19,173 (0.88)
Weak	4,013 (0.03)	2,071 (0.02)	14,185 (0.11)	7,154 (0.06)	99,844 (0.78)
Random	1,561 (0.013)	1,158 (0.01)	4,072 (0.03)	3,625 (0.03)	107,452 (0.92)

Table 6.18: STACY: Link transformation results for DBLP Articles.

	Class1	**Class2**	**Class3**	**Class4**	**Class5**	**Class6**	**Class7**	**Class8**	**Disappear**
Class1	0	1 (0.002)	0	54 (0.09)	0	19 (0.03)	0	0	549 (0.88)
Class2	0	0	0	8 (0.03)	0	9 (0.03)	0	0	238 (0.93)
Class3	0	0	0	0	0	0	0	0	0
Class4	58 (1e-04)	7 (1.39e-05)	0	59,823 (0.12)	0	19,568 (0.04)	0	0	423,247 (0.84)
Class5	0	0	0	0	0	0	0	0	0
Class6	24 (8.9e-05)	4 (1.5e-05)	0	13,465 (0.05)	0	6,329 (0.02)	0	0	249,772 (0.92)
Class7	0	0	0	0	0	0	0	0	0
Class8	0	0	0	0	0	0	0	0	0

Table 6.19: STACY: Link transformation results for DBLP Inproceedings.

	Class1	**Class2**	**Class3**	**Class4**	**Class5**	**Class6**	**Class7**	**Class8**	**Disappear**
Class1	0	0	0	28 (0.06)	0	21 (0.05)	0	0	387 (0.88)
Class2	0	0	0	21 (0.03)	0	7 (0.01)	0	0	596 (0.96)
Class3	0	0	0	0	0	0	0	0	0
Class4	28 (6.79e-05)	5 (1.2e-05)	0	44,665 (0.1)	0	16,425 (0.04)	0	0	351,548 (0.85)
Class5	0	0	0	0	0	0	0	0	0
Class6	26 (3.8e-05)	6 (8.7e-06)	0	19,148 (0.03)	0	10,691 (0.02)	0	0	659,012 (0.95)
Class7	0	0	0	0	0	0	0	0	0
Class8	0	0	0	0	0	0	0	0	0

the edges, here we divide the temporal co-authorship social networks into two time windows of 50% of the timestamp. We apply fast-RECAST and STACY in both parts and then we analyze the link transformation through the classes. Tables 6.14 to 6.17 show the results for fast-RECAST and Tables 6.18 to 6.21 for STACY. The values in each column represent the amount and the proportion (between parentheses) of ties from the past that persists or changes class in the future. For instance, the first values 43,711 and 0.11 in Table 6.14 are the number and the proportion, respectively, of *strong* links in the past that are still *strong* in the present.

Analyzing fast-RECAST results, surprisingly, we cannot see ties classified as *weak* and *random* in DBLP Articles and DBLP Inproceedings in Tables 6.14 and 6.15. This indicates that the features (edge persistence and topological overlap) of these social networks have high (or *social*) values. Furthermore, most ties from the past tend to disappear in the present, especially the *bridges*. This result may be explained by the nature of co-authorships, as researchers collaborate during a period towards a common

Table 6.20: STACY: Link transformation results for PubMed.

	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8	Disappear
Class1	0	0	0	91 (0.14)	0	74 (0.12)	0	0	478 (0.74)
Class2	0	0	0	4 (0.05)	0	3 (0.03)	0	0	75 (0.91)
Class3	0	0	0	344 (0.19)	0	106 (0.06)	0	0	1348 (0.74)
Class4	0	1 (4.1e-05)	0	4,780 (0.2)	0	2,440 (0.1)	0	0	17,192 (0.7)
Class5	0	0	0	27 (0.05)	0	18 (0.03)	0	0	494 (0.9)
Class6	0	0	0	473 (0.09)	0	290 (0.05)	0	0	4,675 (0.86)
Class7	35 (5.7e-05)	7 (1.1e-05)	0	137,563 (0.22)	0	62,939 (0.1)	0	0	416,963 (0.67)
Class8	1 (7.2e-06)	0	0	10,216 (0.07)	0	5,854 (0.04)	0	0	123,557 (0.88)

Table 6.21: STACY: Link transformation results for APS.

	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8	Disappear
Class1	0	0	0	0	0	2 (0.3)	0	0	5 (0.7)
Class2	0	0	0	0	0	0	0	0	3 (1.0)
Class3	0	0	0	0	0	0	0	0	0
Class4	0	0	0	836 (0.03)	0	569 (0.02)	2,219 (0.09)	1,691 (0.07)	19,620 (0.8)
Class5	0	0	0	0	0	0	1 (1.0)	0	0
Class6	0	0	0	450 (0.02)	0	421 (0.02)	918 (0.04)	910 (0.04)	19,170 (0.9)
Class7	11 (1e-04)	2 (1e-05)	0	4,002 (0.03)	0	2,069 (0.02)	14,185 (0.11)	7,154 (0.06)	99,844 (0.8)
Class8	4 (3e-05)	2 (1e-05)	0	1,557 (0.01)	1 (8e-06)	1,156 (0.01)	4,071 (0.03)	3,624 (0.03)	107,452 (0.9)

goal and then, start to collaborate with others. This also reinforces the theory of Granovetter that weak ties are the ones that connect different communities [Granovetter, 1973], which is the case of the bridge edges.

For Tables 6.16 and 6.17, we observe similar behavior between PubMed and APS, and most ties tend to disappear, especially the bridges and random ties. Disregarding disappeared links, most strong and weak ties become weak or random. Surprisingly, the weak ties are the ones that keep more in the same class, comparing to the others in both networks.

Focusing on STACY results, we also cannot see ties classified as weak (*class7*) and random (*class8*) in DBLP Articles and DBLP Inproceedings in Tables 6.18 and 6.19. Thus, co-authorship frequency of these co-authorship SNs also has large (or *social*) value. Furthermore, ties are not classified as transient (*class3*) in DBLP Articles, DBLP Inproceedings and APS (Table 6.21), which reveals the absence of these co-authorships in earlier periods in these networks. Also, DBLP Articles and DBLP Inproceedings do not have ties classified as bursty (*class5*), which indicates that ties with high co-authorship frequency also share a large number of neighbors in this networks in the period covered by the 50% of data (this is also confirmed by the presence of ties in *class3*).

Like fast-RECAST, most ties also tend to disappear when classified by STACY. The difference is that using STACY, we note that ties from different classes tend to change to *class4* (periodic) and *class6* (bridge) over time, specially, in DBLP Articles, DBLP Inproceedings and PubMed (Table 6.20).

Table 6.22: Range of values per class in DBLP Articles.

Class	Range of values
Class1	[0.27; 0.8]
Class2	[0.04; 0.12]
Class3	[0.15; 0.52]
Class4	[0.06; 0.2]
Class5	[0.005; 0.05]
Class6	[0.008; 0.03]
Class7	[0.04; 0.13]
Class8	[0.003; 0.05]

Table 6.23: Range of values per class in DBLP Inproceedings.

Class	Range of values
Class1	[0.32; 0.9]
Class2	[0.06; 0.2]
Class3	[0.19; 0.6]
Class4	[0.06; 0.19]
Class5	[0.02; 0.08]
Class6	[0.008; 0.03]
Class7	[0.03; 0.13]
Class8	[0.003; 0.01]

Table 6.24: Range of values per class in PubMed.

Class	Range of values
Class1	[0.26; 0.67]
Class2	[0.08; 0.15]
Class3	[0.16; 0.5]
Class4	[0.08; 0.19]
Class5	[0.04; 0.07]
Class6	[0.02; 0.04]
Class7	[0.04; 0.15]
Class8	[0.009; 0.02]

Table 6.25: Range of values per class in APS.

Class	Range of values
Class1	[0.66; 1.5]
Class2	[0.11; 0.26]
Class3	[0.29; 0.63]
Class4	[0.08; 0.25]
Class5	[0.015; 0.09]
Class6	[0.015; 0.04]
Class7	[0.04; 0.14]
Class8	[0.006; 0.018]

6.3.4 Deriving temporal_tieness from STACY

As described in Section 6.2.3, STACY classifies ties in eight different classes by combining neighborhood overlap (or topological overlap, $to_{(i,j)}$ – note that we do not consider a modification in neighborhood overlap as in Chapter 5, because the values of neighborhood overlap and a modification of it slightly differentiate from each other over time for pairs of researchers in the four co-authorship SNs), edge persistence ($per_{(i,j)}$) and co-authorship frequency ($coAfrequency_{(i,j)}$). From this combination, we derive a computational model formally defined by Equation 6.3.

$$temporal\_tieness_{(i,j)} = per_{(i,j)}^{\alpha_1} \times to_{(i,j)}^{\alpha_2} \times coAfrequency_{(i,j)}^{\alpha_3} \quad (6.3)$$

, in which α_k (k is 1, 2 or 3) determines the weight that is given to each feature.

Considering $\alpha_1 = 1$, $\alpha_2 = 1$ and $\alpha_3 = 1$ by default, Tables 6.22 to 6.25 present the range of values for temporal_tieness in each class. The temporal_tieness metric is calculated for each pair of researchers by using the values of the metrics (edge persistence, neighborhood overlap and co-authorship frequency) computed by STACY when classifying the ties. To avoid extreme values [Brandão et al., 2014], we get the

first and third quartiles of temporal_tieness in each class to define the beginning and end of the range of values. Note that *class1* (strong ties) has the largest values and *class8* has the smallest ones in the four co-authorship social networks. Also, *class3* has the second largest range of values in all networks. Unfortunately, there are still some overlaps between range of values between some classes, but it can be solved by better analyzing the values of the α parameter that we leave for future work.

These results indicate that temporal_tieness has a pattern of values for each class in co-authorship social networks that have collaboration as an inherent characteristic. Although it is necessary to better define the range of values for some classes, temporal_tieness is able of directly identifying *strong*, *weak* and *random* ties by using such ranges of values. Therefore, this new computational model can be used to measure tie strength in co-authorship social network without the use of STACY, which has more computational cost.

6.4 Concluding Remarks

The concept of tie strength is well understood and analyzed for static networks, but little is known about this concept when applied to temporal networks. In this chapter we characterized the strength of ties in temporal networks by measuring the persistence and the transformation of ties over time. In order to do so, we built four temporal co-authorship social networks considering three real publications datasets. We also proposed fast-RECAST, a parallel and faster version of an existing algorithm (RECAST) that classifies edges into four classes of relationship according to their level of tie strength. Moreover, we propose STACY, a parallel and fast algorithm that classifies the ties into eight different classes. Moreover, we characterize each class according to the number of publications of the researchers. Also, by grouping the edges into these classes, we were able to quantify the dynamism of tie strength over time.

Regarding the results, the link persistence analysis reveals that strong ties and bridges tend to persist over the years more than weak and random ties. Overall, this supports our initial hypothesis that strong ties persist more than the others. Furthermore, STACY was able of finding strong ties that persist more than those found by fast-RECAST. The results of fast-RECAST also show a different pattern for co-authorship social network from APS when the data is divided in 80% and 20%. In this experimental setting, the proportion of strong and bridge ties from the past to the present is very high compared to other social networks. Moreover, the link transformation analysis by using fast-RECAST and STACY revealed that most ties tend to

disappear over time. This may occur due to the co-authorships nature, e.g., researchers tend to publish with students during a period and when the students graduate, they finalize the process of publishing together.

Finally, by using STACY, we defined a new computational model called temporal_tieness and a range of values for each class. Thus, tie strength can be computed with low computational cost when compared to fast-RECAST and STACY.

Chapter 7

Conclusions and Future Work

In this chapter, we summarize the main results achieved so far (Section 7.1) and present the open problems and future work derived from this thesis (Section 7.3).

7.1 Conclusions

In this thesis, we have studied distinct aspects related to the strength of co-authorship ties. Specially, we did analyses, formulated metrics and developed a new computational model. Such studies are categorized in research questions, which are summarized in Sections 7.1.1 to 7.1.5.

7.1.1 RQ1: How to identify which aspects impact on the strength of collaboration ties?

We have studied how nine topological properties (edge betweenness, co-authorship frequency, closeness, eccentricity, clustering coefficient, number of triangles, weight degree, eigenvector and page rank) impact on neighborhood overlap in non-temporal co-authorship social networks from three different research areas (computer science, medicine and sociology). The results showed that each research area has important aspects that impact on the strength of co-authorship ties, since most properties are related to neighborhood overlap in different ways depending on the research area. The results also reveal that edge betweenness, closeness, eccentricity, clustering coefficient, number of triangles and eigenvector are linearly or exponentially dependent of neighborhood overlap in at least one research area. Therefore, such metrics should not be combined with neighborhood overlap to measure the strength of ties. Furthermore, we note that the co-authorship frequency is linearly and exponentially independent of

neighborhood overlap in all social networks. Thus, both metrics can be combined in a computational model to measure tie strength.

7.1.2 RQ2: How to measure the strength of co-authorship ties in non-temporal social networks?

We have measured tie strength in non-temporal and temporal co-authorship social networks. In non-temporal social networks, neighborhood overlap and absolute frequency of interaction (a.k.a. co-authorship frequency or edge weight) have been largely used to measure the strength of ties. Indeed, we have initially measured the strength of ties by using neighborhood overlap (NO) and contrasting it with co-authorship frequency. Such comparison allowed to define a nominal scale to NO. The results showed that different properties influence such metric in a linear way or not. In addition, since we are measuring the strength of ties, we verified if Granovetter's theory governs co-authorship social network when such strength is measured by neighborhood overlap. Our results were positive to such theory. Therefore, our evaluations indicate that neighborhood overlap can be used to measure the strength of ties.

However, by empirically analyzing the results, we identified four main problems with using solely neighborhood overlap and co-authorship frequency to measure tie strength: *(Case 1)* when a pair of collaborators does not have any common neighbor, neighborhood overlap will be zero; *(Case 2)* how determining if two collaborators are from the same community or not is challenging, since co-authorship frequency considers only the absolute frequency of interaction; *(Case 3)* when there is little collaboration between a pair of collaborators and a plenty of common neighbors, neighborhood overlap and co-authorship frequency will present opposite results; and *(Case 4)* when the results are extreme values, neighborhood overlap may not represent the reality. Hence, we proposed a new metric entitled tieness that combines a modified neighborhood overlap with co-authorship frequency. We also defined a nominal scale and verified Granovetter's theory when the strength of ties is measured by tieness. Our analysis validates our metric according Granovetter's theory and shows promising results.

7.1.3 RQ3: How to measure the strength of co-authorship ties in temporal social networks?

In temporal social networks, we have used two algorithms to measure tie strength. They were fast-RECAST and STACY. Both algorithms classify the ties by comparing the values of social networks features with values from random networks. fast-RECAST re-

sulted from an improvement in an existing algorithm called RECAST, whereas STACY was proposed by us. Furthermore, fast-RECAST identifies four relationships classes (strong, weak, bridge and random), while STACY classifies the ties in eight different classes (strong, bridge+, bridge, transient, periodic, bursty, weak and random). Also, the two algorithms differ by the number of considered features: fast-RECAST uses two social network features (edge persistence and neighborhood overlap) and STACY considers three features (edge persistence, neighborhood overlap and co-authorship frequency). As STACY recognizes more tie strength classes, it allows to identify more types of relationships. For example, co-authorships that are bridge+, periodic, transient or bursty. Also, our new algorithm possibilities to observe that most bridges in co-authorship SNs tend to have small co-authorship frequency and that researchers tend to publish together with small frequency in a year with colleagues from the same community. These results follow our intuition of research collaboration. Thus, our new algorithm is able of automatically find diverse kinds of co-authorships. Finally, from STACY, we are able to derive a computational model called *temporal_ tieness* that can classify tie strength with low computational cost.

7.1.4 RQ4: How is tie strength defined for temporal networks?

There are different concepts related to tie strength in non-temporal social networks. However, few studies have addressed the strength of ties in temporal networks. In this thesis, we considered that a strong tie characterizes interactions that are likely to appear in the future, whereas a weak tie occurs sporadically. Our results confirm such claim, since strong ties persist more than weak ones.

7.1.5 RQ5: How much does the strength of ties vary over time?

We have investigated tie strength dynamism over time by analyzing tie persistence and transformation in different classes. To do so, we have applied an existing algorithm called RECAST, whose performance we have improved and called as fast-RECAST. Such algorithm classifies the ties in four different classes (strong, weak, bridge and random). We have also used a new algorithm called STACY that classifies the ties in eight different classes. Surprisingly, most ties tend to perish over time. Moreover, the link persistence analysis reveals that strong ties and bridges tend to persist over the years more than weak and random ties. Also, STACY reveals that more persistent bridges have "social" value to co-authorship frequency. Furthermore, our new algorithm

was able of finding strong ties that persist more than those found by fast-RECAST. All these results show that STACY is able of automatically finding different kinds of relationships in temporal co-authorship social networks.

7.2 Publications

We have the following publications (all of them were published during the PhD) directly and indirectly related to this thesis:

1. BRANDÃO, M. A.; Diniz, M. A. ; de Sousa, G. A. ; Moro, M. M. . Visualizing Co-Authorship Social Networks and Collaboration Recommendations With CNARe. In: Natarajan Meghanathan. (Org.). Advances in Wireless Technologies and Telecommunication. 1ed.: IGI Global, 2017, p. 173-188.

2. BRANDÃO, M. A.; Moro, M. M. The Strength of Co-authorship Ties through Different Topological Properties. Journal of the Brazilian Computer Society (JBCS), 23(1):5, 2017;

3. BRANDÃO, M. A. ; Moro, M. M. Social Professional Network: a Survey and Taxonomy. Journal of Computer Communications (COMCOM), v. 100, p. 20, 2017;

4. BRANDÃO,M. A.; Diniz, M. A.; Moro, M. M. Using Topological Properties to Measure the Strength of Co-authorship Ties. In Proceedings of the V Brazilian Workshop on Social Network Analysis and Mining (BRASNAM–CSBC), 2016;

5. Alves, G.B., BRANDÃO, M.A., Santana, D.M., da Silva, A.P.C. and Moro, M.M., The Strength of Social Coding Collaboration on GitHub. In Proceedings of the 31th Symposium on Databases (SBBD), 2016;

6. BRANDÃO, M. A. ; Moro, M. M. Analyzing the Strength of Co-authorship Ties with Neighborhood Overlap. In Proceedings of the 26th International Conference on Database and Expert Systems Applications (DEXA), 2015;

7. BRANDÃO, M. A. ; Moro, M. M. Neighborhood Overlap: Can This Metric Be Used to Characterize the Strength of Co-authorship Ties?. In ACM Student Research competition & Grace Hopper Celebration, 2015;

8. Sousa, G. A. ; Diniz, M. A. ; BRANDÃO, M. A. ; Moro, M. M. CNARe: Co-authorship Networks Analysis and Recommendations. In Proceedings of the 9th ACM Conference on Recommender Systems (RecSys), 2015;

9. Diniz, M. A. ; Sousa, G. A. ; BRANDÃO, M. A. ; Moro, M. M. CNARe: Uma Ferramenta Online para Análise de Redes Sociais de Co-autoria e Recomendações. In Proceedings of the 30th Simpósio Brasileiro de Bancos de Dados (SBBD), 2015;

10. BRANDÃO, M. A. ; Moro, M. M. ; Almeida, J. M. Experimental Evaluation of Academic Collaboration Recommendation Using Factorial Design. Journal of Information and Data Management (JIDM), v. 5, p. 52, 2014;

11. BRANDÃO, M. A. ; Moro, M. M. ; Almeida, J. M. Análise de Fatores Impactantes na Recomendação de Colaborações Acadêmicas Utilizando Projeto Fatorial. In Proceedings of the 28th Simpósio Brasileiro de Banco de Dados (SBBD), 2013.

Finally, we have also submitted a short paper entitled "Tie Strength Persistence and Transformation" and a full paper called "Strength of Co-authorship Ties in Clusters: a Comparative Analysis" to AMW 2017. Also, we are working to submit the results discussed in Chapter 6.

7.3 Open Problems and Future Work

During this research, we perceived many future directions to this thesis. Beginning with those that can be more easily implemented, we detail them as following.

Expanding the study to other collaboration social networks. The approaches proposed in this thesis to measure tie strength can be applied to other collaboration networks (for example, GitHub and Stack Overflow). We have already started to work on this direction. Specially, we have considered different metrics to measure the strength of social coding collaboration on GitHub [Alves et al., 2016]. One of these metrics is tieness, which shows promising results. While most metrics consider only the network topology, our new metric is able of better differentiating the relationships by considering distinct weights associated to the edges. Note that in such a context edge weight is not co-authorship frequency, but values associated to developers interactions. However, we have not applied STACY neither temporal_tieness on GitHub, i.e., we have not considered the temporal aspect of the relationships on such network. Furthermore, we plan to run the metrics and algorithms in Enron email dataset[1] and other datasets proposed by Barabási [2016].

[1]Enron email dataset: `https://www.cs.cmu.edu/~./enron/`

Using qualitative research to evaluate tie strength. In this thesis, we have evaluated the strength of ties by analyzing Granovetter's theory in non-temporal social networks and link persistence/transformation in temporal social networks. Another direction is asking for users to analyze if they agree or not with their relationships strength generated by our new approaches. Doing so, we are able to build a ground-truth to evaluate our new tie strength metrics and algorithms.

Evaluating tie strength methods by comparing with synthetic data. One of the main problems of working on social network research area is the absence of a ground-truth to evaluate the results. Indeed, a possible solution is to build a synthetic data that represent a completely random and/or perfect social network. Thus, allowing to compare the results from real networks with the synthetic ones. Creating a realistic synthetic data has many challenges related to topologies, data distributions, correlations, attribute values, and so on. To classify tie strength in temporal co-authorship SNs, we have compared the real results with random networks. Nevertheless, we have not done the same for non-temporal social networks. Therefore, we can do similar study to evaluate tieness.

Clustering analyses and evaluation. Due to the common nature of clusters in SN, which is a collection of individuals with dense interactions patterns internally and sparse interactions externally [Mishra et al., 2007]. We believe that tie strength metrics can be used to evaluate clusters quality. In Appendix 4.4, our initial analysis of clustering algorithms (LM, CPM and MCL) by using neighborhood overlap and co-authorship frequency in clustering evaluation showed that MCL is the best clustering algorithm to be applied in co-authorship SN when compared to LM and CPM. Nevertheless, we also conclude that considering only the strength of ties metrics is not enough to define clustering qualities. Therefore, in the next steps, we plan to apply internal measures (like BetaCV, C-index, and so on) to compare with the results generated by the tie strength metrics. Furthermore, we have later identified another clustering algorithm called SCAN (Structural Clustering Algorithm for Networks) [Xu et al., 2007]. In such algorithm, two nodes are assigned to a cluster based on how their share neighbors. SCAN is also able to identify hubs (nodes with high influence in the network) and bursty (have little or no influence) in the social network. Thus, the general concept of this clustering algorithm is related to the concept of neighborhood overlap and tieness. We also intend to analyze the result of SCAN in the collaboration social networks.

Differentiating the α parameter of each property in temporal_tieness. In this thesis, we have evaluated the range of values of temporal_tieness for each class by considering $\alpha = 1$. Although temporal_tieness is able of directly identifying *strong*,

weak and *random* ties, we can study how to better configure such parameter for each topological property. Thus, allowing to classify ties in all the eight classes.

Adding other social network features to STACY. Our new algorithm considers three topological properties to classify tie strength (edge persistence, neighborhood overlap and co-authorship frequency). The main advantage of considering these metrics is that they are free of context. Thus, STACY can be applied to different social networks. Also, as co-authorship frequency represents the edge weight of co-authorship social networks, other properties from different social networks can be considered as edge weight. For example, number of shared repositories for GitHub or frequency of interaction in question & answer forum for Stack Overflow. Nonetheless, other topological and semantic social network properties can be included in STACY. This is not an easy task since the generation of the random graph must also be updated with the new inserted property.

Group recommendation. Another future direction is group recommendation. Overall, our hypothesis is the strength of ties among researchers helps to understand the importance of a relationship to a person, which may improve collaboration recommendation. Thus, a possible application of this thesis is to use the strength of tie metrics associated with clustering algorithms to recommend groups of people to another person. One of the goals of recommendation is to facilitate users find relevant information (about items or people). Hence, the main task is to consider an evaluation metric that measures how "good" recommended groups are to users. Initially, the accuracy of the recommendations may be evaluated using the metrics precision and recall. Then, other evaluation metrics may be considered to better analyzing the characteristics of a target user (who will receive the recommendations), such as diversity or novelty [Brandão et al., 2013; Shani and Gunawardana, 2011].

Bibliography

Abdi, H. (2007). The kendall rank correlation coefficient. *Encyclopedia of Measurement and Statistics*, pages 508--510.

Abdi, H. and Williams, L. (2010). Normalizing data. *Encyclopedia of research design*, pages 935--938.

Abraham, S. M. (2016). Estimating mean time to compromise using non-homogenous continuous-time markov models. In *Procs. of IEEE International Computer Software and Applications Conference (COMPSAC)*, volume 2, pages 467--472, Atlanta, USA.

Acedo et al., F. J. (2006). Co-authorship in management and organizational studies: An empirical and network analysis. *J. of Manag. Studies*, 43(5):957--983.

Adams, B. N. (1967). Interaction theory and the social network. *Sociometry*, pages 64--78.

Adomavicius, G. and Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *TKDE*, 17(6):734--749.

Ahmed, E. B., Nabli, A., and Gargouri, F. (2014). Group extraction from professional social network using a new semi-supervised hierarchical clustering. *Knowledge and information systems*, 40(1):29--47.

Ahmed, M. M., Hafez, A. I., Elwakil, M. M., Hassanien, A. E., and Hassanien, E. (2016). A multi-objective genetic algorithm for community detection in multidimensional social network. In *Procs. of International Conference on Advanced Intelligent System and Informatics (AISI)*, pages 129--139, Beni Suef, Egypt.

Aiello, L. M., Schifanella, R., and State, B. (2014). Reading the source code of social ties. In *Procs. of ACM Conference on Web Science (WebSci)*, pages 139--148, Bloomington, USA.

Akoglu, L. and Dalvi, B. (2010). Structure, tie persistence and event detection in large phone and sms networks. In *Procs. of Workshop on Mining and Learning with Graphs (MLG)*, pages 10--17, Washington, USA. ACM.

Almeida, H., Guedes, D., Meira, W., and Zaki, M. J. (2011). Is there a best quality metric for graph clusters? In *Procs. of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*, pages 44--59, Athens, Greece.

Almeida, V. (2013). Exploring very large data sets from online social networks. In *Procs. of the International Conference on World Wide Web (WWW)*, pages 1165--1166, Rio de Janeiro, Brazil.

Alves, G. B., Brandão, M. A., Santana, D. M., da Silva, A. P. C., and Moro, M. M. (2016). The strength of social coding collaboration on github. In *Procs. of Brazilian Symposium on Databases (SBBD)*, pages 247–252, Salvador, Brazil.

Aral, S. and Walker, D. (2012). Identifying influential and susceptible members of social networks. *Science*, 337(6092):337--341.

Arnaboldi, V., Conti, M., La Gala, M., Passarella, A., and Pezzoni, F. (2016). Ego network structure in online social networks and its impact on information diffusion. *Computer Communications*, 76:26--41.

Atzmueller, M., Ernst, A., Krebs, F., Scholz, C., and Stumme, G. (2016). Formation and temporal evolution of social groups during coffee breaks. In *Procs. of Big Data Analytics in the Social and Ubiquitous Context*, pages 90--108, New York, NY, USA. Springer-Verlag New York, Inc.

Backstrom et al., L. (2006). Group formation in large social networks: Membership, growth, and evolution. In *Procs. of International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 44--54, Philadelphia, USA.

Bagci, H. and Karagoz, P. (2016). Context-aware friend recommendation for location based social networks using random walk. In *Procs. of International Conference Companion on World Wide Web (WWW)*, pages 531--536, Montréal, Canada.

Barabasi, A. (2002). *Linked: The new science of networks*. Perseus Books Group, Cambridge, USA.

Barabási, A.-L. (2016). *Network science*. Cambridge university press.

Barabasi et al., A. L. (2001). Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications*, 311(3):590–614.

Barbosa, E. M., Moro, M. M., Lopes, G. R., and de Oliveira, J. P. M. (2012). Vrrc: web based tool for visualization and recommendation on co-authorship network. In *Procs. of International Conference on Management of Data (SIGMOD)*, pages 865--865, Scottsdale, USA.

Bartusiak et al., R. (2016). Cooperation prediction in github developers network with restricted boltzmann machine. In *Procs. of Asian Conference on Intelligent Information and Database Systems (ACIIDS)*, pages 96--107, Berlin, Heidelberg.

Baumes, J., Goldberg, M. K., Krishnamoorthy, M. S., Magdon-Ismail, M., and Preston, N. (2005). Finding communities by clustering a graph into overlapping subgraphs. In *Procs. of IADIS International Conference on Applied Computing*, pages 97--104, Algarve, Portugal.

Belém, F., Batista, C. S., Santos, R. L. T., Almeida, J. M., and Gonçalves, M. A. (2016). Beyond relevance: explicitly promoting novelty and diversity in tag recommendation. *ACM Transactions on Intelligent Systems and Technology*, 7(3):26:1--26:34.

Bianconi, G. (2013). Statistical mechanics of multiplex networks: Entropy and overlap. *Physical Review E*, 87(6):062806.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Bornmann, L. and Daniel, H.-D. (2007). What do we know about the h index? *Journal of the American Society for Information Science and technology*, 58(9):1381--1385.

Brandão, M. A., Moro, M. M., Lopes, G. R., and Oliveira, J. P. (2013). Using link semantics to recommend collaborations in academic social networks. In *Procs. of International Conference on World Wide Web (WWW) Workshops*, pages 833--840, Rio de Janeiro, Brazil.

Brandão, M. A., Diniz, M. A., and Moro, M. M. (2016). Using topological properties to measure the strength of co-authorship ties. In *Proceedings of Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, pages 199--210, Rio de Janeiro, Brazil.

Brandão, M. A. and Moro, M. M. (2015). Analyzing the strength of co-authorship ties with neighborhood overlap. In *Procs. of the International Conference on Database and Expert Systems Applications (DEXA)*, pages 527--542, Valencia, Spain.

Brandão, M. A. and Moro, M. M. (2017a). Social professional networks: A survey and taxonomy. *Computer Communications*, 100:20 -- 31.

Brandão, M. A. and Moro, M. M. (2017b). The strength of co-authorship ties through different topological properties. *Journal of the Brazilian Computer Society*, 23(1):5.

Brandão, M. A., Moro, M. M., and Almeida, J. M. (2014). Experimental evaluation of academic collaboration recommendation using factorial design. *Journal of Information and Data Management*, 5(1):52.

Brandes, U., Indlekofer, N., and Mader, M. (2012). Visualization methods for longitudinal social networks and stochastic actor-oriented modeling. *Social Networks*, 34(3):291--308.

Brase, C. H. and Brase, C. P. (2012). *Understanding basic statistics*. Cengage Learning.

Bruggeman, J. (2016). The strength of varying tie strength: Comment on aral and van alstyne 1. *American Journal of Sociology*, 121(6):1919--1930.

Burt, R. S. (2004). Structural Holes and Good Ideas. *The American Journal of Sociology*, 110(2):349--399.

Burt, R. S. (2010). *Neighbor networks: Competitive advantage local and personal*. Oxford University Press.

Cameron, A. C. and Windmeijer, F. A. (1997). An r-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics*, 77(2):329 – 342.

Cano et al., P. (2006). Topology of music recommendation networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 16(1):013107–1---013107–6.

Carpineto, C. and Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Computing Surveys*, 44(1):1--50.

Casalnuovo, C., Vasilescu, B., Devanbu, P., and Filkov, V. (2015). Developer onboarding in github: the role of prior social links and language experience. In *Procs. of the Joint Meeting on Foundations of Software Engineering (ESEC/FSE)*, pages 817--828, Bergamo, Italy.

Castilho, D., de Melo, P. O. V., and Benevenuto, F. (2017). The strength of the work ties. *Information Sciences*, 375:155--170.

Cellary, W., Morzy, T., and Gelenbe, E. (2014). *Concurrency control in distributed database systems*. Elsevier.

Chan, H. F., Önder, A. S., and Torgler, B. (2016). The first cut is the deepest: repeated interactions of coauthorship and academic productivity in nobel laureate teams. *Scientometrics*, 106(2):509--524.

Chang, C.-C. and Chin, Y.-C. (2011). Predicting the usage intention of social network games: an intrinsic-extrinsic motivation theory perspective. *International Journal of Online Marketing (IJOM)*, 1(3):29--37.

Chau, D. H., Pandit, S., Wang, S., and Faloutsos, C. (2007). Parallel crawling for online social networks. In *Procs. of International Conference on World Wide Web (WWW)*, pages 1283--1284, Banff, Canada.

Chaudhuri, S., Das, G., Hristidis, V., and Weikum, G. (2004). Probabilistic ranking of database query results. In *Procs. of International Conference on Very Large Data Bases (VLDB)*, pages 888--899, Toronto, Canada.

Chen, H.-H., Gou, L., Zhang, X., and Giles, C. L. (2011). Collabseer: a search engine for collaboration discovery. In *Procs. of Annual International ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 231--240, Ottawa, Canada.

Chen, J., Zaïane, O., and Goebel, R. (2009). Local community identification in social networks. In *Procs. of International Conference on Advances in Social Network Analysis and Mining (ASONAM)*, pages 237--242, Athens, Greece.

Chen, X., Yang, C.-Z., Lu, T.-K., and Jaygarl, H. (2008). Implicit social network model for predicting and tracking the location of faults. In *Procs. of Annual IEEE International Computer Software and Applications Conference (COMPSAC)*, pages 136--143, Turku, Finland.

Chen, Z. and Ji, H. (2010). Graph-based clustering for computational linguistics: A survey. In *Procs. of the Workshop on Graph-based Methods for Natural Language Processing*, pages 1--9, Uppsala, Sweden.

Chen, Z., Xia, F., Jiang, H., Liu, H., and Zhang, J. (2015). Aver: Random walk based academic venue recommendation. In *Procs. of International Conference on World Wide Web (WWW)*, pages 579--584, Florence, Italy.

Cheng, C.-B., Day, M.-Y., Shih, S.-P., and Chang, W. (2014). Study of scientific collaborations in the intelligence and security informatics research community by social network analysis. In *Procs. of Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, Rio de Janeiro, Brazil.

Chok, N. S. (2010). *Pearson's versus Spearman's and Kendall's correlation coefficients for continuous data.* PhD thesis, University of Pittsburgh.

Chung, F. and Lu, L. (2002). Connected components in random graphs with given expected degree sequences. *Annals of combinatorics*, 6(2):125--145.

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

Corbellini, A., Mateos, C., Zunino, A., Godoy, D., and Schiaffino, S. (2017). Persisting big-data: The nosql landscape. *Information Systems*, 63:1--23.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2009). *Introduction to algorithms.* MIT press.

Dabbish, L., Stuart, C., Tsay, J., and Herbsleb, J. (2012). Social coding in github: Transparency and collaboration in an open software repository. In *Procs. of ACM Conference on Computer Supported Cooperative Work (CSCW)*, pages 1277--1286, Seattle, USA.

D'agostino, G. and Scala, A. (2014). *Networks of networks: the last frontier of complexity*, volume 340. Springer.

Dale, M. R. T. and Fortin, M.-J. (2010). From graphs to spatial graphs. *Annual Review of Ecology, Evolution, and Systematics*, 41.

Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukherjea, S., Nanavati, A. A., and Joshi, A. (2008). Social ties and their relevance to churn in mobile telecom networks. In *Procs. of international conference on Extending database technology: Advances in database technology (EDBT)*, pages 668--677, Nantes, France.

Dawande, M., Keskinocak, P., Swaminathan, J. M., and Tayur, S. (2001). On bipartite and multipartite clique problems. *Journal of Algorithms*, 41(2):388--403.

de la Maza, M. (2007). Luv: A programming language for describing human relationships. In *Procs. of Annual IEEE International Computer Software and Applications Conference (COMPSAC)*, volume 1, pages 645--646, Beijing, China.

de Sousa, G. A., Diniz, M. A., Brandao, M. A., and Moro, M. M. (2015). Cnare: Co-authorship social network analysis and recommendations. In *Procs. of ACM Conference on Recommender Systems (RecSys)*, pages 329–330, Vienna, Austria.

Deb, D., Vishveshwara, S., and Vishveshwara, S. (2009). Understanding protein structure from a percolation perspective. *Biophysical journal*, 97(6):1787--1794.

Delis, A., Ntoulas, A., and Liakos, P. (2016). Scalable link community detection: A local dispersion-aware approach. In *Procs. of the 2016 IEEE International Conference on Big Data (Big Data)*, pages 716--725, Washington, USA.

Dickison, M. E., Magnani, M., and Rossi, L. (2016). *Multilayer Social Networks*. Cambridge University Press.

Digiampietri, L. and Maruyama, W. (2014). Predição de novas coautorias na rede social acadêmica dos programas brasileiros de pós-graduação em ciência da computação. In *Procs. of Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, pages 243--248, Rio de Janeiro, Brazil.

Dom, B., Eiron, I., Cozzi, A., and Zhang, Y. (2003). Graph-based ranking algorithms for e-mail expertise analysis. In *Procs. of ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 42--48, San Diego, USA.

Dom, B. E. (2002). An information-theoretic external cluster-validity measure. In *Procs. of Conference on Uncertainty in artificial intelligence (UAI)*, pages 137--145, Alberta, Canada.

Ductor, L. (2015). Does co-authorship lead to higher academic productivity? *Oxford Bulletin of Economics and Statistics*, 77(3):385--407.

Easley, D. and Kleinberg, J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press.

Elmacioglu, E. and Lee, D. (2005). On six degrees of separation in dblp-db and more. *ACM SIGMOD Record*, 34(2):33--40.

Erdös, P. and Rényi, A. (1959). On random graphs i. *Publicationes Mathematicae (Debrecen)*, 6:290--297.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3):75--174.

Fortunato, S. and Hric, D. (2016). Community detection in networks: A user guide. *Physics Reports*, 659:1--44.

Fouss, F. and Saerens, M. (2008). Evaluating performance of recommender systems: An experimental comparison. In *Procs. of Web Intelligence and Intelligent Agent Technology (WIIAT)*, pages 735--738, Sydney, Australia.

Freeman, L. C. (2000). Visualizing social networks. *Journal of social structure*, 1(1):4.

Freire, V. P. and Figueiredo, D. R. (2011). Ranking in collaboration networks using a group based metric. *Journal of the Brazilian Computer Society*, 17(41):255--266.

Fu, T. Z., Song, Q., and Chiu, D. M. (2014). The academic social network. *Scientometrics*, 101(1):203--239.

Fujino, H., Hasida, K., and Matsubara, Y. (2017). Music exploration by impression based interaction. In *Procs. of ACM Workshop on Exploratory Search and Interactive Data Analytics (ESIDA)*, pages 55--58, Limassol, Cyprus.

Garcia-Molina, H., Ullman, J. D., and Widom, J. (2000). *Database system implementation*, volume 654. Prentice Hall Upper Saddle River, NJ:.

Ge, M., Delgado-Battenfeld, C., and Jannach, D. (2010). Beyond accuracy: Evaluating recommender systems by coverage and serendipity. In *Procs. of ACM Conference on Recommender Systems (RecSys)*, pages 257--260, Barcelona, Spain.

Ghosh, R. and Lerman, K. (2009). Structure of heterogeneous networks. In *Procs. of International Conference on Computational Science and Engineering (CSE)*, volume 4, pages 98--105, Vancouver, Canada.

Gilbert, E. and Karahalios, K. (2009). Predicting tie strength with social media. In *Procs. of Conference on Human Factors in Computing Systems (SIGCHI)*, pages 211--220, Boston, USA.

Giridhar, P., Wang, S., and Abdelzaher, T. (2017). Demo abstract: Event tracker using social networks. In *Procs. of the IEEE/ACM Second International Conference on Internet-of-Things Design and Implementation (IoTDI)*, pages 285--286, Pittsburgh, USA.

Girvan, M. and Newman, M. E. (2002). Community structure in social and biological networks. *National Academy of Sciences*, 99(12):7821--7826.

Gjoka, M., Butts, C. T., Kurant, M., and Markopoulou, A. (2011a). Multigraph sampling of online social networks. *IEEE Journal on Selected Areas in Communications*, 29(9):1893--1905.

Gjoka, M., Kurant, M., Butts, C. T., and Markopoulou, A. (2011b). Practical recommendations on crawling online social networks. *Selected Areas in Communications*, 29(9):1872--1892.

Glänzel, W. and Schubert, A. (2005). Analysing scientific networks through co-authorship. In *Handbook of quantitative science and technology research*, pages 257--276. Springer.

Gómez et al., D. (2015). A divide-and-link algorithm for hierarchical clustering in networks. *Information Sciences*, 316:308--328.

Gonçalves, G. D., Figueiredo, F., Almeida, J. M., and Gonçalves, M. A. (2014). Characterizing scholar popularity: A case study in the computer science research community. In *Procs. of ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, Piscataway, USA.

Goodman, L. A. (1961). Snowball sampling. *The annals of mathematical statistics*, pages 148--170.

Goulas, A., Schaefer, A., and Margulies, D. S. (2015). The strength of weak connections in the macaque cortico-cortical network. *Brain Structure and Function*, 220(5):2939--2951.

Granovetter, M. S. (1973). The strength of weak ties. *The American Journal of Sociology*, 78(6):1360--1380.

Gross, E. (1956). Symbiosis and consensus as integrative factors in small groups. *American Sociological Review*, 21(2):174--179.

Guerra-Gomez, J. A., Wilson, A., Liu, J., Davies, D., Jarvis, P., and Bier, E. (2016). Network explorer: Design, implementation, and real world deployment of a large network visualization tool. In *Procs. of International Working Conference on Advanced Visual Interfaces (AVI)*, pages 108--111, Bari, Italy.

Guille, A., Hacid, H., Favre, C., and Zighed, D. A. (2013). Information diffusion in online social networks: A survey. *ACM SIGMOD Record*, 42(2):17--28.

Gupte, M. and Eliassi-Rad, T. (2012). Measuring tie strength in implicit social networks. In *Procs. of Annual ACM Web Science Conference (WebSci)*, pages 109--118, Evanston, USA.

Han, J., Haihong, E., Le, G., and Du, J. (2011). Survey on nosql database. In *Procs. of International Conference on Pervasive Computing and Applications (ICPCA)*, pages 363--366, Port Elizabeth, South Africa.

Harman, D. (1992). Ranking algorithms. In Frakes, W. B. and Baeza-Yates, R., editors, *Information retrieval*, pages 363--392. Prentice-Hall, Inc., Upper Saddle River, USA.

Harman, M., Swift, S., and Mahdavi, K. (2005). An empirical study of the robustness of two module clustering fitness functions. In *Procs. of Annual Conference on Genetic and Evolutionary Computation (GECCO)*, pages 1029--1036, Washington, USA.

Harth, A., Umbrich, J., and Decker, S. (2006). Multicrawler: A pipelined architecture for crawling and indexing semantic web data. In Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., and Aroyo, L., editors, *The Semantic Web - ISWC*, volume 4273, pages 258–271. Springer Berlin Heidelberg.

He, J. and Chu, W. W. (2010). *A social network-based recommender system (SNRS)*. Springer.

Heide, J. B. and Wathne, K. H. (2006). Friends, businesspeople, and relationship roles: A conceptual framework and a research agenda. *Journal of Marketing*, 70(3):90--103.

Holten, D. and Van Wijk, J. J. (2009). Force-directed edge bundling for graph visualization. In *Procs. of Eurographics / IEEE - VGTC Conference on Visualization (EuroVis)*, pages 983--990, Berlin, Germany.

Hristova, D., Williams, M. J., Musolesi, M., Panzarasa, P., and Mascolo, C. (2016). Measuring urban social diversity using interconnected geo-social networks. In *Procs. of the International Conference on World Wide Web (WWW)*, pages 21--30, Montréal, Canada.

Huang, T.-H. and Huang, M. L. (2006). Analysis and visualization of co-authorship networks for understanding academic collaboration and knowledge domain of individual researchers. In *Procs. of International Conference on Computer Graphics, Imaging and Visualisation (CGIV)*, pages 18--23, Sydney, Australia.

Huynh, T., Luong, H., and Hoang, K. (2012). Integrating bibliographical data of computer science publications from online digital libraries. In Pan, J.-S., Chen, S.-M., and Nguyen, N., editors, *Intelligent Information and Database Systems*, volume 7198, pages 226–235. Springer Berlin Heidelberg.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for clustering data*, volume 6. Prentice hall Englewood Cliffs.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, 31(3):264--323.

Jain, R. (1991). *The Art of Computer Systems Performance Analysis: techniques for experimental design, measurement, simulation, and modeling*. John Wiley & Sons.

Jiang, J., Sheng, W., Yu, S., Xiang, Y., and Zhou, W. (2016). Rumor source identification in social networks with time-varying topology. *IEEE Trans. Dependable and Secure Computing*, PP(99):1--15.

Jin, Y., Lin, C.-Y., Matsuo, Y., and Ishizuka, M. (2012). Mining dynamic social networks from public news articles for company value prediction. *Social Network Analysis and Mining*, 2(3):217--228.

Kadushin, C. (2012). *Understanding social networks: Theories, concepts, and findings*. Oxford University Press.

Kahanda, I. and Neville, J. (2009). Using transactional information to predict link strength in online social networks. In *Procs. of International AAAI Conference on Weblogs and Social Media*, pages 74--81, San Jose, USA.

Kang, C., Pugliese, A., Grant, J., and Subrahmanian, V. (2014). Stun: querying spatio-temporal uncertain (social) networks. *Social Network Analysis and Mining*, 4(1):1--19.

Karsai, M., Perra, N., and Vespignani, A. (2014). Time varying networks and the weakness of strong ties. *Scientific reports*, 4.

Kempe, D., Kleinberg, J., and Tardos, E. (2003). Maximizing the spread of influence through a social network. In *Procs. of ACM International Conference on Knowledge discovery and data mining (SIGKDD)*, pages 137--146, Washington, USA.

Keyes, C. (2015). *Data mining: Analysis of social network' s: Facebook, Twitter, LinkedIn, Google+, and more.* Meteor Content Providers.

Khan, K. S., Kunz, R., Kleijnen, J., and Antes, G. (2003). Five steps to conducting a systematic review. *Journal of the Royal Society of Medicine*, 96(3):118--121.

Kim, P. and Kim, S. (2014). A detection of overlapping community in mobile social network. In *Procs. of ACM Symposium On Applied Computing (SAC)*, pages 175--179, Gyeongju, Republic of Korea.

Kitchenham, B., Pearl Brereton, O., Budgen, D., Turner, M., Bailey, J., and Linkman, S. (2009). Systematic literature reviews in software engineering - a systematic literature review. *Inf. Softw. Technol.*, 51(1):7--15.

Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., and Porter, M. A. (2014). Multilayer networks. *Journal of complex networks*, 2(3):203--271.

Kleinberg, J. M. (2007). Challenges in mining social network data: Processes, privacy, and paradoxes. In *Procs. of ACM International Conference on Knowledge discovery and data mining (SIGKDD)*, pages 4--5, San Jose, USA.

Knoke, D. and Yang, S. (2008). *Social network analysis*. Sage.

Koo, D.-M. (2016). Impact of tie strength and experience on the effectiveness of online service recommendations. *Electronic Commerce Research and Applications*, 15:38--51.

Kostakos, V. (2009). Temporal graphs. *Physica A: Statistical Mechanics and its Applications*, 388(6):1007--1023.

Kotsiantis, S. B., Kanellopoulos, D., and Pintelas, P. E. (2006). Data preprocessing for supervised leaning. *International Journal of Computer Science*, 1(2):111--117.

Kramer, A. D. (2010). An unobtrusive behavioral model of gross national happiness. In *Procs. of Conference on Human Factors in Computing Systems (SIGCHI)*, pages 287--290, Atlanta, USA.

Kshitij, A., Ghosh, J., and Gupta, B. M. (2015). Embedded information structures and functions of co-authorship networks: Evidence from cancer research collaboration in india. *Scientometrics*, 102(1):285--306.

Lancichinetti, A., Fortunato, S., and Kertész, J. (2009). Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015.

Laurent, G., Saramäki, J., and Karsai, M. (2015). From calls to communities: a model for time-varying social networks. *The European Physical Journal B*, 88(11):1--10.

Lee, H. (2015). Uncovering the multidisciplinary nature of technology management: Journal citation network analysis. *Scientometrics*, 102(1):51--75.

Lewis, P. and McKenzie, E. (1988). *Simulation methodology for statisticians, operations analysts, and engineers*, volume 1. CRC press.

Li et al., K. (2012). Efficient algorithm based on neighborhood overlap for community identification in complex networks. *Physica A: Statistical Mechanics and its Applications*, 391(4):1788--1796.

Liben-Nowell, D. and Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019--1031.

Lima, H., Silva, T. H., Moro, M. M., Santos, R. L., Meira, Jr., W., and Laender, A. H. (2013). Aggregating productivity indices for ranking researchers across multiple areas. In *Procs. of ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, pages 97--106, Indianapolis, USA.

Lopes, G. R., Moro, M. M., Wives, L. K., and De Oliveira, J. P. M. (2010). Collaboration recommendation on academic social networks. In *Procs. of International Conference on Conceptual Modeling (ER)*, pages 190--199, Berlin, Heidelberg.

Lopes et al., G. R. (2011). Ranking strategy for graduate programs evaluation. In *Procs. of IEEE International Conference on Information Technology and Applications (ICITA)*, pages 56--64, Sydney, Australia.

Lops, P., de Gemmis, M., and Semeraro, G. (2011). Content-based recommender systems: State of the art and trends. In Ricci, F., Rokach, L., Shapira, B., and Kantor, P. B., editors, *Recommender Systems Handbook*, pages 73–105. Springer US.

Lotero, L., Cardillo, A., Hurtado, R., and Gómez-Gardeñes, J. (2016). Several multiplexes in the same city: The role of socioeconomic differences in urban mobility. In *Interconnected Networks*, pages 149--164. Springer.

Luna, J. E. O., Revoredo, K., and Cozman, F. G. (2013). Link prediction using a probabilistic description logic. *Journal of the Brazilian Computer Society*, 19(108).

Malliaros, F. D. and Vazirgiannis, M. (2013). Clustering and community detection in directed networks: A survey. *Physics Reports*, 533(4):95--142.

Manikandan, S. (2011). Measures of central tendency: Median and mode. *Journal of Pharmacology and Pharmacotherapeutics*, 2(3):214.

Meilă, M. (2007). Comparing clusterings–an information based distance. *Journal of multivariate analysis*, 98(5):873--895.

Melamed, D. (2015). Communities of classes: A network approach to social mobility. *Research in social stratification and mobility*, 41:56--65.

Meng, L., Hulovatyy, Y., Striegel, A., and Milenković, T. (2016). On the interplay between individuals' evolving interaction patterns and traits in dynamic multiplex social networks. *IEEE Transactions on Network Science and Engineering*, 3(1):32--43.

Mihalcea, R. and Radev, D. (2011). *Graph-based natural language processing and information retrieval*. Cambridge University Press.

Miller, J. C. and Hagberg, A. (2011). Efficient generation of networks with given expected degrees. In *Procs. of International Workshop on Algorithms and Models for the Web-Graph (WAW)*, pages 115--126, Barcelona, Spain.

Miritello, G., Moro, E., and Lara, R. (2011). Dynamical strength of social ties in information spreading. *Physical Review E*, 83(4):045--102.

Mishra et al., N. (2007). Clustering social networks. In Bonato, A. and Chung, F. R. K., editors, *Algorithms and Models for the Web-Graph*, pages 56--67. Springer.

Montolio, S. L., Dominguez-Sal, D., and Larriba-Pey, J. L. (2013). Research endogamy as an indicator of conference quality. *ACM SIGMOD Record*, 42(2):11–16.

Mrvar, A. and Batagelj, V. (2016). Analysis and visualization of large networks with program package pajek. *Complex Adaptive Systems Modeling*, 4(1):1--8.

Mucha, P. J. and Porter, M. A. (2010). Communities in multislice voting networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 20(4):041108.

Murray, S. (2013). *Interactive Data Visualization for the Web*. O'Reilly Media, Inc.

Narayanan, A., Shi, E., and Rubinstein, B. I. (2011). Link prediction by de-anonymization: How we won the kaggle social network challenge. In *Procs. of International Joint Conference on Neural Networks (IJCNN)*, pages 1825--1834, San Jose, USA.

Newman, M. E. (2001a). Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical review E*, 64(1):016132.

Newman, M. E. J. (2001b). The structure of scientific collaboration networks. *National Academy of Sciences*, 98(2):404--409.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, 45(2):167--256.

Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113.

Nicosia et al., V. (2013). Graph metrics for temporal networks. In Holme, P. and Saramäki, J., editors, *Temporal Networks*, pages 15--40. Springer Berlin Heidelberg, Berlin, Heidelberg.

Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., and Barabási, A.-L. (2007). Structure and tie strengths in mobile communication networks. *Procs. of National Academy of Sciences*, 104(18):7332--7336.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Procs. of International Conference on Language Resources and Evaluation (LREC)*, pages 1320--1326, Valletta, Malta.

Palla, G., Barabási, A., and Vicsek, T. (2007). Quantifying social group evolution. *Nature*, 446(7136):664--667.

Palla et al., G. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814--818.

Pan, R. K. and Saramäki, J. (2012). The strength of strong ties in scientific collaboration networks. *EPL (Europhysics Letters)*, 97(1):18007.

Paraschiv, I. C., Dascalu, M., Dessus, P., Trausan-Matu, S., and McNamara, D. S. (2016). A paper recommendation system with readerbench: The graphical visualization of semantically related papers and concepts. In *State-of-the-Art and Future Directions of Smart Learning*, pages 445--451. Springer.

Park, S., Kim, I., Lee, S. W., Yoo, J., Jeong, B., and Cha, M. (2015). Manifestation of depression and loneliness on social networks: A case study of young adults on facebook. In *Procs. of ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW)*, pages 557--570, Vancouver, Canada.

Pendlebury, D. A. (2009). The use and misuse of journal metrics and other citation indicators. *Archivum immunologiae et therapiae experimentalis*, 57(1):1--11.

Peng, S., Yang, A., Cao, L., Yu, S., and Xie, D. (2017). Social influence modeling using information theory in mobile social networks. *Information Sciences*, 379:146--159.

Pettenati, M. C. and Cigognini, M. E. (2007). Social networking theories and tools to support connectivist learning activities. *International Journal of Web-Based Learning and Teaching Technologies (IJWLTT)*, 2(3):42--60.

Protasiewicz, J., Pedrycz, W., Kozłowski, M., Dadas, S., Stanisławek, T., Kopacz, A., and Gałężewska, M. (2016). A recommender system of reviewers and experts in reviewing problems. *Knowledge-Based Systems*.

Pu, P., Chen, L., and Hu, R. (2012). Evaluating recommender systems from the user's perspective: survey of the state of the art. *User Modeling and User-Adapted Interaction*, 22(4-5):317--355.

Rabbany et al., R. (2014). Relative validity criteria for community mining algorithms. In *Encyclopedia of Social Network Analysis and Mining*, pages 1562--1576. Springer.

Raeder, T., Lizardo, O., Hachen, D., and Chawla, N. V. (2011). Predictors of short-term decay of cell phone contacts in a large scale communication network. *Social Networks*, 33(4):245--257.

Rahman, M. and Karim, R. (2016). Comparative study of different methods of social network analysis and visualization. In *Procs. of International Conference on Networking Systems and Security (NSysS)*, pages 1--7, Dhaka, Bangladesh.

Rana, J., Kristiansson, J., and Synnes, K. (2014). The strength of social strength: An evaluation study of algorithmic versus user-defined ranking. In *Procs. of ACM Symposium On Applied Computing (SAC)*, Gyeongju, Republic of Korea.

Rezvanian, A. and Meybodi, M. R. (2015). Sampling social networks using shortest paths. *Physica A: Statistical Mechanics and its Applications*, 424:254–268.

Ribas et al., S. (2015). Using reference groups to assess academic productivity in computer science. In *Procs. of International Conference on World Wide Web (WWW)*, Florence, Italy.

Rosenberg, A. and Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *Procs. of Conference on Empirical Methods in Natural Language Processing-Conference on Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410--420, Prague, Czech Republic.

Różewski, P., Jankowski, J., Brodka, P., and Michalski, R. (2015). Knowledge workers' collaborative learning behavior modeling in an organizational social network. *Computers in Human Behavior*, 51, Part B:1248–1260.

Russell, M. A. (2013). *Mining the social web: Data mining Facebook, Twitter, LinkedIn, Google+, GitHub, and more*. O´Reilly Media, Inc.

Salehi, M., Sharma, R., Marzolla, M., Magnani, M., Siyari, P., and Montesi, D. (2015). Spreading processes in multilayer networks. *IEEE Transactions on Network Science and Engineering*, 2(2):65--83.

Salehi-Abari, A. and Boutilier, C. (2015). Preference-oriented social networks: Group recommendation and inference. In *Procs. of ACM Conference on Recommender Systems (RecSys)*, pages 35--42, Vienna, Austria.

Sales-Pardo, M., Guimera, R., Moreira, A. A., and Amaral, L. A. N. (2007). Extracting the hierarchical organization of complex systems. *National Academy of Sciences*, 104(39):15224--15229.

Satuluri, V. and Parthasarathy, S. (2009). Scalable graph clustering using stochastic flows: Applications to community discovery. In *Procs. of ACM International conference on Knowledge discovery and data mining (SIGKDD)*, pages 737--746, Paris, France.

Satuluri, V., Parthasarathy, S., and Ucar, D. (2010). Markov clustering of protein interaction networks with improved balance and scalability. In *Procs. of ACM International Conference on Bioinformatics and Computational Biology (BCB)*, pages 247--256, Niagara Falls, New York.

Scellato, S., Noulas, A., and Mascolo, C. (2011). Exploiting place features in link prediction on location-based social networks. In *Procs. of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 1046--1054, San Diego, USA. ACM.

Schall, D. (2014). Who to follow recommendation in large-scale online development communities. *Information and Software Technology*, 56(12):1543--1555.

Scott, J. and Carrington, P. J. (2011). *The SAGE handbook of social network analysis*. SAGE publications.

Selassie, D., Heller, B., and Heer, J. (2011). Divided edge bundling for directional network data. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2354--2363.

Seo, Y.-D., Kim, Y.-G., Lee, E., and Baik, D.-K. (2017). Personalized recommender system based on friendship strength in social network services. *Expert Systems with Applications*, 69:135--148.

Shani, G. and Gunawardana, A. (2011). Evaluating recommendation systems. In Ricci, F., Rokach, L., Shapira, B., and Kantor, P. B., editors, *Recommender Systems Handbook*, pages 257–297. Springer US, Boston, USA.

Sharma, A. and Yan, B. (2013). Pairwise learning in recommendation: Experiments with community recommendation on linkedin. In *Procs. of ACM conference on Recommender systems (RecSys)*, pages 193--200, Hong Kong, China.

Sharma, J. (2012). *Business statistics*. Pearson Education India.

Silva, T. H., da Silva, A. P. C., and Moro, M. M. (2015a). tc-index: A new research productivity index based on evolving communities. In Kapidakis, S., Mazurek, C., and Werla, M., editors, *Research and Advanced Technology for Digital Libraries*, volume 9316, pages 209–221. Springer International Publishing.

Silva, T. H. P., Moro, M. M., and Silva, A. P. C. (2015b). Authorship contribution dynamics on publication venues in computer science: An aggregated quality analysis. In *Procs. of ACM Symposium On Applied Computing (SAC)*, pages 1142--1147, Salamanca, Spain.

Silva, T. H. P., Moro, M. M., Silva, A. P. C., Meira, Jr., W., and Laender, A. H. F. (2014). Community-based endogamy as an influence indicator. In *Procs. of ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, pages 67--76, London, United Kingdom.

Simon, R. J. (1974). The work habits of eminent scholars. *Work and Occupations*, 1(3):327--335.

Skeels, M. M. and Grudin, J. (2009). When social networks cross boundaries: a case study of workplace use of facebook and linkedin. In *Procs. of ACM International*

Conference on Supporting Group Work (GROUP), pages 95--104, Sanibel Island, USA.

Song, I., Dillon, D., Goh, T., and Sung, M. (2011). A health social network recommender system. *Agents in Principle, Agents in Practice*, pages 361--372.

Sorenson, O. (2005). Social networks and industrial geography. In Cantner, U., Dinopoulos, E., and Lanzillotti, R., editors, *Entrepreneurships, the New Economy and Public Policy*, pages 55–69. Springer Berlin Heidelberg.

Subbian, K., Prakash, B. A., and Adamic, L. (2017). Detecting large reshare cascades in social networks. In *Procs. of the International Conference on World Wide Web (WWW)*, pages 597--605, Perth, Australia.

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

Sun, Y., Han, J., Zhao, P., Yin, Z., Cheng, H., and Wu, T. (2009). Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In *Procs. of International Conference on Extending Database Technology: Advances in Database Technology (EDBT)*, pages 565--576, Saint Petersburg, Russia.

Sun et al., Z. (2015). Recommender systems based on social networks. *Journal of Systems and Software*, 99:109--119.

Surian, D., Liu, N., Lo, D., Tong, H., Lim, E.-P., and Faloutsos, C. (2011). Recommending people in developers' collaboration network. In *Procs. of Working Conference on Reverse Engineering (WCRE)*, pages 379--388, Limerick, Ireland.

Tabarzad, M. A. and Hamzeh, A. (2017). A heuristic local community detection method (hlcd). *Applied Intelligence*, 46(1):62--78.

Tan, P.-N., Steinbach, M., and Kumar, V. (2006). *Introduction to data mining*, volume 1. Pearson Addison Wesley Boston.

Tang, J., Hu, X., and Liu, H. (2013). Social recommendation: a review. *Social Network Analysis and Mining*, 3(4):1113--1133.

Tang, J., Zhang, D., and Yao, L. (2007). Social network extraction of academic researchers. In *Procs. of IEEE International Conference on Data Mining (ICDM)*, pages 292--301, Omaha, USA.

Tang, L., Long, B., Chen, B.-C., and Agarwal, D. (2016). An empirical study on recommendation with multiple types of feedback. In *Procs. of ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 283--292, San Francisco, USA.

Thung, F., Bissyande, T. F., Lo, D., and Jiang, L. (2013). Network structure of social coding in github. In *Procs. of European Conference on Software Maintenance and Reengineering (CSMR)*, pages 323--326, Genova, Italy.

Trusov, M., Bucklin, R. E., and Pauwels, K. (2009). Effects of word-of-mouth versus traditional marketing: Findings from an internet social networking site. *Journal of marketing*, 73(5):90--102.

Turian, J., Ratinov, L., and Bengio, Y. (2010). Word representations: a simple and general method for semi-supervised learning. In *Procs. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 384--394, Uppsala, Sweden.

Valente, T. W., Palinkas, L. A., Czaja, S., Chu, K.-H., and Brown, C. H. (2015). Social network analysis for program implementation. *PloS one*, 10(6):e0131712.

Van Dongen, S. M. (2000). *Graph clustering by flow simulation.* PhD thesis, Utrecht University.

Vaz de Melo et al., P. O. S. (2015). Recast: Telling apart social and random relationships in dynamic networks. *Performance Evaluation*, 87:19--36.

Viana, W., da Silva, A. P. C., and Moro, M. M. (2016). Pick the right team and make a blockbuster: a social analysis through movie history. In *Procs. of Annual ACM Symposium on Applied Computing (SAC)*, pages 1108–1114, Pisa, Italy.

Viégas, F. B. and Donath, J. (2004). Social network visualization: Can we go beyond the graph. In *Procs. of Workshop on Social Networks*, volume 4, pages 6--10, Chicago, USA.

Vural, A. G., Cambazoglu, B. B., and Karagoz, P. (2014). Sentiment-focused web crawling. *ACM Trans. Web*, 8(4):22:1--22:21.

Wal, T., LJ, A., and Boschma, R. A. (2009). Applying social network analysis in economic geography: framing some key analytic issues. *The Annals of Regional Science*, 43(3):739--756.

Wang, D., Li, J., Xu, K., and Wu, Y. (2017). Sentiment community detection: Exploring sentiments and relationships in social networks. *Electronic Commerce Research*, 17(1):103--132.

Wang, D., Yan, K.-K., Rozowsky, J., Pan, E., and Gerstein, M. (2016). Temporal dynamics of collaborative networks in large scientific consortia. *Trends in Genetics*, 32(5):251--253.

Wang et al., P. (2016). Social selection models for multilevel networks. *Social Networks*, 44:346--362.

Wasserman, S. (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge university press.

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*, 393(6684):409--10.

Weng et al., J. (2010). Twitterrank: finding topic-sensitive influential twitterers. In *Procs. of ACM international conference on Web search and data mining (WSDM)*, pages 261--270, New York, USA.

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.

Wiese, J., Min, J.-K., Hong, J. I., and Zimmerman, J. (2015). You never call, you never write: Call and sms logs do not always indicate tie strength. In *Procs. of ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW)*, pages 765--774, Vancouver, Canada.

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

Wu, W., He, L., and Yang, J. (2012). Evaluating recommender systems. In *Procs. of International Conference on Digital Information Management (ICDIM)*, pages 56--61, Macau, China.

Xia, F., Chen, Z., Wang, W., Li, J., and Yang, L. T. (2014). Mvcwalker: Random walk-based most valuable collaborators recommendation exploiting academic factors. *IEEE Transactions on Emerging Topics in Computing*, 2(3):364--375.

Xie, J., Kelley, S., and Szymanski, B. K. (2013). Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM Computing Surveys*, 45(4):43:1--43:35.

Xu, X., Yuruk, N., Feng, Z., and Schweiger, T. A. J. (2007). Scan: a structural clustering algorithm for networks. In *Procs. of ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 824--833, San Jose, USA.

Xuan, Y., Chen, Y., Li, H., Hui, P., and Shi, L. (2016). Lbsnshield: Malicious account detection in location-based social networks. In *Procs. of the Conference on Computer Supported Cooperative Work and Social Computing Companion (CSCW)*, pages 437--440, New York, USA.

Yan, R., Huang, C., Tang, J., Zhang, Y., and Li, X. (2012). To better stand on the shoulder of giants. In *Procs. of ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, pages 51--60, Washington, USA.

Yang, C., Sun, J., Ma, J., Zhang, S., Wang, G., and Hua, Z. (2015). Scientific collaborator recommendation in heterogeneous bibliographic networks. In *Procs. of Hawaii International Conference on System Sciences (HICSS)*, pages 552--561, Kauai, USA.

Yang, X., Guo, Y., Liu, Y., and Steck, H. (2014). A survey of collaborative filtering based social recommender systems. *Computer Communications*, 41:1--10.

Yasin, A. and Liu, L. (2016). An online identity and smart contract management system. In *Procs. of Annual IEEE International Computer Software and Applications Conference (COMPSAC)*, volume 2, pages 192--198, Atlanta, USA.

Yu, S., Liu, M., Dou, W., Liu, X., and Zhou, S. (2017). Networking for big data: A survey. *IEEE Communications Surveys & Tutorials*, 19(1):531--549.

Yu, Y., Wang, H., Yin, G., and Ling, C. X. (2014). Reviewer recommender of pull-requests in github. In *Procs. of IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 609--612, Victoria, Canada.

Yu, Y., Wang, H., Yin, G., and Wang, T. (2016a). Researchgate: An effective altmetric indicator for active researchers? *Computers in Human Behavior*, 55:1001--1006.

Yu, Y., Wang, H., Yin, G., and Wang, T. (2016b). Reviewer recommendation for pull-requests in github: What can we learn from code review and bug assignment? *Information and Software Technology*, 74:204---218.

Zaki, M. J. and Meira Jr, W. (2014). *Data mining and analysis: Fundamental concepts and algorithms*. Cambridge University Press.

Zhang, A. X., Bhardwaj, A., and Karger, D. (2016a). Confer: A conference recommendation and meetup tool. In *Procs. of ACM Conference on Computer Supported Cooperative Work and Social Computing Companion (CSCW)*, pages 118--121, San Francisco, USA.

Zhang, H., Nguyen, D. T., Zhang, H., and Thai, M. T. (2016b). Least cost influence maximization across multiple social networks. *IEEE/ACM Transactions on Networking*, 24(2):929--939.

Zhang, L., Zou, Y., Xie, B., and Zhu, Z. (2014). Recommending relevant projects via user behaviour: An exploratory study on github. In *Procs. of International Workshop on Crowd-based Software Development Methods and Technologies (CrowdSoft)*, pages 25--30, Hong Kong, China.

Zhong, J., Peter, W. T., and Wei, Y. (2017). An intelligent and improved density and distance-based clustering approach for industrial survey data classification. *Expert Systems with Applications*, 68:21--28.

Zhuang, Z., Wagle, R., and Giles, C. L. (2005). What's there and what's not?: Focused crawling for missing documents in digital libraries. In *Procs. of ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, pages 301--310, Denver, USA. ACM.

Zignani, M., Gaito, S., and Rossi, G. P. (2016). Predicting the link strength of newborn links. In *Procs. of International Conference on World Wide Web (WWW)*, pages 147--148, Montreal, Canada.

Zuo, X., Blackburn, J., Kourtellis, N., Skvoretz, J., and Iamnitchi, A. (2016). The power of indirect ties. *Computer Communications*, 73:188--199.

Appendix A

CNARe

Social networks represent individuals and the interactions among them, and studying such networks allows to discover different social patterns [Ahmed et al., 2016; Brandão and Moro, 2017a]. For instance, Chang and Chin [2011] study factors that affect user intention to use a social network game, and Pettenati and Cigognini [2007] use social networks theories to elaborate new e-learning practices. Furthermore, the social networks features can also be used to improve the quality of recommendation algorithms, such as those for friends, music, books and collaborators [He and Chu, 2010; Tang et al., 2013].

Specifically, recommending collaborators is a specific type of people recommendation in which the main goal is to recommend a pair of individuals to collaborate in a determined context. For instance, Surian et al. [2011] extract information from Source forge[1] and build a developer collaboration network. Then, they propose a new algorithm to recommend developers candidate to projects in Source forge. Likewise, Protasiewicz et al. [2016] propose an architecture to recommend reviewers to evaluate researchers' proposals and publications. In this context, this chapter focuses on recommendation of co-authors by considering algorithms that use information available in co-authorship social networks. A co-authorship social network is a type of social network in which nodes are authors and edges represent that they have publications in common.

Advances in collaboration recommendation algorithms have shown the potential to improve researchers' productivity and their groups through establishing new research connections [Brandão et al., 2013; Lopes et al., 2010; Xia et al., 2014]. The recommendation strategies include analyses of the topological features from the co-authorship social networks, semantic properties of the relationship between researchers and math

[1]Source forge: `http://sourceforge.net`

formalizations. Such algorithms provide as result a recommendation list with the top ranked researchers that may collaborate with another researcher.

Besides characteristics of the recommendation algorithms, another relevant aspect of a full system is the visualization of the recommendations results. Generally, the recommendations are presented in sorted lists (according to the recommendation function's result). For instance, Confer (used in IJCAI-16[2]) is a tool that uses recommendation approaches in order to help conference attendees to find talks and papers, to discover people with common interest and manage their time in the conference [Zhang et al., 2016a]. It presents the recommendations as a list, and the users can attribute a star to each recommended item. However, these lists are often not enough to understand how the result was defined or to verify the potential of the recommendations to improve the network as a whole.

Here, the authors present an online tool called CNARe (Co-authorship Networks Analysis and Recommendations) - the pronounce is scenery [de Sousa et al., 2015]. CNARe helps researchers to choose collaborators through automatic recommendations, visualize recommendations, compare the results from different recommendation algorithms and analyze the impact of the recommended researchers in their current network. The tool implements three recommendation algorithms [Brandão et al., 2013; Lopes et al., 2010; Xia et al., 2014]. CNARe also provides other visualizations, for example, comparing the relationship between two or more co-authorship networks from different institutions and analyzing the strength of the co-authorships classified as social link (weak, strong or bridges - a co-authorship that connects researchers from different communities) or random relationship.

After discussing related work on recommender systems and social networks visualizations (Section A.1), the contributions of this chapter are summarized as follows. The CNARe architecture and the processes of collecting and building a dataset from the ACM digital library[3] (Section A.2). The description of the main functionalities and interfaces of CNARe, including the use case diagram and the main features of CNARe's pages (Section A.3). The visualizations of large co-authorship social networks emphasizing the strength of co-authorship links (Section A.4).

A.1 Related Work

In this section, we discuss the related work on recommender systems focusing on people recommendation and social network visualizations.

[2]Confer in IJCAI-16: `http://confer.csail.mit.edu/ijcai2016/schedule`
[3]ACM digital library: `http://dl.acm.org`

A.1.1 Recommender Systems

There are many recommender systems for different contexts, from social networks to e-commerce. These systems can provide recommendations of items (books, papers, songs) or people (friends, co-workers, partners). For instance, Paraschiv et al. [2016] propose a model that considers semantic overlap to recommend papers (items), and Bagci and Karagoz [2016] use the data available on location-based social networks to recommend friends.

Regarding people recommendation, co-authorship social networks have been used to make research teams more productive. The current version of CNARe implements three recommendations algorithms that combine topological properties from co-authorship social networks with academic metrics: Affin [Brandão et al., 2013] considers the shortest path between researchers and the researchers' institutional affiliation; CORALS [Lopes et al., 2010] combines the shortest path between researchers and their research area; and MVCWalker [Xia et al., 2014] uses a random walk model with three academic metrics (the co-author order in the publication, the time of last collaboration and the collaboration frequency).

Regarding similar tools, there are two more related to CNARe: VRRC, which shows the results of only one recommendation algorithm [Barbosa et al., 2012]; and CollabSeer, which recommends researchers considering the co-authorship social networks topology and the interests' areas of a user [Chen et al., 2011]. In CNARe, the generated recommendations consider not only the research area of the researchers, but also the affiliation, co-author order in the publication, the last collaboration time and the collaboration frequency. Moreover, CNARe provides various visualizations (ego-network and global social networks) aiming to show how a recommended collaborator may change an existing co-authorship social network of the researchers who received the recommendation.

A.1.2 Social Network Visualizations

Visualizing social networks may easily provide new insights about users and their interactions in such environment [Viégas and Donath, 2004]. In other words, a visualization is more than simply plotting pictures, it may also facilitate learning and generate new knowledge. According to Freeman [2000], there are two ways to create social network images: drawing graphs in which nodes represent individuals and edges are the connections between them, and plotting a matrix in which rows and columns represent people and the number or color intensity in the cells stands for the amount of social

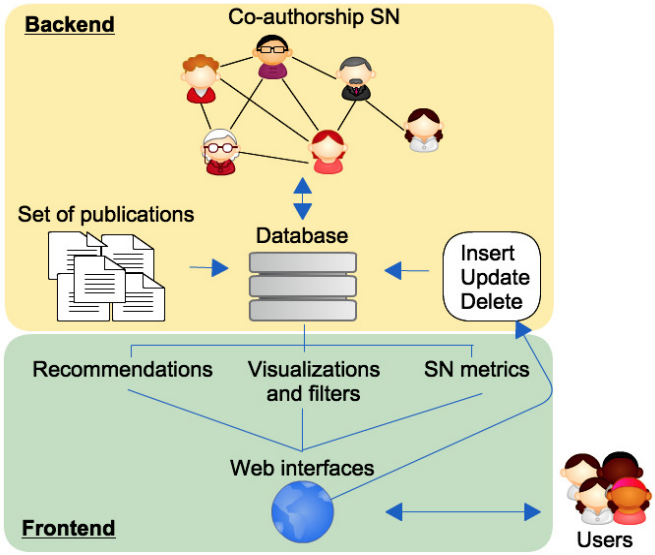


Figure A.1: CNARe architecture.

interaction (e.g., frequency of message exchange, number of co-authorships, the time of interaction) between people. CNARe presents visualizations in both ways.

Furthermore, there are studies investigating the methods that better provide visualizations of large networks. For instance, Rahman and Karim [2016] compare three layouts (force directed drawing, spherical and clustered graph) and provide insights about the three methods that help to identify the best one according to datasets' properties. In addition, Brandes et al. [2012] present different methods to visually explore dynamic social networks. The layout of the visualizations in CNARe is Force-Layout [Holten and Van Wijk, 2009] from D3.js[4] and it was selected through empirical analysis, as it allows analyzing nodes and their interactions in the best way.

Existing social networks visualization tools allow the analysis of different networks. For example, Pajek is a program package that enables analysis and visualizations of large networks [Mrvar and Batagelj, 2016]. Likewise, Network Explorer is a large-network visualization tool that enables users to find clusters of nodes and to identify important nodes in the network [Guerra-Gomez et al., 2016]. CNARe differs from such tools by providing recommendations associated with social network visualizations.

A.2 CNARe Architecture

This section presents how CNARe is built regarding its architecture, data storage and collection, and main functionalities.

[4]http://D3.js: `d3js.org`

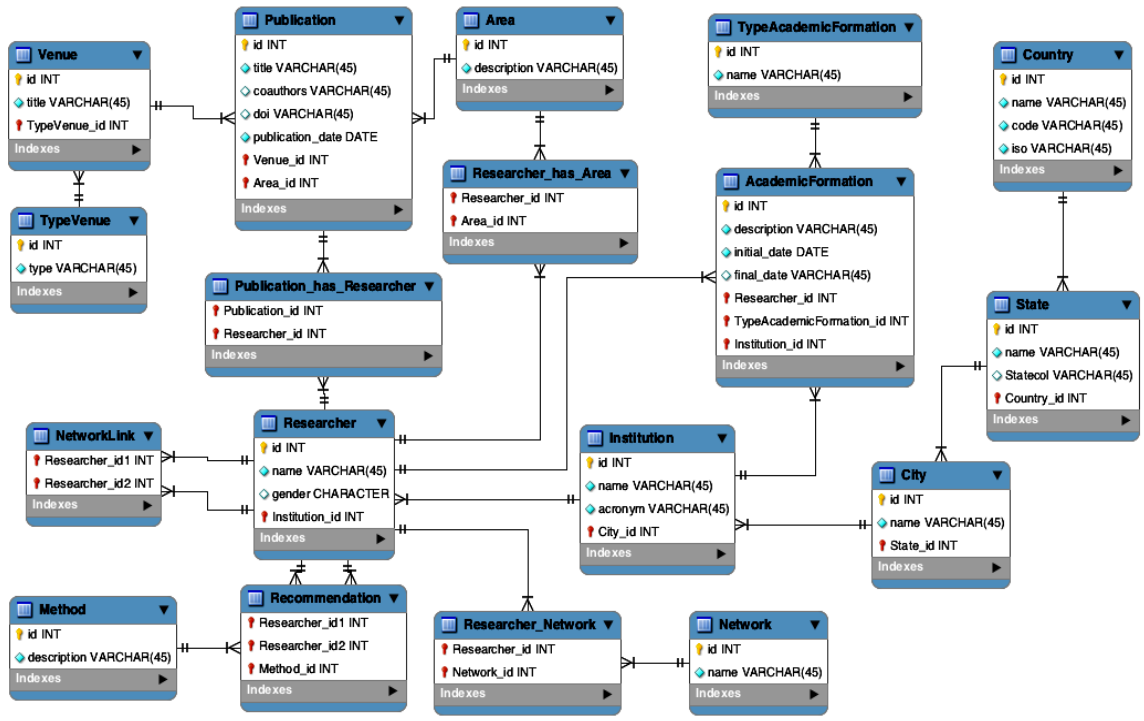


Figure A.2: Relational schema of CNARe database: 16 main tables and two associative tables (Publication_has_Researcher and Researcher_has_Area).

Figure A.1 illustrates a general view of the main components of CNARe divided in backend and frontend. In the backend, CNARe stores researchers' publications in a SQL database, whose relational schema includes tables for researchers, publications and recommendations, as presented in Figure A.2. Each table has B-tree indexes for primary and foreign keys. Other indexes were not created because the performance of the necessary SQL queries is fast enough and could (potentially) harm the performance of inserts, deletes and updates. For each researcher, the recommendations are stored in a table that also identifies what method has generated them. Using data previously stored in the database, the users generate the recommendations for each researcher. However, collaborators can also be recommended in real time when one of the three recommendation algorithm is selected.

Furthermore, each researcher may belong to more than one co-authorship social network (since a user can add various networks with different clusters, for example, by research group, graduate program, institution, and so on). Finally, it is important to note that publications are in the range [2000-2015] in order to ensure recommendations of researchers that have recent work in the area.

The initial database includes publications from Computer Science. The data collecting procedure uses the snowball sampling strategy [Goodman, 1961] and considers

available information in the researchers' page at the ACM digital library. This library was chosen because it presents the area of each publication according to ACM Classification System. Each researcher's page has a publications list, in which each publication has DOI (Digital Object Identifier System), the co-authors list, the date and location of the publication. From the DOI, the specified URL is accessed to get the research area of each publication and information about each co-author: institution, total number of publications and co-author names (since the co-authors list provides the name in citation format). After inserting the co-authors in the database, a new query is executed to obtain the co-author with the largest number of publications whose page has not been visited yet. Then, the collecting process starts again from the page of such an author.

Initially, the data collecting procedure considers researchers from Brazilian institutions (COPPE/UFRJ, PUC/RIO, UFMG, UFPE, UFRGS, UNICAMP, USP/SC, UFF, USP, UFCG, UFES, UFPR, UFRJ, UFRN, UFSC, UFSCAR, UNB, UNISINOS, PUC/PR, PUC/RS, UFAM, UFBA, UFC, PUC/MG, UCPEL, UFG, UFPA) and international institutions (University of Carnegie-Mellon, of Illinois at Urbana-Champaign, of California - Berkeley, of Singapore, Stanford, Chinese Academy of Sciences, of Southampton, of Los Angeles, Tsinghua, among others)[5].

Reducing the noise in the input of the recommendation algorithms requires to filter the number of researchers. In this case, the noise is given by researchers with few publications, for example, as they may be students or not active researchers in their area. Thus, considering these researchers in the recommendation algorithms may generate not useful recommendations (i.e., they are noise data).

For the Brazilian institutions, only researchers with at least 10 publications were considered (such value excludes most students). Regarding international institutions, the previously mentioned ones accounts for 100 researchers with the largest number of publications in the ACM digital library. Such researchers were reached and collected from a seed researcher (Hector Garcia-Molina from Stanford University, one of the researchers with more publications in the ACM). Hence, from Hector Garcia-Molina, the other researchers directly or indirectly linked to him were collected, and the database has researchers from other international institutions not mentioned as well.

Table A.1 summarizes the statistics of the CNARe default database[6] and presents the number of researchers, institutions, publications, average number of co-author per publication, quantity of publications' area and period of the gathered publications as

[5]In the future, the authors plan to consider other institutions as well.

[6]A dump of the relational database is available on `http://www.dcc.ufmg.br/~mirella/projs/apoena`

Table A.1: Description of the dataset stored in CNARe database.

Collected Data	
#Reseachers	6,112
#Institutions	681
#Publications	4,259
Co-authors average	3.98
#Research areas	61
Period	2000-2015

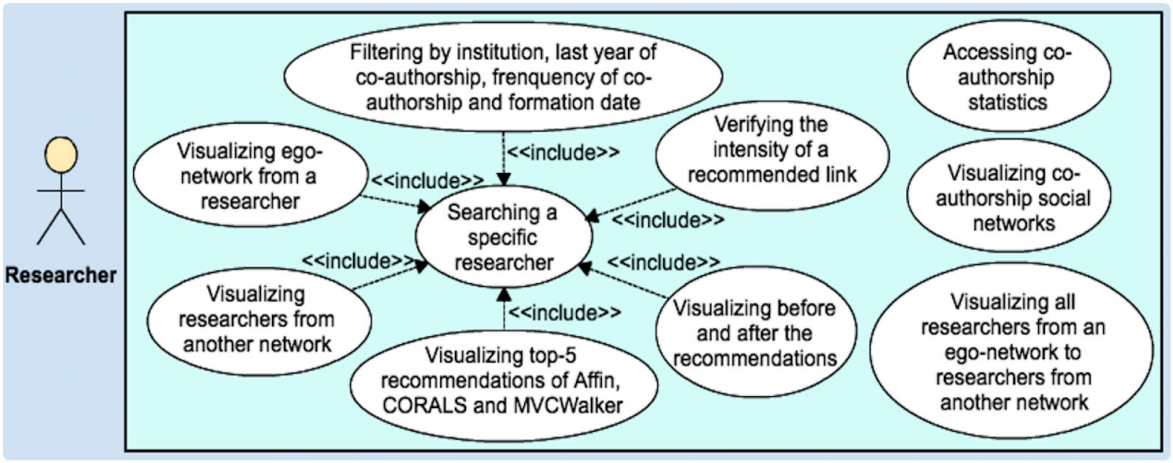


Figure A.3: Use case diagram: a researcher can execute all actions. The include indicates that those actions depending on the search of a researcher.

collected in July 2015. Observe that there are more researchers than publications, because most publications have more than one author. It is important to emphasize that this is a default database in CNARe, as users can also upload data themselves.

In the frontend (Figure A.1), the three main functionalities of the tool are: visualization of the recommendations according to the three algorithms, visualizations with filters, and results of metrics of social network analysis. Note that the last two features aim to improve the presentation and understating of the recommendations. Moreover, CNARe also allows users to import new researchers and their publications through files in CSV format (Comma Separated Values). In order to import a researcher, the CSV file must have the following columns: researcher name, research area, institution, link to the researcher homepage and the year of the last academic degree. Regarding the import of publications, the user has to inform the following columns in the CSV file: title, publication date, research area, authors and venue. This feature allows anyone with publications to use the tool and build new networks. Next section details the functionalities through examples.

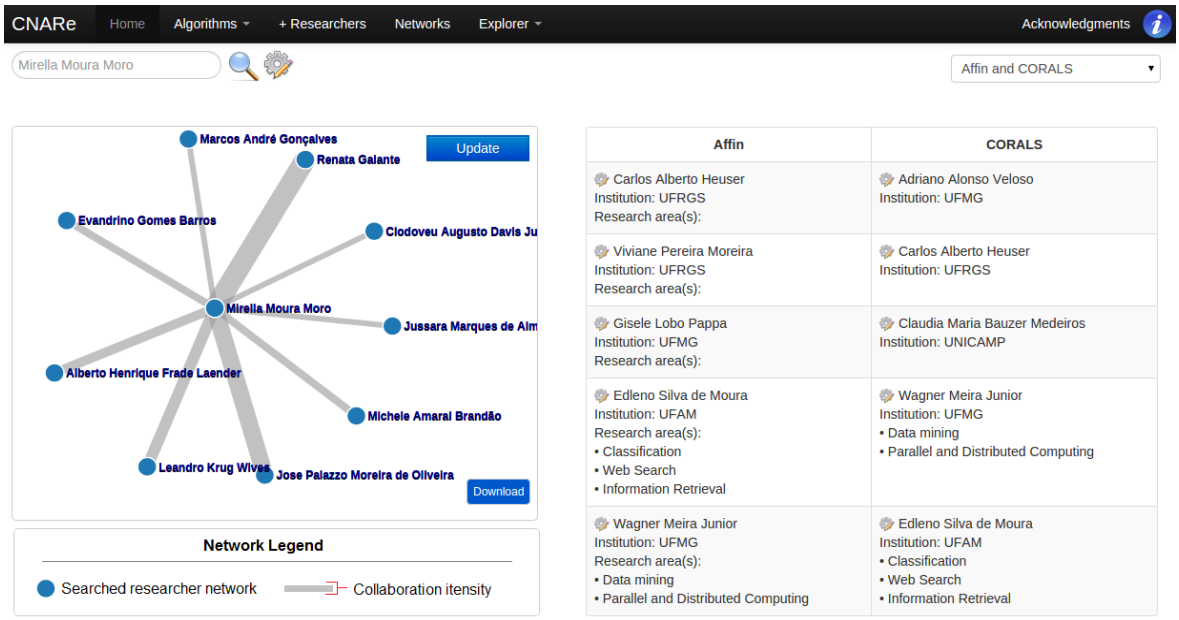


Figure A.4: Main interface of CNARe with recommendations to Mirella M. Moro.

A.3 Design and Interfaces

In order to understand the main functionalities provided by CNARe, Figure A.3 shows a use case diagram. Note that we consider our user as a researcher for simplicity, as the tool could be used by a hiring committee of a department as to make its collaboration network stronger. Next, we detail each functionality (collaboration recommendation, visualizations and filters, and social networks metrics).

A.3.1 Collaboration Recommendation

Figure A.4 shows the initial page for collaboration recommendation. It presents the field to search a researcher from which the user can visualize and compare the top-5 generated recommendations. The comparison is fulfilled in pairs, i.e., three combinations: Affin and CORALS, Affin and MVCWalker, CORALS and MVCWalker. In this example, the ego-network of the researcher is on the left and the recommendations on the right according to Affin and CORALS algorithms.

The page also allows to edit or add information about a researcher stored in the database. To do so, the user clicks on the editing icon that is in the right side of each researcher name. There, the user can change the institution, add new academic formation (it is necessary to insert the start date, conclusion date, the institution of the academic degree) and the research area in each researcher's profile. Such functionality is important to keep CNARe database updated (e.g., correct old information).

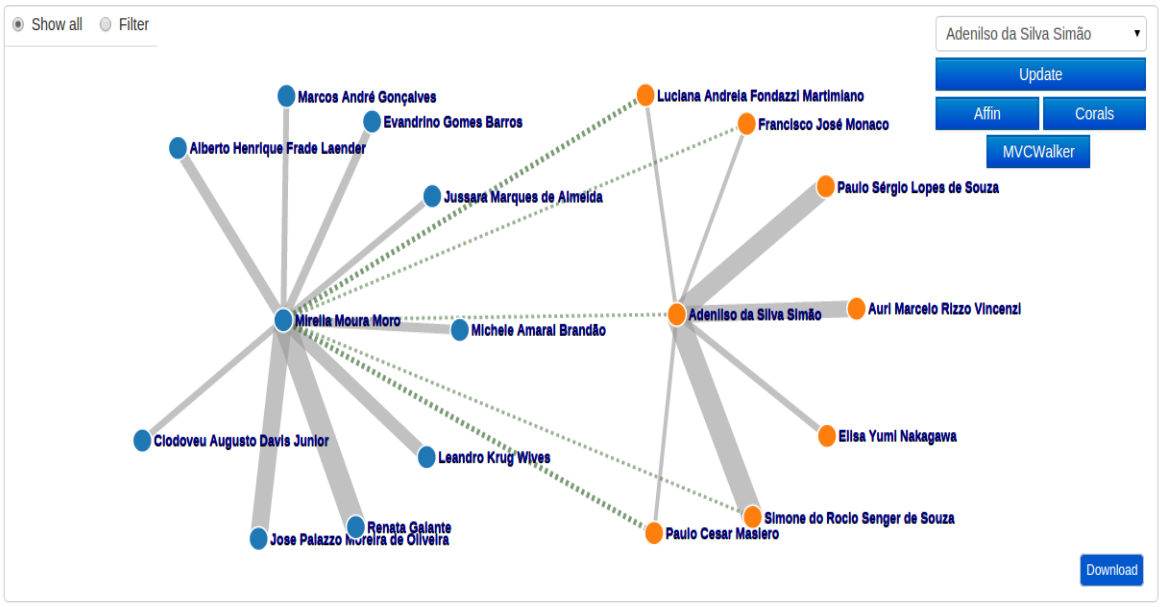


Figure A.5: Green lines represent recommended collaborations: the more intense more has been recommended by the algorithm. The recommendations are generated by clicking in one of the options with the algorithms' name.

For each recommended researcher, the tool presents the institution and research areas, which allow to know more about them. It is also possible to visualize the intensity (score) of each recommendation when moving the mouse over the edge that represents the recommendation and the co-authors of a recommended researcher, and analyze the strength of the recommended collaborations using social networks metrics. Such visualizations contribute to the user understanding how a recommended collaboration may change a co-authorship social network. For example, Figure A.5 shows the recommendations considering the co-authors of a selected researcher, also presents the collaborators with whom a user may have contact by using the recommended researchers as "bridge".

A.3.2 Visualizations and Filters

In order to visualize the co-authorship social networks, the tool provides two options: ego-network of a researcher (examples in Figures A.4 and A.5) and global co-authorship network (example in Figure A.6). The ego-network presents a researcher with her/his co-authorships, aiming to visualize the current collaborations and the recommended ones. On the other hand, the global networks show an eagle-eye vision of all co-authorships from an institution, a researcher (the relationships among co-authors of a researcher), or a network inserted by user. For instance, a user can visualize a global

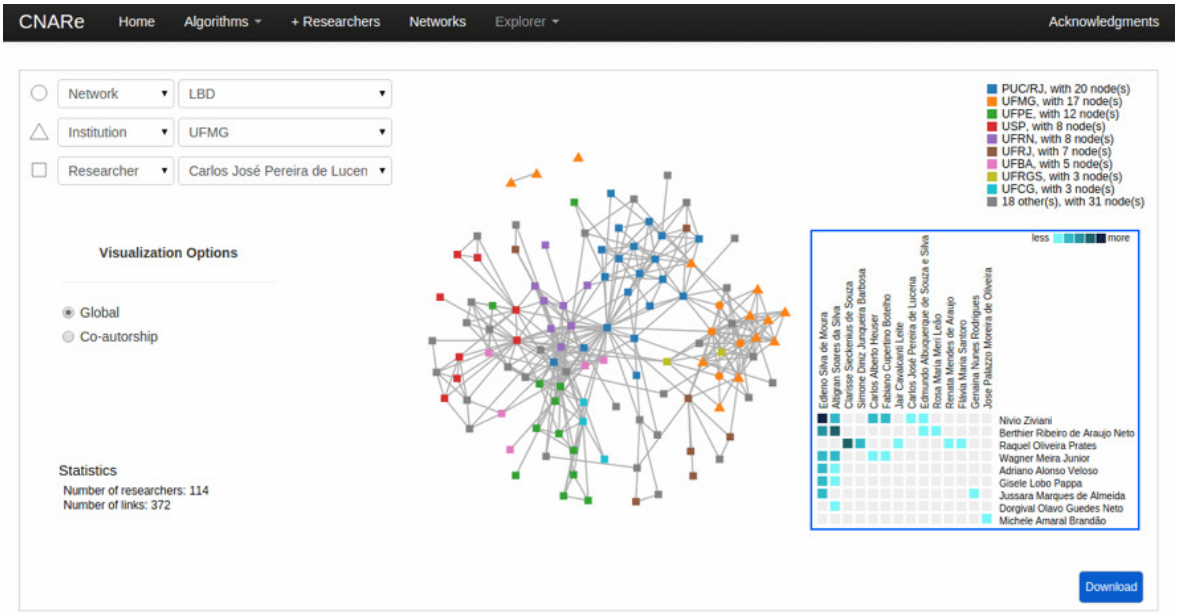


Figure A.6: Global network example: in the Visualization Options menu, when the Co-authorship option is selected, a co-authorship matrix is presented instead of the social network. Here, the matrix is in a blue rectangle.

co-authorship social network of an institution from a recommended researcher.

Moreover, CNARe also allows to compare three co-authorship social networks (accessing the option Compare More). This enables, for instance, to analyze which institution of the five recommended researchers has the densest co-authorship network. The visualizations can be personalized through four filters applied to the co-authorships: (1) by institution, it shows only the links of co-authors from a selected institution; (2) the last time that the co-authorship happened, it focuses on co-authorships in a period of years (e.g., the last two years); (3) the co-authorship frequency, it presents the links between researchers that published together a selected number of times (e.g., from 2 to 5 publications in common); and (4) the date of the last academic formation, it allows to consider researchers from similar academic "generations" (e.g., removing researchers who have retired or are too young).

A.3.3 Social Networks Metrics

CNARe also presents the results for social networks metrics. Considering the ego-networks, it may be hard to find relevant metrics, because the networks are relatively small, with not enough information. For ego-networks, CNARe applies three metrics (Easley & Kleinberg, 2012; Wasserman & Faust, 1994): (1) neighborhood overlap is the number of nodes that are neighbors of both researchers involved in a co-authorship

Table A.2: Description of the large social networks stored in CNARe database.

Dataset	Number of nodes	Number of edges
PubMed	443,784	5,550,294
DBLP Articles	837,583	2,935,590
DBLP Inproceedings	945,297	3,760,247
APS	180,718	821,870

divided by the number of nodes that are neighbors of at least one of the researchers in a co-authorship; (2) clustering coefficient is the probability that two randomly selected co-authors of a researcher are also connected to each other; and (3) affiliation homophily is the measure of the similarity between a pair of researchers considering their institution.

Specifically, the neighborhood overlap metric presents the strength of the recommended links, which allows to analyze if each recommended link will be a bridge (i.e., an edge responsible for connecting different communities and not connected yet) or not. The clustering coefficient and homophily metrics show how the recommendations affect the researchers' networks from different perspectives.

CNARe also presents statistics of the global co-authorship social networks, including the number of nodes in each network (of a researcher, institution or uploaded by a user) and in the global network, the amount of connections and the frequency of co-authorships (Figure A.6). These statistics allow to understand the topology of the social networks.

A.4 Advanced Social Networks Visualizations

The goal is now to provide the visualizations of large social networks and distinguish the links according to their strength. Thus, CNARe also allows the visualization of co-authorships from different datasets, such as PubMed (US National Library of Medicine National Institutes of Health), DBLP (Computer Science Bibliography) and APS (American Physical Society), which offer insights on the organization of these different social networks. Table A.2 presents the number of nodes and edges in each social network, as collected in April 2016 for PubMed, September 2015 for DBLP, and March 2016 for APS.

The data from PubMed was gathered through the e-utilities offered by the National Center for Biotechnology Information. The e-fetch utility allows to make queries to the NCBI's database. The queries aim to collect data from publications from the most prestigious venues in the health and medical sciences according to h5-index (h-index of those papers published in the last five years [Bornmann and Daniel, 2007]).

Likewise, DBLP's dataset was taken from Universität Trier website, which is then split into two different datasets (due to its large size): one for the social network considering authors' common articles as the edges, and another considering inproceedings. Regarding the APS, the authors get access to a sample dataset in JSON format. Then, such file was parsed in order to insert the data in a relational database and to build a social network.

In CNARe, the visualizations of those social networks show the edges classified according to their strength. In order to do such classification, $fast-RECAST$ *Random rElationship ClASsifier sTrategy with Multiprocessing modules* is applied to the social networks. Such algorithm classifies the edges as social links (friends, acquaintances or bridges) or random links [Vaz de Melo et al., 2015]. Here, the edges classified as friends are called strong links and acquaintances as weak ones. Bridges and random links maintain the same name.

Overall, $fast-RECAST$ classifies an edge as social when two characteristics are present in the relationship: regularity and similarity. The regularity indicates that a relationship repeats over time, whereas the similarity means that two individuals in a relationship have common neighbors. Such characteristics can be mapped into social network metrics as edge persistence and topological overlap (also known as neighborhood overlap), respectively.

Considering such properties, Vaz de Melo et al. [2015] evaluate which combination of values of edge persistence and topological overlap define a relationship as friends (strong), bridges, acquaintances (weak) or random, and compare the results with a random graph (a random version of with the same number of nodes, edges and degree distribution; the only difference is the way that nodes are connected to each other).

In the CNARe page that shows the visualization of the edges classified, the selection of nodes from each dataset can be done by researcher ego-network, publication venue or publication area, enabling comparison of different social networks topology. Also, the representation of co-authorships has been modified through distinct shapes and colors to show different properties, allowing visualization of bridges (a co-authorship that connects researchers from different communities) or classification of co-authorships as weak or strong. Figure A.7 presents the visualization of a social network from a venue in the PubMed dataset.

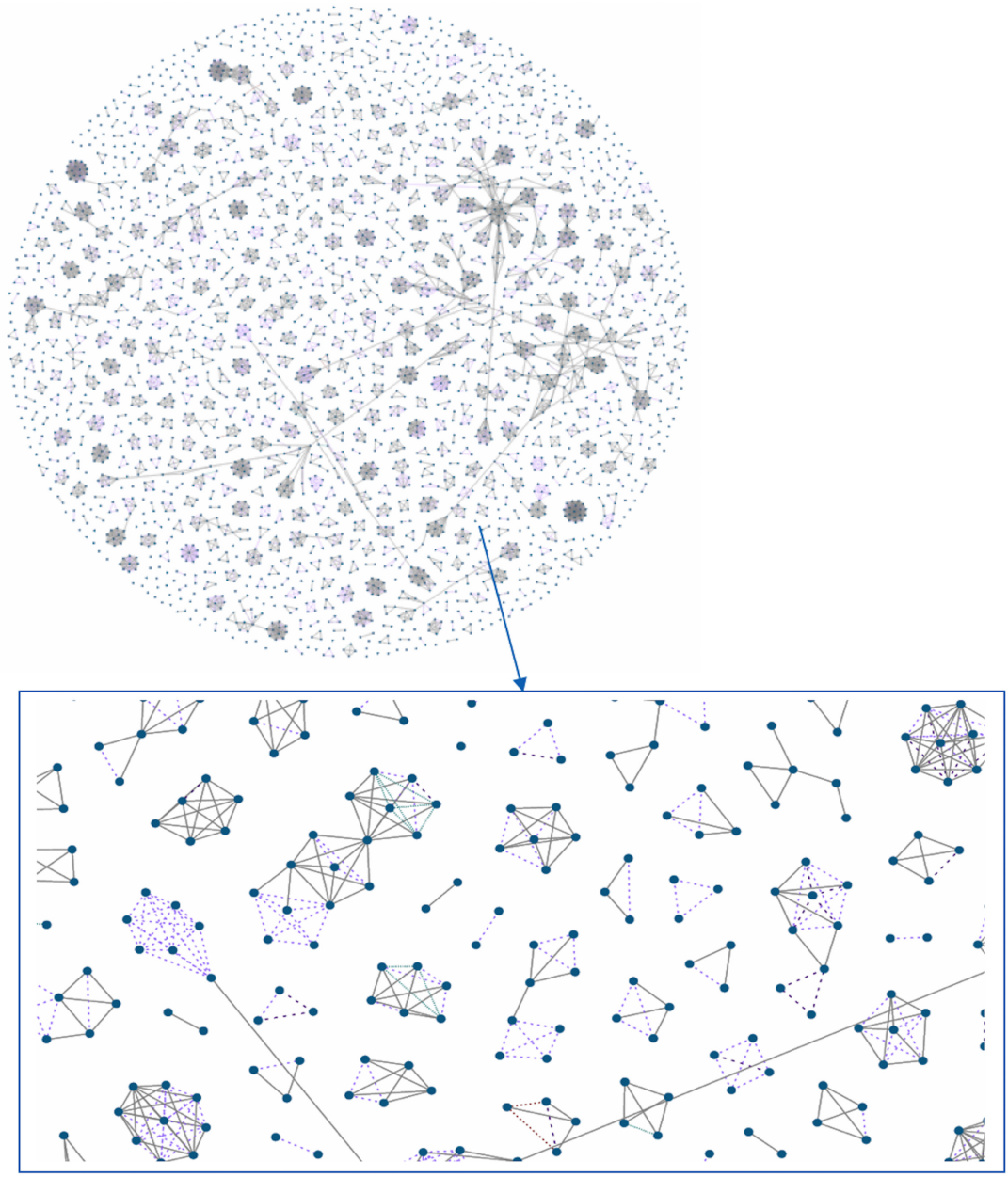


Figure A.7: Visualization of PubMed social network from the venue Lancet Medical Journal (London, England). The green edges are strong links, red edges are bridges, purple edges are weak links and dark purple edges are random links. The gray edges are links that do not received any classification.

A.5 Concluding Remarks

This chapter presented CNARe, an online tool that shows the collaboration recommendations of three different algorithms (Affin, CORALS and MVCWalker). Visualizations and metrics of social networks are also used in order to show how the recommendations may modify researchers' ego networks. The visualizations reveal

that new recommended links may work as bridges to co-authorship social networks. All these visualizations represent a step-forward in the collaboration recommendation tools, because CNARe considers three recommendation algorithms instead of only visualizing the results. Furthermore, besides CNARe having initial datasets, others can be easily uploaded. The only requirement is that the data have the fields needed by the recommendation algorithms. Finally, CNARe also provides visualizations of large networks differentiating the edges classified (strong, weak, bridges or random) by $fast - RECAST$ algorithm.

In the future, the authors plan to include other recommendation algorithms and social networks metrics. The next steps also comprise a better differentiation of the edges regarding their strength (i.e., consider algorithms different from $fast - RECAST$).